Imperial College London

EPSRC Centre for Doctoral Training in NEUROTECHNOLOGY for life and health

Doctoral Thesis

Imperial College London

Department of Bioengineering

Variational tomography

A low-cost approach to estimate reconstruction uncertainty with an application to full-waveform inversion

Author:

Oscar Bates

Supervisors:

Prof. Mengxing Tang

Dr. Lluis Guasch

Dr. Oscar Calderon Agudo

Dr. Matthew Williams

3rd January 2023

Abstract

Bayesian methods are a popular research direction for inverse problems, but this framework has not achieved widespread use in tomographic imaging. Primarily, this is because traditional tomographic uncertainty estimators are computationally expensive, but also because there is no well established way to interpret tomographic uncertainty estimates. This thesis presents a Bayesian model for tomography, where the tomographic image is assumed to follow a mean-field Gaussian distribution. Using this assumption, an iterative reconstruction algorithm can be derived which produces two images, 1) a *maximum a posteriori* (MAP) estimator, which is equivalent to a typical tomographic image reconstruction, and 2) a no computational cost tomographic uncertainty estimate, where each pixel of the reconstruction is associated with an individual standard deviation or variance estimate. In practice, the algorithm is derived using stochastic variational inference, which motivated its name; variational tomography (VT). VT is demonstrated by application to full-waveform inversion (FWI). FWI is a computationally complex reconstruction algorithm because the tomographic data are simulated using a full numerical simulation of an acoustic wavefield. The mean-field Gaussian is a restrictive assumption, which has implications for the accuracy of the posterior. However, it is demonstrated that the MAP estimator is equivalent to a traditional iterative reconstruction and the no cost uncertainty estimate has some useful properties. Specifically, the tomographic uncertainty estimates are evaluated *in-silico* and *in-vitro*, qualitatively and quantitatively, for the purpose of image quality analysis (IQA). A comparison between different methods of uncertainty estimation is not provided because different uncertainty estimators make different assumptions; there is no reliable ground truth.

Instead, this study demonstrates that VT uncertainty is a useful tool for no-reference IQA. Thus, VT does not alter the dynamics of a traditional iterative reconstruction algorithm, but provides an additional tomographic uncertainty estimator with no additional computational cost.

Statement of Originality

The material in this dissertation is my own work or has been properly acknowledged.

Oscar Bates

Copywrite declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law

Acknowledgement

I would like to start by thanking my supervisors, Mengxing Tang, Lluis Guasch, Oscar Calderon-Agudo, and Matthew Williams. This has been an exciting and interesting research project, and I am hugely grateful to them for the opportunity to work on it. I am also grateful to Mike Warner. At one stage or another, each of them has also provided me with helpful advice and support and I would not have succeeded without them.

No-one does their research in a vacuum, and I am fortunate to conduct my research with many colleagues from the *Ultrasound Laboratory for Imaging and Sensing* and the *Transmission and Reflection Ultrasound Laboratory*. These groups cover much ground in ultrasound and it has been brilliant to spend time with each and every lab member. I am particularly grateful to colleagues who have helped me directly by sharing their software and datasets: George Strong, Carlos Cueto, Javier Cudiero, and Thomas Caradoc Robins.

My work was supported by the EPSRC Centre for Doctoral Training in Neurotechnology under Grant No. EP/L016737/1 and the Department of Bioengineering at Imperial College London.

Doctoral research is demanding personally and I am truly grateful for the support of my family and friends. I want to especially thank mum and dad for teaching me the value of hard work and almost everything else I know. I also want to thank my wife for her endless support and my young son for brightening up the difficult final months of my project. I rely on all of them, and my other friends and family, for their help and advice. I would not have achieved anything without them.

Contents

Abstract i

Statement of Originality iii

Copywrite declaration v

Acknowledgement vii

1 Introduction 1

1.1 Ultrasound computed tomography 2

1.2 Full-waveform inversion in Geophysics 6

1.3 Image quality assessment 8

1.4 Uncertainty estimation 9

1.5 Thesis structure and contributions 12

2 Numerical methods and full-waveform inversion 15

2.1 The physics of ultrasound propagation 15

2.2 Numerical modelling of wave propagation 20

2.3 Full-waveform inversion 24

2.4 A PyTorch inversion algorithm 30

2.5 Transducer configuration 32

3 Variational tomography: theory and practice 35

3.1 Introduction 36

3.2 Theory . . . 40
3.2.1 The maximum likelihood interpretation of FWI . . . 40
3.2.2 The variational Bayesian interpretation of FWI . . . 44
3.3 Method . . . 51
3.3.1 An optimisation perspective . . . 52
3.3.2 The mean-field Gaussian approximation . . . 53
3.3.3 The adjoint-state method and stochastic variational inference . . . 57
3.3.4 The MIDA *in-silico* dataset . . . 59
3.4 Results . . . 61
3.4.1 Starting sound speed . . . 61
3.4.2 Optimisation hyperparameters . . . 63
3.5 Discussion . . . 65
3.5.1 Comments on the results . . . 66
3.5.2 Future work . . . 69

4 Qualitative validation *in-vitro* 71
4.1 Introduction . . . 72
4.2 Method . . . 75
4.2.1 The *in-vitro* dataset . . . 75
4.3 Results . . . 77
4.3.1 Progressive decoding . . . 78
4.3.2 Detecting reconstruction failure . . . 80
4.3.3 Artefact detection . . . 81
4.4 Discussion . . . 82
4.4.1 Comments on the results . . . 82
4.4.2 Usefulness . . . 84
4.4.3 Future work . . . 86

5 Quantitative validation *in-silico* 89
5.1 Introduction . . . 90

5.2 Method . . . 93
5.2.1 Image quality analysis metrics . . . 94
5.2.2 The circle *in-silico* dataset . . . 95
5.3 Results . . . 98
5.3.1 Density . . . 98
5.3.2 Absorption . . . 101
5.3.3 Position errors . . . 104
5.4 Discussion . . . 108
5.4.1 Comments on the results . . . 108
5.4.2 Usefulness . . . 110
5.4.3 Future work . . . 111

6 Conclusions **115**

7 Future work **117**
7.1 Prior information . . . 118
7.2 Learnt statistical priors . . . 120

List of Figures

2.1 The unit cube of the stress tensor s_{ij}. When a force is applied along the axis i, it provokes stresses in all the directions j. These stresses have parallel and perpendicular components. 16

2.2 The isotropic modes of propagation for the unit cube. (a) shows that strain F induces a compression mode which is supported by elastic and fluid materials, (b) shows that strain G induces a shear mode which is only present in elastic materials. 20

2.3 (a-d) The simulated wavefield, shown for different timesteps, surrounded by a ring of receivers on the black oval. (e) Wavefield data are obtained when the displacement or pressure is measured at the receivers. The arrows indicate common features of the wavefield and wavefield data 26

2.4 (a) A simulated example of observed data using a 2-D slice of the MIDA head phantom [126]. (b) Predicted data is generated assuming a constant sound speed of 1480$[m\,s^{-1}]$, and (c) the L2 loss, showing the difference between the observed and predicted data 27

2.5 The adjoint method produces an estimate of the gradient by correlating the predicted and adjoint wavefield. The black arrows show where the predicted and adjoint wavefield have the strongest correlation. 29

3.1 (a) The sound speed reconstruction (left), variance reconstruction (centre), and absolute error (right) for a water start. (b) The sound speed reconstruction (left), variance reconstruction (centre), and absolute error (right) for a start where the skull is correct. 62

3.2 A grid search over batch size. Batch size is optimised between 6 and 48 sources, step length is 5. 63

3.3 A grid search over step length. Batch size is 24 sources, step length is optimised between 2 and 15. 64

4.1 (a) The ultrasound tomography apparatus with the PVA tissue-mimicking phantom (centre) surrounded by two P4-1 transducers which are held mechanically in suspension from two rotary motors. (b) The design-image used to produce the moulds for the phantom. The transmitted source wavelet is shown in the time (c) and frequency (d) domains. 76

4.2 Reconstructions using VT-FWI and the *in-vitro* dataset, low-pass filtered at 700 $[kHz]$. (a) The average variance per iteration for 20 different reconstruction processes. (b)-(d) Examples of the mean (left) and variance (right) reconstructions corresponding to the process indicated by the cyan line in (a). 79

4.3 Reconstructions using VT-FWI and the *in-vitro* dataset, low-pass filtered at 1 $[MHz]$. (a) The average variance per iteration for 20 different reconstruction processes. (b)-(d) Examples of the mean (left) and variance (right) reconstructions corresponding to the process indicated by the cyan line in (a). 80

4.4 (a) An uncropped replica of the result in Figure 4.2(d), where the crop-box is indicated by the blue dotted square. (b) Another example reconstruction produced using the 700$[kHZ]$ filtered data, but in this case the posterior reconstruction contains an artefact due to source positioning error. 81

5.1 A circlular phantom is used for the quantitative evaluation which has (a) a sound speed of 1550$[ms^{-1}]$. (b) The circle can have variable alternate physics, for example density or attenuation. The black circle indicates the location of the transducers. 96

5.2 VT-FWI reconstructions of the circle dataset. MSDM per iteration is shown for (a) constant density of 1010 $[kg\,m^{-3}]$ and (b) variable density where the circle phantom is 1700 $[kg\,m^{-3}]$ for 3 different reconstruction processes. (c-f) reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue ring, indicating the region of interest. 98

5.3 The quantitative relationship between mean standard deviation magnitude and mean absolute error where density incudes an artefact. (a) MSDM vs density, (b) MAE vs density, (c) MSDM vs MAE, and (d) MSDM and MAE vs density for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue ring, indicating the region of interest. 100

5.4 VT-FWI reconstructions of the circle dataset. MSDM per iteration is shown for (a) no absorption and (b) variable absorption where the circle phantom is 10$[dB\,cm^{-1}]$ for 3 different reconstruction processes. (c-f) Reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue circle, indicating the region of interest. 102

5.5 The quantitative relationship between mean standard deviation magnitude and mean absolute error where absorption includes an artefact. (a) MSDM vs absorption, (b) MAE vs absorption, (c) MSDM vs MAE, and (d) MSDM and MAE vs density for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue circle, indicating the region of interest. 103

5.6 VT-FWI recosntructions of the circle dataset. MSDM per iteration is shown for (a) correct transducer positions and (b) random transducer perturbations sampled from 0.2[mm] wide Gaussian for 3 different reconstruction processes. (c-f) Reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue circle, indicating the region of interest. 104

5.7 The quantitative relationship between mean standard deviation magnitude and mean absolute error where resolution is reduced by transducer misplacement. (a) MSDM vs position error, (b) MAE vs position error, (c) MSDM vs MAE, and (d) MSDM and MAE vs position error for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue circle, indicating the region of interest. 106

5.8 VT-FWI reconstructions of the circle dataset, where 10 transducers are systematically displaced 1 [mm] vertically. Reconstructions of (a) mean, (b) standard deviation magnitude, and (c) the absolute error with 3 repetitions. . 107

5.9 Wavefield data are shown when there is (a) no absorption and (b) absorption of 10 [dB]. Plots of the (c) reflected component (receiver ID 50) and (d) transmitted component (receiver ID 178) of the wavefield for different absorption coefficients, the approximate location of the plots are indicated by the red lines in (a). 112

7.1 L2 regularisation is introduced progressively by adjusting the weight term. (a) shows the effect of regularisation on the mean, (b) shows the unregualarised standard deviation of the likelihood, and (c) shows the regularised standard deviation of the posterior. 121

7.2 Examples from the RSNA dataset (a-i) before preprocessing and (j-r) after preprocessing. Sub-figures show saggital (left), coronal (centre), and axial (right) slices. 123

List of Tables

3.1 The time taken to compute the mean and standard deviation update on different devices when batch size is 24. 61

4.1 The time taken to compute the mean and standard deviation update on different devices for a batch size of 10. 77

Chapter 1

Introduction

In almost all areas of observational science and technology, a measurement consists of two components: 1) the expected or mean value of the parameter being measured and 2) an estimate of its variability, reliability, or uncertainty. Both components are vital to interpreting the measurement; indeed, in many disciplines, a mean estimate without an uncertainty estimate is considered useless. In medical imaging, the mean estimate is a picture representing the anatomy under investigation. Mean estimate examples include T1 or T2 contrast images magnetic resonance imaging (MRI) or relative absorption in x-ray computed tomography (CT). Unfortunately, in most imaging studies, uncertainty estimates are never provided. The absence of uncertainty estimates are not because the images are flawless or because practitioners are unaware of their value; uncertainty estimates can be used for image quality analysis (IQA) and other tasks. Unfortunately, tomographic uncertainty estimates (i.e. pixel-wise or voxel-wise) have proven extremely difficult and computationally unaffordable in practice.

Historically, technological developments in medical imaging try to reduce uncertainty to negligible levels by escalating the volume of acquired data and, therefore, improving the image quality. This approach is effective, but can have negative consequences. First, these quality improvements come at the expense of the patient (i.e. long scan times for MRI [1], or higher radiation-dose in CT [2]) and, second, because many image quality issues,

for example artefacts, are the direct result of failing physics approximations. The latter problems is difficult to resolve because physics approximations are required to solve the inverse problem.

1.1 Ultrasound computed tomography

Ultrasound computed tomography (US-CT) is a novel imaging modality which promises safe, low-cost, high-resolution imaging of the breast [3]–[9] , bone [10]–[12], and brain [13], [14]. These qualities mean that US-CT has the potential to transform patient care, as detailed below. Prominent among US-CT reconstruction algorithms is Full-Waveform Inversion (FWI), a technique which uses a full numerical simulation of the wavefield and has revolutionised Geophysical imaging [15] due to its high-resolution limit and ability to compensate for non-linear acoustic effects. The nascent field of US-CT offers an opportunity to introduce tomographic uncertainty estimation because the novelty of US-CT will necessitate a period of clinical adoption, during which clinicians may be more open to new ways of working. US-CT imaging is also a testing ground for the more established imaging modalities such as CT or MRI; demonstrating effective tomographic uncertainty estimation in US-CT will inevitably increase interest from other modalities.

US-CT brain imaging. Globally, in 2016, neurological disorders caused 9 million deaths and 276 million disability-adjusted life-years, making neurological disorders the second leading cause of death and primary cause of acquired disability [16]. Alone, stroke contributed 12 million cases [17], 5.5 million deaths (61% of neurological, 10% of total), and 116 million disability-adjusted life-years (42% of neurological). This means stroke was the second leading cause of death and disability [17] and the leading neurological contributor [16]. The main diagnostic tool for stroke is structural brain imaging, namely CT and MRI [18]. Both of these modalities are non-portable and relatively expensive, MRI reconstruction is slow, and CT uses ionising radiation. In other words, the patient must be brought to these systems and they are unsuitable for real-time monitoring. This is a critical problem because the treatment of stroke is time-sensitive; any reduction in the time between onset and treatment is known to improve outcomes [19]. Thus, there is a need for an imaging system which is

portable, because this would enable the imaging system to be taken to the patient. There is also a need for a system which is inherently safe, meaning it could used to continually monitor the health of the patient. Finally, 70% of strokes occur in low- and middle- income countries [20], [21]; it is essential that such a system be cheap to maximise benefit. US-CT fits these parameters; a "helmet" style system would be portable, and ultrasound is well-known for being safe and cost efficient.

The divide between B-mode ultrasound imaging and US-CT is not clearly defined because, technically speaking, B-mode ultrasound is a tomographic imaging modality and there is a current trend towards wide angle apertures using multiple probes [22], [23]. Early work by Smith *et al* [24] proposed the use of B-mode ultrasound imaging for neuroimaging. This development recognised the need for low-frequency ($0.5-1\,[MHz]$) transducers so that skull tissue does not entirely reflect and absorb the wavefield [25] and stood in contrast to standard protocols at the time, which used $2.25\,[MHz]$ transducers. Research continues on transcranial B-mode ultrasound with attempts made to correct the aberration from the human skull bone [26], [27], for the particular task of haematoma [28], or for detecting foreign bodies in blast injury patients [27]. Unfortunately, the B-mode approach is fundamentally flawed because the skull-tissue boundaries have a high reflection coefficient, but brain tissues do not [29]. This means that brain tissues have very low contrast using conventional B-mode imaging.

I will define US-CT by the use of transmission information to drive the reconstruction algorithm. This is unconventional compared to B-mode ultrasound, and there are far fewer US-CT studies. Using this definition, Dines *et al* [30] were the first to attempt US-CT neuroimaging through the skull. This study used the filtered back-projection algorithm to produce absorption maps for an infant and adult head *ex-vivo*. The US-CT scan of the infant brain showed reasonable fidelity compared to a contemporaneous CT scan, but the US-CT scan of the adult brain was severely degraded. This is because the skull is also a formidable barrier for US-CT and was the reason that an earlier publication aimed to characterise the tissue properties of an excised brain without the skull [31]. That study demonstrated that grey and white matter could be differentiated according to absorption and sound speed, a

finding which is supported by a modern study of the 2-D acoustical properties of a sheep brain [32].

To date, only two other proposals exist. First, a belt-type conformal ring of ultrasound transducers [33], where diffraction tomography is used to reconstruct the data [34]. No results have been provided to validate the success of this approach. Second, FWI was proposed and validated *in-silico* [13], with further results available for *in-vitro* models [35]. This approach uses FWI to invert the observed wavefield data from a 3-D, hemispherical array of transducers. Crucially, this study showed that a prior estimate of the sound speed of the skull and a 3-D inversion are necessary to achieve convergence.

The difficulty specifying prior information was the initial motivation for the development of the low-cost Bayesian framework for FWI explored in thesis: a Bayesian framework will enable the incorporation of strong statistical (i.e. not patient specific) prior information without significant additional computational cost. However, further proposals suggest that starting models may not be necessary if an appropriate algorithm design is chosen [14]. Only the accurate inversion of an *in-vivo* dataset will prove whether priors are necessary for FWI neuroimaging.

US-CT breast imaging. The advantages of US-CT are also important for medical screening. Today, mammography is the most widespread medical screening programme and is established in dozens of countries. Mammographic screening is designed to identify breast cancers before the onset of symptoms because early intervention is known to improve patient outcomes [36]. Unfortunately, mammographic screening suffers from significant problems; for example, false negatives, prominently the difficultly identifying cancers in women with dense breast tissue [36], [37], and false negatives, such as the diagnosis of non-life threatening tumours [36], degrade the benefits of the programme. Mammography is also fundamentally flawed due to its reliance on ionising radiation, which induces one cancer per 400-800 detected [38]. Clinical trials suggest US-CT can resolve these problems [39] and US-CT been given premarket approved by the FDA for use additionally to mammography for women with dense breasts [9], [40] and for monitoring treatment response [41].

Issues with safety and specificity in mammography have led researchers to develop US-

CT for breast imaging applications. The literature here is more extensive that in brain imaging, with clinical systems already being marketed [9], [40]. The earliest method used the algebraic reconstruction technique [42] where the system is assumed to follow a linear forward model and the sinogram of the ultrasound data is inverted using a predefined expression for the update. Another contemporary approach was to use time of flight tomography [43], [44], where the time of arrival of the transmitted wavefront is selected and the sound speed is estimated directly or iteratively based on the distance between the source and receivers. Since then, there have been a huge variety of proposals for US-CT mammography reconstruction algorithms [3]–[7], [9], [41], [45]–[69]. Of these methods, 3 systems have presented *in-vivo* reconstructions.

First, the QT Ultrasound system [9], [11], [48], [49], [52], [64]. Many ultrasound systems use transducers which are capable of transmitting and receiving ultrasound waves. In contrast, the QT Ultrasound system uses a specifically designed transmitter and an independent receiver in a transmission configuration [48]. The system incorporates additional reflectors to obtain long offset wave propagation, and the whole unit rotates to give full 3-D coverage of the breast. The system has been demonstrated for full 3-D reconstructions *in-vivo*, with approximately 500 volunteers studied [9]. The reconstruction algorithm uses a variant of FWI, where the paraxial approximation is used to speed up the reconstructions [49]. The paraxial approximation assumes most of the wave propagates forward from the source (i.e. the transmitter) to the receiver. 3-D reconstructions of the data, as opposed to reconstructions of 2-D slices, was found to be critical for image quality [9].

Second, the Delphinius system [3], [4], [41], [51], [53], [57], [62], [67]. This system uses a ring array of transducers which is rastered vertically to obtain 3-D information [4], [57], [62]. It has been demonstrated for 2-D reconstructions *in-vivo* [51], [53] and 3-D reconstructions using FWI [41], [67]. This technique uses frequency domain FWI, where the time stepping scheme commonly used to calculate the gradient (e.g. see in Figure 2.5) is replaced with a Fourier transform. FWI is a popular reconstruction algorithm for US-CT data and additional FWI breast imaging studies include [5]–[7], [59].

Finally the KIT system [45], [47], [63], [69]. This system consists of a hemispherical

array of transducers and has been used to obtain *in-vivo* images of several patients [69]. Reflection-mode reflectivity images and transmission-mode sound speed images were produced. The reconstruction algorithm for this system used synthetic aperture beam forming for the reflection images and the algebraic reconstruction technique for the sound speed images [47], [69], but several other methods have been proposed including time of flight tomography, contrast source inversion, and bent ray tomography [45], [63].

It is apparent that 3-D reconstructions are necessary to improve image quality. This strengthens the need for automatic diagnostics because, in addition to the flaws of intuitive image analysis previously discussed for mammography, evaluating volumetric images significantly increases the time it takes a clinician to analyse an image. The problem is that automatic image analysis tools can become confounded by image quality; they can learn to identify image quality problems rather than disease to the extent that clinical trails of e-ASPECTs - a tool used to detect blood clots - excluded artefactual images [70]. Therefore, mitigating confounding is the second motivation for tomographic uncertainty because if tomographic uncertainty can locate and quantify image quality problems, then it might improve intuitive analysis and enable automatic analysis on poor-quality images.

1.2 Full-waveform inversion in Geophysics

Full-waveform inversion is a tomographic reconstruction algorithm originally developed for Geophysical applications [71], [72]. Previous seismological investigations relied on travel time information only, but seismic waves contain temporal and amplitude information. An approach to access amplitude information, called migration, had been developed a decade before FWI, which back-propagated the observed seismic data and correlating this with a predicted wavefield [73]. The key development of FWI was to recast migration as a local optimisation problem, which reunifies amplitude and phase information into a single solution. FWI operates by making predictions of the seismic data using a numerical simulator, and comparing these predictions to the observed seismic data using a loss function.

Unfortunately, FWI is computationally demanding. For this reason, numerical examples in 2-D geometries were not immediately available. By 1986, the first 2-D study was pro-

duced [74], which included examples of single and dual point scatterers, and a circular disk phantom. The study demonstrated that multiple sources were important for removing artefacts and transmission information was required to correctly reconstruct the circular phantom. The cycle-skipping phenomenon was also investigated, which occurs when the observed and predicted wavefield data arrive more than 90^o out-of-phase. Cycle-skipping causes the reconstruction to fail and is a common difficulty for FWI imaging, but it can be resolved by low-pass filtering the wavefield data [75].

Despite these findings, FWI took a long time to be recognised as an efficient seismic imaging modality. The traditional setting for FWI was as a local, gradient-based, optimisation algorithm. This improves the computational efficiency of FWI but could not solve reflection datasets, which are common in seismology. For this reason, global and semi-global optimisers were tried [76]–[79] but this approach requires long compute time to succeed.

The key problem is that most seismic imaging datasets gather only short-offset, reflection, data (i.e. see section 2.3) but FWI requires long-offset and transmission data to function efficiently [80]–[83]. The underlying problem is that the reflection terms in the loss function are insensitive to wavelength because sound speed modulates the amplitude but not the phase of reflected waves. Therefore, long-offset and transmission information, which are sensitive to phase, are required for successful inversions.

However, the requirement for long-offset data creates a further difficulty, because wave-propagation over dozens of wavelengths increases the effect of non-linear physics [84] (e.g. multiple scattering) and introduces numerical errors, which are introduced by numerical approximations during simulation. Many different numerical approximations have been tried including finite difference methods [85], finite element methods [86], pseudospectral methods [87], finite volume methods [88], and spectral element methods [89]. A brief introduction is available in Section 2.2, and a full review is available in [90].

The most popular strategy to combat physical non-linearities and numerical errors is the multiscale method [75], which band-limits the wavefield to low wavenumbers by low-pass filtering. The approach stabilises the inversion by reducing the number of wavelengths between the source and receivers and increases the convexity of the objective function.

Another problem is the difficulty matching amplitudes of the predicted and observed data. This occurs for two reasons; first, because it is challenging to calibrate the receiver instruments to match the pressure field from a simulation and, second, because visco-elastic effects attenuate the wavefield and are difficult to model correctly. Instead the wavefield data are normalised, which has proven effective on US-CT data [59].

1.3 Image quality assessment

The substantial potential of US-CT is clear, but this modality is likely to suffer from degraded image quality compared to modern MRI and CT protocols. This is because the physics approximations underlying CT and MRI reconstruction algorithms are well understood in comparison to US-CT, meaning that the most important physical effects are accounted for appropriately. Chapter 5 illustrates the effect of erroneous physics approximations in US-CT. It is possible to realise to potential of US-CT if image quality can be quantified. For example, this knowledge can be used to reduce confounding during automatic (i.e. machine learning) and intuitive (i.e. clinician lead) image analysis; uncertainty estimation will allow clinicians and machines to reliably mitigate artefacts and other image quality issues which can be confounded with disease incidence. However, quantitative IQA provides further motivation for tomographic uncertainty estimation; these estimates will provide a metric which can be used to optimise the system. This is because lower uncertainty will mean better image quality and, therefore, uncertainty can be targeted directly for image quality improvement.

In addition, IQA is a general problem in medical imaging. On calibration, medical imaging equipment undergoes a series of tests to quantify its resolution, contrast to noise ratio, and signal to noise ratio. Largely ignored from this procedure is the fact that these values will vary between patients and acquisition settings. As a result, imaging artefacts are a common problem in clinical settings, and reconstructions are routinely inspected following acquisition to ensure their fidelity [91], [92]. If the scan is severely degraded, the patient will be required to undergo an additional examination. This is problematic because they may be exposed to additional risks from radiation or contrast agents, and they may have

already left the facility, doubling costs because an additional appointment at the clinic is required. These issues could be mitigated by real-time automated IQA, which can evaluate reconstructions as they are acquired and provide clinical feedback while the patient is in position. Real-time automated IQA would enable solutions such as acquiring additional slices during the scan, reducing patient exposure and clinical costs.

Unfortunately, current image quality metrics are unsuitable for these tasks. Most IQA methods used are full reference methods [92]–[96], which compare the reconstruction to a reference image representing the ground truth. In a clinical setting, it is difficult to define a suitable reference image; either the images are of phantoms, which may not correctly reflect images of patients, or the images are of patients, who lack a ground truth measurement. No-reference IQA methods exist [91], [97], but there is no well established technique due to methodological limitations. Statistical methods, which rely on uncertainty estimation, are a potential solution [98]. However, current proposals to reconstruct tomographic uncertainty are computationally expensive or rely on unproven deep learning technologies (see Sections 1.4 and 3.1). Thus, the final motivation to develop a Bayesian model for tomographic imaging is to enable reliable, automatic, and real-time IQA by obtaining accurate uncertainty estimates.

1.4 Uncertainty estimation

Uncertainty estimation is not routine in either Geophysical or medical imaging but, in geophysics, the Monte-Carlo method is relatively well studied [76], [78], [99]. This method generates large ensembles of synthetic images which are used to build a probability mass function for the likelihood distribution and, therefore, provide pixel-wise estimates of the variance. Unfortunately, to build an accurate estimate of the likelihood distribution requires several samples of the likelihood per dimension. Because, in effect, images have one dimension per pixel, the Monte-Carlo method cannot be applied directly to images and correlation functions must be specified to reduce the dimensionality of the image.

Another approach is to calculate the Hessian in full [15], in subsets [100], or by the BFGS algorithm [101]. By the Cramer-Rao lower bound, the inverse of the Hessian can be shown

to be a covariance matrix. Calculating the full Hessian has a similar computational cost to Monte-Carlo, but taking subsets or using the BFGS algorithm are both computationally cheaper. In particular the BFGS algorithm does not add significantly to computation time and can speed convergence, but this comes at the cost of greater demand for memory.

Finally, flavours of variational inference such as using the Stein approximation [102] and automatic differentiation variational inference [103] have been investigated. These flavours of variational inference are also unsuitable for medical imaging because of their computational cost; automatic differentiation variational inference requires the inversion of the Jacobian matrix, which is expensive for FWI, and the Stein approximation is slow to converge.

Clearly, computational cost is a common limitation for uncertainty estimation, which has limited its uptake in Geophysics. Furthermore, all of the traditional Geophysical solutions are impossibly expensive for medical applications, which require reconstructions within minutes.

In medical imaging, early work emphasised the theory of *maximum likelihood estimation* to derive Cramer-Rao bounds [104] or deterministic expectation maximisation algorithms for direct covariance estimation [105]–[107]. The direct expressions for the covariance estimator divide into iterative algorithms [105], where the covariance matrix is updated in conjunction with the mean estimate, or fixed point analysis [106], where the covariance matrix is estimated once the mean has converged. These techniques have several difficulties. First, they involve multiple terms where derivatives of the objective function must be found, which is computationally demanding. Second, for the fixed point analysis, the covariance is estimated assuming the reconstruction has converged, which is often incorrect. Finally, for the iterative algorithm, deterministic expressions are calculated manually after particular distributions have been assumed. These expressions can be difficult to derive and must be revised when the distributions are altered.

These computational difficulties have impeded the development of uncertainty estimators in tomographic reconstruction, but developments in deep learning have renewed interest in tomographic uncertainty estimation. The latest proposals directly use Bayesian deep learning to learn an algorithm which produces a tomographic reconstruction and an corresponding

uncertainty estimate [108]–[110]. The problems with this approach are that the whole tomographic reconstruction algorithm must be replaced with a learnt counterpart, these learnt reconstruction algorithms are unproven in tomography, and Bayesian neural networks have a large memory footprint [111], which will inhibit reconstruction of 3-D images.

Instead of using a Bayesian deep leaning algorithm directly, I will apply stochastic variational inference, the technique underlying Bayesian deep learning, to a conventional tomographic reconstruction algorithm. The assumptions used to construct this scheme are not specific to FWI, so I will call this scheme variational tomography (VT) to indicate that the algorithm can be generalised to other modalities and reconstruction algorithms. Throughout the thesis, when I am referring specifically to a US-CT system with the FWI algorithm, I use the term VT-FWI. Stochastic variational inference has had significant success in deep learning [112]–[114] and, superficially, it is reasonable to apply it to FWI because differential equation solvers (i.e. the solver used by FWI) and convolutional deep neural networks share many properties [115]. Less clear is how to interpret the resulting uncertainty estimates because this is application specific. Thus, I propose to use VT uncertainty for image quality analysis.

In summary, this thesis shows the development and validation of variational tomography; a useful, computationally affordable tomographic uncertainty estimator based on stochastic variational inference. The VT uncertainty estimate is orders of magnitude cheaper than alternatives; it requires less than 1% additional computation time and no additional data compared with a traditional reconstruction algorithm. VT can also be used in conjunction with traditional, proven reconstruction algorithms. One such algorithm is full-waveform inversion, which is used to in ultrasound computed tomography. The adjoint state method is the key technique which makes the FWI inverse problem tractable [15], [71], so it will be essential for the new tomographic uncertainty estimator to integrate this mechanism. It is also important that the estimator provides useful information for clinical practice. To evaluate the utility of the estimates for medical imaging, the estimator is assessed qualitatively and quantitatively in relation to perceived and measured image quality. Comparing the tomographic uncertainty estimates obtained from different uncertainty estimators is not

necessary; the need for low-cost uncertainty estimates is clear, if this estimator has useful properties, then it is less important that it is accurate in comparison to other approaches.

1.5 Thesis structure and contributions

In summary, the objective of this thesis is to develop a low-computational-cost Bayesian framework for FWI. This objective is motivated by 4 benefits of the proposed scheme:

- To incorporate statistical prior information without significant additional computational cost.
- To reduce confounding due to image quality during automatic (i.e. machine learning) and intuitive (i.e. clinician lead) image analysis.
- To use uncertainty estimation as a metric for optimising the acquisition and system settings.
- To enable reliable, automatic, real-time IQA for adaptive reconstruction protocols.

Chapter 2 starts with an introduction to FWI. This is important because understanding the physics of wave propagation and the inversion of wavefield data is important to understand the discussions in the subsequent chapters. The chapter begins by developing the isotropic acoustic wave equation from the principles of stess and strain via the isotropic elastic wave equation. Both these wave equations are in use for Geophysical FWI, but this thesis inverts only the isotropic acoustic wave equation. These wave equations are analytically intractable for inhomogeneous sound speed, so numerical simulators are used to make predictions. Therefore the numerical modelling of wave equations is also introduced. The chapter concludes with an intuitive and visual introduction to FWI.

The objective of the Thesis is met across 3 chapters. In Chapter 3, a Bayesian model is developed for FWI by assuming the sound speed and likelihood distributions follow a mean-field Gaussian distribution. Then, it is demonstrated that this Bayesian model can be solved using stochastic variational inference; by taking derivatives of the likelihood function with respect to the mean and standard deviation of the sound speed, two gradient expressions

can be obtained, one for each parameter. This scheme is called Variational Tomography. Using the adjoint state method on the mean and standard deviation expressions shows that the standard deviation reconstruction requires few additional computations compared to the mean. The Chapter concludes with an *in-silico* example; the two gradient expressions are used to obtain a tomographic mean estimate, which is equivalent to the traditional FWI reconstruction, and a tomographic variance estimate. It is shown that the tomographic variance reconstruction is qualitatively comparable to the absolute error between the ground truth sound speed and the FWI sound speed (i.e. tomographic mean) reconstruction.

FWI is optimised by batch- or minibatch- stochastic gradient descent. This introduces some variability into the reconstruction algorithm. Furthermore, the standard deviation gradient contains an additional stochastic term, which might make the VT reconstruction algorithm unstable. In reality the VT reconstruction algorithm is numerically stable and, when the sound speed reconstruction matches the ground truth, the tomographic variance estimate is also reliable. These findings are demonstrated in Chapter 4 using an *in-vitro* dataset. The use of an *in-vitro* dataset is important because FWI is known to be challenging on real data. These results also show that stochastic perturbations to the sound speed, necessary for VT, do not affect the stability of the FWI reconstruction of experimental data. Chapter 4 also expands on the uses of tomographic uncertainty by presenting a qualitatively relationship between tomographic uncertainty and the image quality of the tomographic mean reconstruction.

Chapter 5 presents the final contribution; evidence of a quantitative relationship between tomographic uncertainty and quantitative IQA. Errors are introduced which violate the assumptions of the experimental configuration or the FWI physics-model and, therefore, these errors reduce image quality. The resulting reconstructions are investigated using tomographic uncertainty and quantitative IQA. The results often demonstrate a linear relationship between the tomographic standard deviation magnitude and the absolute error, with some specific exceptions. By enabling quantitative no-reference IQA *in-vitro* and *in-vivo*, the results suggest that tomographic uncertainty could be used to reduce confounding during image analysis, as a metric to optimise image quality, and for adaptive reconstruction

algorithms. The chapter finishes by discussing the role that implicit prior information will play in the VT framework, this is expected to improve the results and will enable the final motivation; to incorporate statistical prior information and regularise the FWI reconstruction process.

Chapter 6 contains the conclusions of this thesis. First, a Bayesian model can be constructed for tomography which results in a low computational cost uncetainty estimator. Second, the tomographic uncertainty reconstructions are qualitatively related to *in-vitro* image quality. Finally, tomographic uncertainty can be expressed as a new image quality metric called mean standard deviation magnitude, which is quantiatively related to the mean absolute error between the ground truth image and *in-silico* FWI reconstructions.

Chapter 7 focuses on future work and specifically the specification of a prior to constrain the FWI reconstruction process. Data-driven, non-parametric models are studied, with statistical deformation models and variational autoencoders recommended for further study. The prior will be trained using a database of images and details regarding preprocessing an open-source database of CT images are provided.

Chapter 2

Numerical methods and full-waveform inversion

Ultrasound computed tomography, full-waveform inversion, and uncertainty estimation were introduced in the previous chapter, with a view to developing a novel tomographic uncertainty estimator for real-time image quality analysis on FWI reconstructions. FWI relies on a full numerical simulation of the propagation of an ultrasound wavefield in tissue. Therefore, this chapter begins by reviewing the physics of wave propagation for different media. These equations are analytically intractable in all but the simplest of cases [90], so there is a review of numerical methods for solving the acoustic wave equation. Next, a conceptual introduction to FWI is provided, but the derivation of the adjoint state method is included in Section 3.2. Finally, it is important that the uncertainty can be estimated under realistic experimental conditions, so specific details regarding the configuration of the transducer array are discussed.

2.1 The physics of ultrasound propagation

Ultrasound waves propagate when the stresses and strains on a material or medium are locally out of equilibrium. Following the description by Feynman [116], the stress tensor s_{ij}

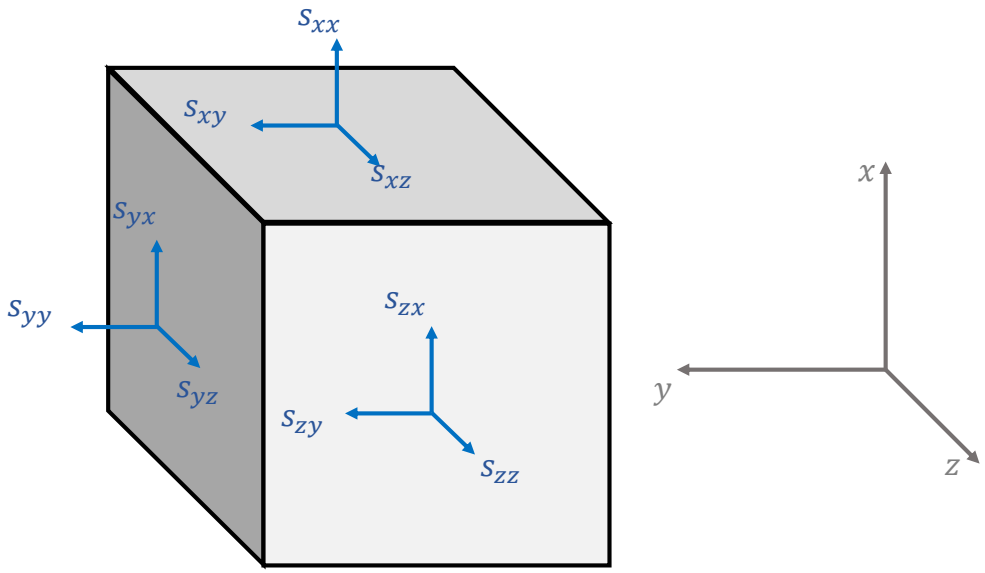


Figure 2.1: The unit cube of the stress tensor s_{ij}. When a force is applied along the axis i, it provokes stresses in all the directions j. These stresses have parallel and perpendicular components.

and the strain tensor e_{kl} are related by the tensor of elasticity c_{ijkl} such that;

$$s_{ij} = \sum_{k,l} c_{ijkl}\, e_{kl} \tag{2.1}$$

In Equation 2.1 the directional components i, j, k, l have 2 or 3 dimensions depending on whether the desired reconstruction is 2-D or 3-D. For example, in 2-D, $i = \{x, y\}$.

In terms of their physical meaning, stress describes the force on the interior of the material per unit area, and strain is the change in the length of the material per unit length. The tensor of elasticity describes the dependencies between the stress and the strain as they are applied in different directions.

Figure 2.1 shows an example of the components of the stress tensor. This illustrates how tensors encode the relationship between the direction of action (the applied force) and the direction of response (the internal forces the action provokes). For example, a force acting in the x-direction causes a compressive stress s_{xx} and two shear stresses s_{xy} and s_{xz}. By definition, compressive stress acts on surfaces facing in the same direction as the applied

force and shear stress acts on surfaces facing in the perpendicular directions.

Thus, the stress tensor has 9 unique components, because the applied force has 3 unique components and the direction of response also has 3 components. Because a stress and strain are described by 9 unique components, the tensor of elasticity has 81 unique components. For example, a stress s_{xx} provokes strains in all 9 components of the strain tensor. It is apparent that Equation 2.1 describes a hugely complex system and that solving for the components of the tensor of elasticity would be ill-posed based on measurements of external stress and strain.

This dilemma is resolved by symmetries in the material. The first assumption is that stress and strain tensors are symmetric, which leaves 36 unique components in the tensor of elasticity. Furthermore, the symmetry $c_{ijkl} = c_{klij}$ of the tensor of elasticity leaves 21 unique components. Finally symmetries are present within the material. For example, isotropic and cubic crystalline materials are symmetric under 90^o rotation, meaning

$$\begin{aligned} c_{xxxx} &= c_{yyyy} = c_{zzzz} \\ c_{xxyy} &= c_{yyzz} = c_{zzxx} = ... \\ c_{xyxy} &= c_{yzyz} = c_{zxzx} = ... \end{aligned} \tag{2.2}$$

These materials are also symmetric under reflection about a plane perpendicular to a chosen axis, meaning the shear components of the stress and strain tensors must be aligned, or their components must be zero. For example, $s_{x,y} \neq s_{x,-y}$ unless $s_{x,y} \neq s_{x,-y} = 0$, or when the strain tensor aligns $s_{x,y}\, c_{x,y,x,y}\, e_{x,y} = s_{x,-y}\, c_{x,-y,x,-y}\, e_{x,-y}$. Thus,

$$c_{xxxy} = c_{yyyz} = c_{xyxx} = ... = 0 \tag{2.3}$$

which means only 3 components of the tensor of elasticity are non-zero, as defined in Equation 2.2.

Finally, isotropic crystals are symmetric under any rotation (rather than simply 90^o rotation), which means the stress and strain must be linearly related (i.e. related by scalar quantities). There are two parameters, called the Lamé parameters, which meet this cri-

terium such that,

$$s_{ij} = 2\mu\, e_{ij} + \lambda \left(\sum_k e_{kk} \right) \delta_{ij} \tag{2.4}$$

This is a form of Hooke's law. The first Lamé parameter λ scales the trace of the strain matrix, before being diagonalised. Therefore, this parameter scales the compressive strain, while remaining symmetric regardless of the direction of application of the force. The second Lamé parameter μ is a scalar which relates each element of the stress matrix e_{ij} to the corresponding element in the strain matrix s_{ij}. By convention, μ is scaled by two so it is directly equivalent to the shear modulus.

By definition, the strain is the derivative of the displacement u, which means,

$$s_{ij} = 2\mu \frac{1}{2} \left(\partial_j u_i + \partial_i u_j \right) + \lambda \left(\sum_k \frac{1}{2} \left(\partial_m u_l + \partial_l u_m \right)_{kk} \right) \delta_{ij} \tag{2.5}$$

Using linear algebra notation for simplicity and the relationship,

$$F = \nabla \cdot S \tag{2.6}$$

where F is an external force and S is the strain in matrix notation, then Equation 2.5 can be written,

$$\begin{aligned} F &= (\lambda + \mu) \nabla \left(\nabla \cdot u \right) + \mu \nabla \cdot \nabla u \\ F &= (\lambda + \mu) \nabla \left(\nabla \cdot u \right) + \mu \nabla^2 u \end{aligned} \tag{2.7}$$

which is the isotropic elastic wave equation.

Additionally, noting Newton's law such that,

$$F = \rho \frac{\partial^2}{\partial t^2} u \tag{2.8}$$

where ρ is the density, then,

$$\rho \frac{\partial^2}{\partial t^2} u = (\lambda + \mu) \nabla (\nabla \cdot u) + \mu \nabla^2 u \tag{2.9}$$

This approximation can be further reduced if shear forces are assumed to be negligible; in other words, if the medium can only support compressive forces, then only the components $c_{xxxx} = c_{yyyy} = c_{zzzz}$ are non-zero. In this case,

$$\begin{aligned} \rho \frac{\partial^2}{\partial t^2} u &= K \nabla^2 u \\ c^{-2} \frac{\partial^2}{\partial t^2} u &= \nabla^2 u \end{aligned} \tag{2.10}$$

which is the isotropic acoustic wave equation. In Equation 2.10, c is the sound speed, ρ is the density (assumed to be constant), $K = \lambda + \frac{2}{3}\mu$ is the bulk modulus, and u is the displacement. Note that pressure p is correlated 90^o out of phase with displacement [116], so Equation 2.10 can also be written,

$$c^{-2} \frac{\partial^2}{\partial t^2} p = \nabla^2 p \tag{2.11}$$

Figure 2.2 shows a physical description of the isotropic modes of wave propagation. This illustrates that compressive forces cause the unit cube to bulge in directions aligned with the surface normals, whereas shear forces causes the unit cube to distort in directions which are perpendicular to the surface normals.

This thesis usually assumes isotropic, acoustic media. Therefore, Equation 2.11 is used. In medical imaging, this assumption is valid for soft tissues, which struggle to support long distance shear modes. It remains unclear how suitable it will be in the skull because, strictly speaking, bone is crystalline and elastic. However, research from focused ultrasound therapy suggests that the acoustic approximation is suitable for waves at near-perpendicular incident angle [117], [118]. Furthermore, if the isotropic assumption fails, then FWI could be performed without it. Anisotropic schemes are used in Geophysics [119], but they are more computationally expensive.

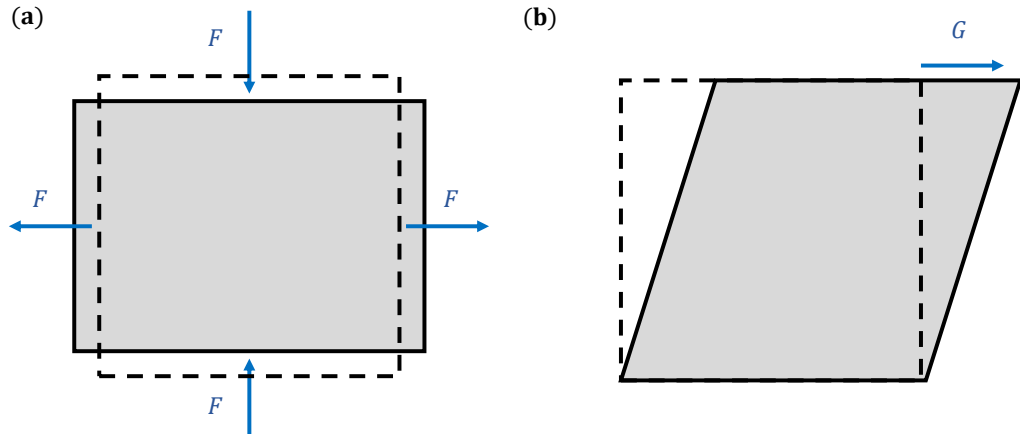


Figure 2.2: The isotropic modes of propagation for the unit cube. (a) shows that strain F induces a compression mode which is supported by elastic and fluid materials, (b) shows that strain G induces a shear mode which is only present in elastic materials.

2.2 Numerical modelling of wave propagation

Geophysicists have been studying methods of wave-propagation for decades and a wide variety of approaches exist, each with an extensive background literature [90]. Here, I will outline some common approaches, which are distinguished on the basis of the grid geometry and the numerical approximation of the derivative. I will conclude by discussing the finite-difference method in detail.

The first distinction is between methods which use structured or unstructured grids. Fundamentally, these approaches differ according to the connectivity between adjacent cells. Connectivity is directly related to the distance between the centre-points of each cell. For example, finite difference methods use grids with regular connectivity. This means each grid element is similar (i.e. they can be transformed by scaling only) and the grid can be represented by a vector (1-D), matrix (2-D), or tensor (n-D). The most common form is the regular equispaced grid in Cartesian geometry, which corresponds directly to the pixel representation used in digital imaging.

In contrast, finite element methods have irregular connectivity. In practice, this means

each element is dissimilar but isotropic; the elements come from a consistent class of shapes, for example a tetrahedral mesh. The irregular connectivity means the grid has greater memory requirements, because the connectivity matrix must also be sorted, but also that tissue boundaries can be correctly represented. This is a key consideration when the boundary is sharp or when there are significant changes in contrast within the image (e.g. the skull). Unfortunately, the grid also requires prior knowledge about the position of the boundaries, which can be difficult to obtain in practice.

It is clear that a regular grid is preferable for medical FWI. First, this corresponds closely to existing methods of digital image presentation (e.g. CT or MRI). Second, the grid can be implemented without prior anatomical knowledge, which would defeat the purpose of tomographic imaging. The key limitation is the specification of the skull-tissue boundaries. Fortunately, an *in-vitro* demonstration using a high contrast, dental resin phantom suggests that FWI of the head is possible on a structured grid [35].

The second distinction is between a grid point or series expansion approach for the spatial derivative. The archetypal series expansion approach is the pseudospectral method. In this approach, the wavefield is decomposed using the Fourier series. This is particularly desirable because an arbitrary function can be represented exactly using a finite set of bases. Given the structure of the wavefield, it appears that this is ideal for FWI and that the method is capable of producing low memory representations of the wavefield. The approximation also leads to a compact, fast, and precise representation of the derivative;

$$\partial_x^{(n)} u(x,t) = \mathcal{F}^{-1}\left[(ik)^n U(k,t)\right] \tag{2.12}$$

where $U(k,t)$ is the Fourier representation of the wavefield $u(x,t)$. The derivative expression takes this form because the n^{th} order derivative operator $\partial_x^{(n)}$ can be exchanged with the Fourier operator $\mathcal{F}(\cdot)$ to act on the term $\exp(ikx)$, which is contained in $U(k,t)$. Thus, the inverse Fourier operator $\mathcal{F}^{-1}(\cdot)$ is sufficient to recover the derivative in the x-domain. This scheme still performs time-stepping because a finite-difference method is still used to approximate the time derivative.

The key limitation of the pseudo-spectral method is boundaries, which are computa-

tionally demanding because they require a large number of basis functions [90]. This is significant for "free surface" boundaries, where there is effectively infinite impedance. Examples of free-surface boundaries in medical imaging might include air-tissue boundaries in the nasal cavity, and transducer-water boundaries if the system needs to be included in the model.

An alternative is the finite difference scheme. Finite difference schemes are well established in Geophysics, have been used for US-CT of the breast [5], and have also been used to model wavefield propagation in the head for focused ultrasound [120]. The finite difference scheme uses a numerical approximation which is composed of function evaluations at discrete steps along the axis. The simplest example uses the geometric definition of the gradient,

$$\begin{aligned} \frac{\partial f(x)}{\partial x} &= D_x\, f(x) \\ &\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \end{aligned} \tag{2.13}$$

where D_x is the geometric derivative, f(x) is the function which is evaluated at point x, and h is a discrete distance in the x direction. Removing the limit $h \to 0$ yields the forward finite difference approximation.

Higher order derivatives are achieved by Taylor expansion [121],

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + ... \tag{2.14}$$

For the wave-equation, second order derivatives are required. The numerical order of the stencil must also be specified. For example, the 2^{nd} order approximation requires 5 function evaluations,

$$\begin{aligned} D_x^2\, f(x) = \alpha_{(-2)} f(x-2h) + \alpha_{(-1)} f(x-h) + \alpha_{(0)} f(x) \\ + \alpha_{(1)} f(x+h) + \alpha_{(2)} f(x+2h) \end{aligned} \tag{2.15}$$

where $\alpha_{(n)}$ is the coefficient of the nth function evaluation at the point $x + nh$. Taking the Taylor expansion of Equation 2.15 and rearranging,

$$\begin{aligned} D_x^2 f(x) = &(\alpha_{(-2)} + \alpha_{(-1)} + \alpha_{(0)} + \alpha_{(1)} + \alpha_{(2)}) f(x) \\ &+ (2\alpha_{(-2)} + \alpha_{(-1)} - \alpha_{(1)} - 2\alpha_{(2)}) h f'(x) \\ &+ (4\alpha_{(-2)} + \alpha_{(-1)} + \alpha_{(1)} + 4\alpha_{(2)}) h^2/2! f''(x) \\ &+ O(h^3) \end{aligned} \tag{2.16}$$

Solving for f(x) = 0, f'(x) = 0, and f"(x) = 1, gives the fourth order approximation to the Laplacian. The coefficients are,

$$h = \frac{1}{12},\ \alpha_{(-2)} = -1,\ \alpha_{(-1)} = 16,\ \alpha_{(0)} = -30,\ \alpha_{(1)} = 16,\ \alpha_{(2)} = -1 \tag{2.17}$$

Repeating this process for arbitrary length approximations gives approximations with different numerical order. Using the numerical approximations, the time-order 2, space-order 4 (OT2-OS4) 2-D acoustic wave equation takes the form,

$$\begin{aligned} u(t+dt,x,y) = &-u(t-dt,x,y) + 2u(t,x,y) + \frac{dt^2}{12dx^2} c^{-2}[-60u(t,x,y) \\ &- u(t,x-2dx,y) + 16u(t,x-dx,y) + 16u(t,x-dx,y) \\ &- u(t,x-2dx,y) - u(t,x,y-2dy) + 16u(t,x,y-dy) \\ &+ 16u(t,x,y-dy) - u(t,x,y-2dy)] \end{aligned} \tag{2.18}$$

where dt is the time step (i.e. the discrete distance in the time domain) and $dx = dy$ is the grid spacing. This is a time-stepping scheme because the current and previous wavefields are used to calculate the next wavefield.

These coefficients have been obtained by Taylor expansion, but it is possible to find coefficients which provide a more accurate numerical approximation of the derivative. Such schemes optimise the coefficients of the finite difference stencil so that they are optimal

for simulating the propagation of a particular band of wavenumbers. A set of optimised coefficients for the Laplacian were used to generate predicted wavefields (e.g. [122]–[125]), although I am unable to disclose the values in full due to commercial confidentiality. Note that if different coefficients are used in the Laplacian stencil, then the results in this thesis are reproducible by adapting the spatial and time discretisation.

Numerical stability and dispersion are important considerations when developing numerical simulators. A simulation is numerically stable if the Courant-Friedrichs-Lewy (CFL) condition is met. Mathematically, the CFL condition is [90],

$$c\frac{dt}{dx} \leqslant 1 \tag{2.19}$$

where c is the magnitude of the sound speed in the model. If this condition is not met, then the solution to the finite-difference equation will be unstable; its solutions will increase until they are too large to be represented by the numerical register. The practical consequence of the CFL condition is to limit the range of values that the time discretisation dt and space discretisation dx can take.

Simulations are less sensitive to numerical dispersion, but it is impossible to eliminate numerical dispersion from finite-difference simulations with heterogeneous sound speeds. Numerical dispersion occurs because the phase velocity of the simulated wave is dependent on the wavenumber [90] and, therefore, the numerically simulated wavefield will suffer from non-physical dispersion. This means that some frequencies of the numerical simulated wavefield will arrive early and others will arrive late. A typical relationship is that low-frequency components of the wavefield will travel slower than the physical reality. The effect can be compensated for by adjusting the time discretisation dt to increase the number of grid points per wavelength.

2.3 Full-waveform inversion

An FWI experiment starts with measurements of acoustic pressure or displacement, taken using receiver sensors. An example of these measurements can be seen in Figure 2.3(e), where

the measurements are taken from a simulation which uses a 2-D slice of the MIDA head phantom [126]. MIDA is a numerical model of the human head, derived from MRI imaging of a 29-year-old female volunteer. The phantom consists of 153 segmented structures, which were derived from by T1, T2, magnetic resonance angiography (MRA) and diffusion tensor imaging (DTI) sequences. Together, the 153 structures cover the entire head and can be adapted to acoustic, electric, or other units by associating each class to values from the IT'IS database of human tissue properties [29].

In medical imaging, the receiver sensors are ultrasound transducers which measure the displacement of the surface of a piezoelectric crystal using voltage. The surface displacement is caused by pressure changes on the surface of the piezoelectric. Thus, the receivers measure the pressure changes which are produced by an incident acoustic wavefield. The source of the acoustic wavefield is a different transducer, which applies an external force to a medium and causes an the wavefield to propagate outwards. Information is contained in the pressure measurements made by the receivers because the acoustic wavefield is modulated by the structures between the source and the receivers.

Figure 2.3 shows the relationship between (a-d) the wavefield and (e) the wavefield data. The receivers are on the black oval surrounding the head phantom and the receivers can be described by their azimuthal angle relative to the source transducer. In Figure 2.3(a) the source applies an external force to the water and a wavefield is produced. Figure 2.3(e, red arrow) shows that the receivers on an acute azimuthal angle (receiver ID's 250-350) measure the direct water wave (i.e. the sound wave that travels only in water). Figure 2.3(b) shows that the outgoing wavefield reflects off the skin and skull. Figure 2.3(e, green arrow) shows that the same acute azimuthal angle receivers (receiver ID's 250-350) measure pressure of this reflected component of the wavefield. Some of the energy of the outgoing wavefield propagates through the skull, arriving at the receivers with oblique, near-perpendicular azimuthal angles. Figure 2.3(c) shows the arrival of the wavefield at the receivers with oblique, near-perpendicular azimuthal angles. This forms the component of the wavefield data indicated in Figure 2.3(e, blue arrow) for receiver ID's 125-215. Finally, Figure 2.3(d) shows the arrival of the wavefield at the receivers with oblique, near 180^o azimuthal angle.

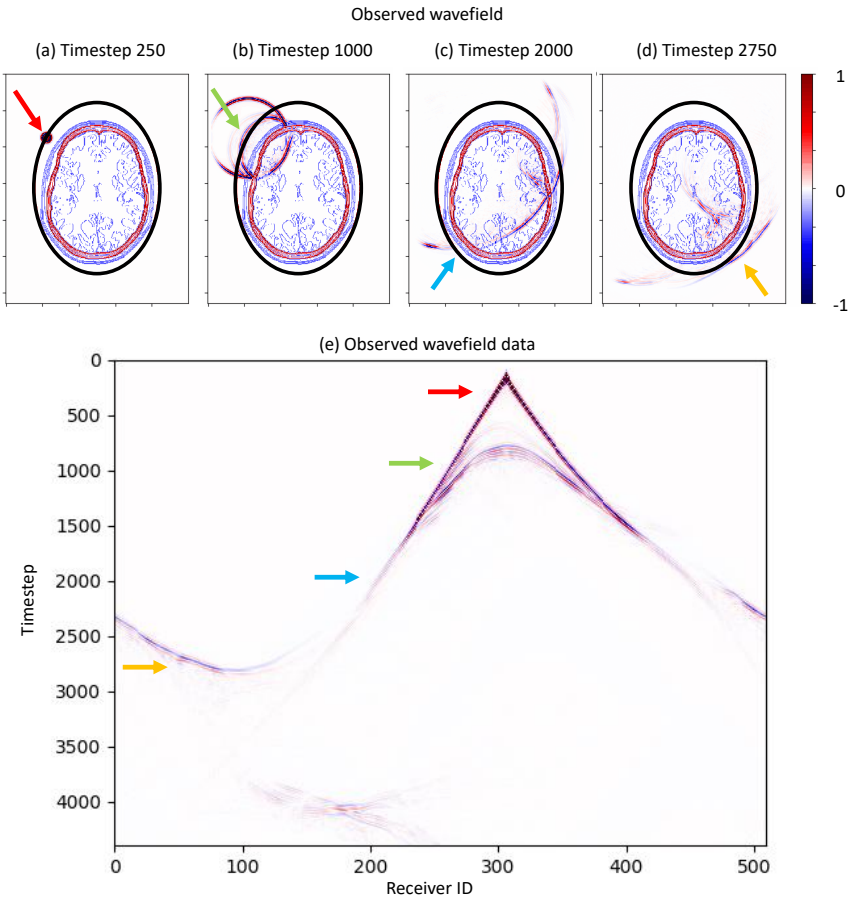


Figure 2.3: (a-d) The simulated wavefield, shown for different timesteps, surrounded by a ring of receivers on the black oval. (e) Wavefield data are obtained when the displacement or pressure is measured at the receivers. The arrows indicate common features of the wavefield and wavefield data

The first arrivals at these receivers are direct transmissions through the head indicated in Figure 2.3(e, orange arrow) for receiver ID's 25-100.

The aim of the FWI algorithm is to exactly match some numerically simulated wavefield data to the observed wavefield data, where the numerically simulated data can be changed by changing the parameters of the simulation (e.g. sound speed). The numerically simulated wavefield data is called the predicted data. Figure 2.4(a) shows some observed wavefield data and Figure 2.4(b) shows some predicted wavefield data for medium with a uniform sound speed of $1480\,[m\,s^{-1}]$, which is approximately the sound speed of water [29]. The numerical simulator used to generate this predicted data used a finite-difference time-domain

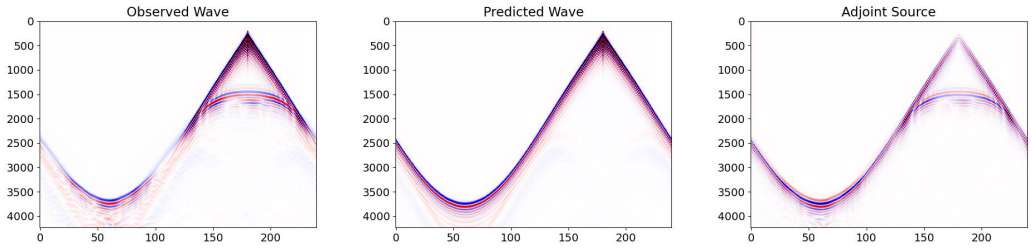


Figure 2.4: (a) A simulated example of observed data using a 2-D slice of the MIDA head phantom [126]. (b) Predicted data is generated assuming a constant sound speed of 1480$[m\,s^{-1}]$, and (c) the L2 loss, showing the difference between the observed and predicted data

approximation of the acoustic wave-equation, for example see Equation 2.18. In Figure 2.4, the 2^{nd} order time- and 10^{th} order space-approximation was used, and the Laplacian used optimised coefficients. Note that the wavefield data has been low-pass filtered to prevent cycle skipping [75].

By comparing the observed wavefield data to the predicted wavefield data, the model parameters can be adjusted to ensure a close match between them. The predicted data is compared to the observed data using a loss. This thesis adopts the L2-norm for the loss function between the predicted and observed data. Mathematically, this is defined,

$$\mathcal{L}(c) = -\frac{1}{2}[d - d_p(c)]^T\,[d - d_p(c)] \tag{2.20}$$

where d is the observed data from the receivers and d_p is the predicted data. In this thesis, d_p is dependent only on the acoustic sound speed c. Figure 2.4(c) shows how the predicted and observed data are compared using the sum of squares loss function.

It is difficult to match the amplitude of the predicted and observed data in the case when the observed data is obtained experimentally. This is because it is difficult to assign a pressure to an observed voltage; the observed voltage is modulated by piezoelectric response, amplifiers, and internal electrical impedance. In reality, the wavefield data are normalised. This thesis uses source normalisation, which means that the amplitude of the predicted and observed data for each source is scaled into the same range. Mathematically,

$$d^{norm}(x_r, t; x_s) = \frac{\sum_{x_r,t} |d(x_r, t; x_s)|}{\sum_{x_r,t} |d_p(x_r, t; x_s)|} d(x_r, t; x_s) \tag{2.21}$$

where the predicted data d_p observed data d are dependent on the sources x_s, receivers x_r, and time t. The result is d^{norm}; the observed data standardised into the same range of amplitudes as the predicted data. An alternative approach is trace normalisation, which means that the amplitude of the predicted and observed data for each receiver is scaled into the same range. Trace normalisation effectively removes amplitude information and was found to reduce the quality of the reconstructions *in-silico*. Figure 2.4(c) shows the adjoint source after source normalisation.

Thus, the FWI algorithm has 4 inputs. First, a numerical approximation of a wave equation. For example, this thesis assumes the constant density acoustic wave equation defined in Equation 2.18. Second, some observed data. As discussed previously, in medical ultrasound this is the radio-frequency data, which is the measurement of displacement or pressure on the surface of the receiving transducers. Third, the displacement or pressure of the emitting or source transducer, also called the source wavelet. Fourth, the starting model, which is a digital (i.e. discretised) image or volume containing prior sound speed information.

FWI is an adjoint method. Adjoint methods describe a class of solutions to non-linear inverse problems. These solutions often involve a non-linear forward problem, the linearisation of the inverse problem, and an iterative gradient-based solver to converge on the true non-linear solution. One example is back-propagation in neural networks, which calculates the gradient of the loss with respect to each weight of the network by cascading the loss through each layer using the chain rule. Similarly, FWI linearises the inverse problem by ignoring high-order scattering in the forward problem, which makes the non-linear operator invertible. Section 3.3.3 contains a full derivation of the FWI adjoint method for the d'Alembert operation (Equation 3.1), which is equivalent to the gradient with respect to the mean of a Gaussian probability distribution. Below is an outline of these equations and their practical meaning.

In the case of the acoustic wave equation, the FWI adjoint method is,

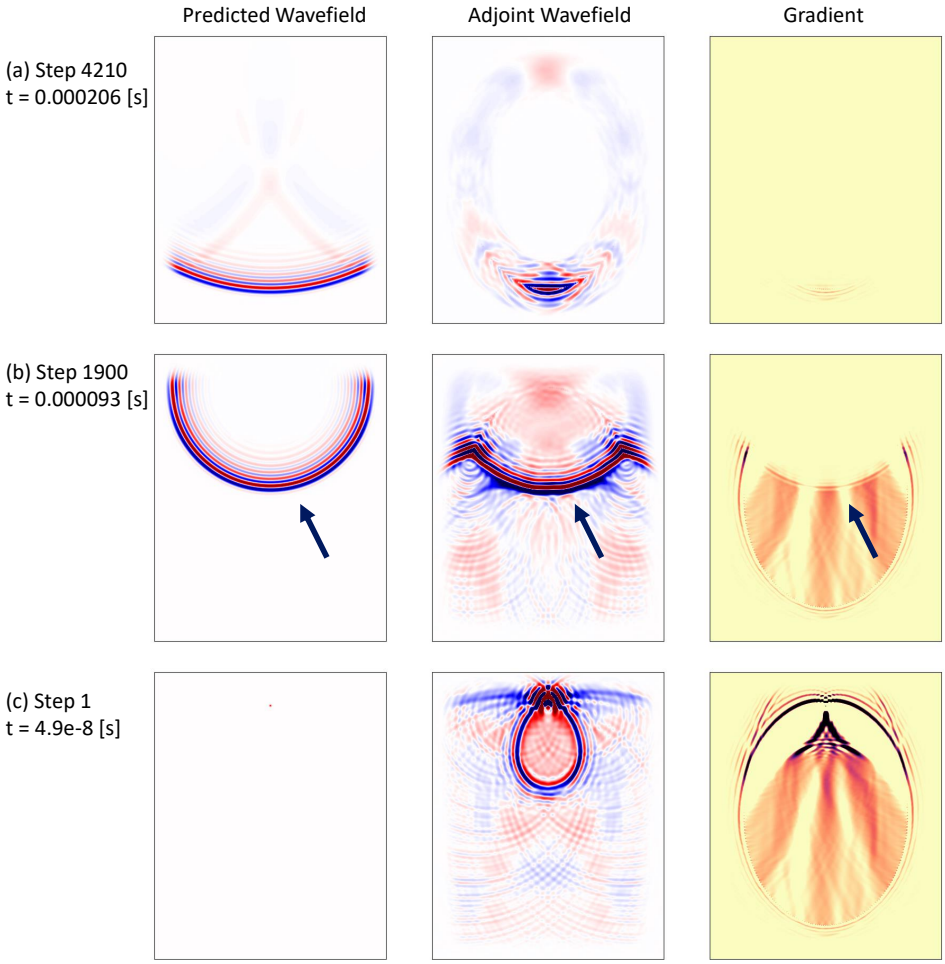


Figure 2.5: The adjoint method produces an estimate of the gradient by correlating the predicted and adjoint wavefield. The black arrows show where the predicted and adjoint wavefield have the strongest correlation.

$$\frac{\partial}{\partial c}\mathcal{L}(c) = u_p(c)^T \frac{\partial H(c)}{\partial c} u_{adj}(c) \tag{2.22}$$

where $u_p(c)^T$ is the predicted wavefield, which is calculated numerically using Eq 2.18 and stored in memory, $\frac{\partial H(c)}{\partial c}$ is the derivative of the d'Alembert operator with respect to the sound speed, and $u_{adj}(c)$ is the adjoint wavefield. See Equation 3.54 in the full derivation.

The adjoint wavefield is a new wavefield which is propagating back from the receivers to the source. It is non-physical (it does not exist in reality) but it can be derived mathematically, see Equation 3.51 in the full derivation. The equation for the adjoint wavefield

is,

$$u_{adj}(c) = \left[H(c)^{-1}\right]^T \varsigma_{adj} \tag{2.23}$$

where $H(c)$ is the acoustic wave equation operator and ς_{adj} is the adjoint source. The adjoint source ς_{adj} is calculated using the loss shown in Equation 2.20 and shown in Figure 2.4(c).

Equation 2.22 has several interpretations [71], [127]. Figure 2.5 shows an interpretation for the explicit finite-difference scheme in Equation 2.18. In this case a Hadamard product between the predicted and adjoint wavefield vectors is taken at each time-step. This produces a matrix which has $(n \times m)$ entries, where n is the number of model parameters and m is the number of time-steps. In practice, as the adjoint wavefield propagates from the receiver to the source, the multiplication of the two wavefields adds features to the gradient, which is indicated by the arrows in Figure 2.5(b). By integrating over the time-steps, the gradient for a single source is obtained, which is shown in Figure 2.5(c).

2.4 A PyTorch inversion algorithm

Existing FWI libraries involve a trade-off between computational performance and usability during development. For example, some libraries [128], [129] use high-level programming languages, which ease development at the expense of computation time, while others [130] can be used on supercomputers (via the message passing interface) at the expense of development time. GPU computing offers a solution to this trade-off; GPUs have high computational performance for numerical simulation, due to their parallel computing capability, and have been integrated with high-level programming languages. High-level programming languages make the development and prototyping stage simple because they abstract the parts of the code (e.g. multithreading, memory management) which are required to control the parallel computing system.

The development of VT and the inversion of real datasets requires control over simulation inputs, gradient descent parameters, the wavefield data, and the loss function. This degree of control was achieved without a significant trade-off in computation time because a finite-

difference, time domain FWI solver was developed using the PyTorch library [131]. PyTorch is a tensor algebra library written for Python, a high-level programming language. This means the FWI solver was simple to implement, due to the simple Python interface, and PyTorch provided access to GPU acceleration for the requisite mathematical operations (e.g. addition, multiplication, convolution). The PyTorch library was developed for deep learning, so this FWI solver has the added benefit that it can be integrated directly with machine learning architectures. This feature will be useful during the development of a data-driven prior (see Chapter 7).

The theory of numerical finite-difference methods is available in Section 2.2. In practice, numerical simulators are plagued by problems with numerical stability and dispersion. In FDTD implementations the derivative is approximated with an arbitrary number of polynomials, obtained by a Taylor expansion. The truncation of this series leads to approximation errors, but using additional terms becomes increasingly expensive with limited improvement in the approximation. In Chapter 3, the inverse crime means that the observed and predicted data suffer from the same effects of numerical dispersion. This means that, provided the solver is within the CFL limit [90], numerical dispersion does not affect the inversion. However, numerical dispersion will affect the inversion of the *in-vitro* data. Several finite-difference approximations were tried for the Laplacian, specifically Taylor expansions using the (OT: time order, OS: space order) OT2-OS4, OT2-OS10, OT2-OS20, and OT4-OS20 approximations. Optimised coefficients (e.g. see [125]) for the OT2-OS10 approximation were also tested. The best approximation was identified by comparing the water-shot with the simulation result, which showed that the OT2-OS10 approximation with optimised coefficients performed the best.

As discussed in Section 2.3, amplitudes are difficult to match in FWI. This is because of 1) the complexity of the displacement of the piezoelectric, which depends on the incident angle, pressure, and the relationship between displacement and voltage, and 2) the inability to fully account for visco-elastic effects. The PyTorch inversion algorithm adopts source normalisation because it retains relative amplitudes between receivers of each source, although differences in amplitude between sources are lost. Trace normalisation, which is aggressive

because it effectively leaves only the phase information, was found to degrade the quality of the results. However, trace normalisation can be useful if strong amplitude perturbations are present from density and absorption.

2.5 Transducer configuration

Recently FWI has been applied to medical imaging, and shows great potential to enhance the resolution and contrast of US-CT data [3]–[10], [13]. Despite the potential promised by US-CT image reconstruction, there is not yet a consensus on the optimal configuration of the transducer aperture.

The most common transducer configuration is a 2-D ring array which can be rastered vertically, this had been demonstrated *in-vivo* for breast imaging [57], [62], *in-silico* for bone imaging [10], and *in-vitro* for brain imaging [35]. This configuration is popular because it offers full transmission information, which is important for sound speed imaging [15]. The configuration also reduces the financial cost in comparison to 3-D apertures, which typically require far more transducers to prevent sparsity errors [69].

There is evidence that full 3-D apertures might provide improvements. The simplest is a hemispherical configuration [13], [69], but this requires numerous transducers to avoid problems with sparsity (increasing financial costs) and modelling ultrasound wave propagation in 3-D increases computation time, particularly for FWI [13]. An alternative is to configure two transducers directly opposite each-other. Given a fixed number of transducers, this increases the density of the receive array and reflectors can be used to obtain long-offset information [52]. This configuration also permits 3-D acquisitions without excessive financial costs [9]. However, for FWI, the computational cost of this configuration is unchanged.

The objective of variational tomography is to provide a Bayesian model with low computational cost, so that uncertainty estimation is accessible for 3-D configurations during the FWI reconstruction process. However, the computational cost of 3-D FWI means developing VT-FWI using 3-D models is impractical. Instead, the *in-silico* examples use a 2-D ring configuration (Chapters 3 & 5), which provides full coverage of the sample and it is a realistic configuration *in-vivo* [57], [62]. The density of the transducers is realistic, but

point transducers are assumed which may be inappropriate for *in-vivo* acquisitions [132]. For the *in-vitro* example (Chapter 4), two transducers will be placed in a transmission-style configuration. This reduces the financial cost of the design, increases the transducer density, and is a realistic configuration in practice [9], [52].

Chapter 3

Variational tomography: theory and practice

Chapter 1 introduced the background research which is relevant to this thesis, while Chapter 2 developed the physical and numerical methods which are used for the US-CT FWI reconstructions. The latter chapter also includes a visual, conceptual introduction to the FWI algorithm. Chapter 3 is the first novel contribution. Primarily, this is a Bayesian, *maximum a-posteriori* model with an additional uncertainty estimator with no additional computational cost. It shows that the uncertainty estimator is derived using appropriate assumptions (i.e. the mean-field Gaussian distribution) and stochastic variational inference.

The structure of this chapter is as follows. First, the introduction presents alternative approaches to the uncertainty estimation problem in tomographic imaging. Second, in order that variational tomography is understandable, context is provided by a description of *maximum likelihood estimation* and stochastic variational inference. The main contribution of this chapter (i.e. the Bayesian model) is presented in the method, where variational tomography is applied to FWI. Subsequently, *in-silico* results pertaining to neuroimaging are presented in order to demonstrate some properties of the estimators, and these results are discussed.

I have presented the equations resulting from the theory from this chapter at the Inter-

national Ultrasonics Symposium.

O.Bates *et al* (2021) No-cost uncertainty estimation for ultrasound tomography of the human brain. IEEE IUS, Virtual, 2021.

I have also included the theoretical description and methods in the first part of a research paper in Inverse Problems.

O.Bates *et al* (2022) A probabilistic approach to tomography and adjoint state methods. Inverse Problems, vol.38, 045008.

This work lead to a patent.

O.Bates (2021), Patent, *GB.2104792*. A method for uncertainty estimation in medical imaging.

3.1 Introduction

Tomographic reconstruction algorithms rarely provide uncertainty estimates because pixel-wise uncertainty estimation is computationally expensive. Iterative reconstruction is a commonly used tomographic reconstruction algorithm and has its statistical basis in *maximum likelihood estimation* (MLE), Section 3.2.1 demonstrates MLE in the context of FWI. Fessler [106] and Fessler and Hero [104] use the MLE framework to show that covariance matrix of the reconstruction can be calculated by taking the inverse of the Hessian of the likelihood distribution. Unfortunately, the Hessian is computationally intractable for realistic problems and particularly for 3-D tomography [98].

MLE is a Frequentist statistical technique. Bayesian methods offer numerous practical advantages and are a popular alternative [133]. Bayesian methods which neglect uncertainty estimation are called *maximum a posteriori* (MAP) estimation. The main difference

between MLE and MAP estimation is the use of prior information in the form of a regulariser. The typical regulariser is an additional term in the objective function which penalises reconstructions that do not have favourable properties, this is also called a weak constraint. The simplest example is Tikhonov regularisation, which uses the L2-norm of the parameter vector to penalise the objective. Further details about regularisation can be found in [133].

This paper favours Bayesian inference because it combines regularisation with uncertainty estimation. Bayesian inference is less common than MAP in tomographic reconstruction. One approach to Bayesian inference is to replace the iterative reconstruction with a neural network: the data-to-model approach. Barbano *et al* [108] use deep gradient unrolling, a method which replaces the gradient descent process with a neural network. The network is broken into blocks, where each block corresponds to a single iteration of gradient descent. A block takes a predicted reconstruction and the gradient of the likelihood function as inputs, and uses the neural network to produce the predicted reconstruction for the next block. The whole neural network is trained using a likelihood function, which is the L2-norm of the difference between the predicted reconstruction (the output of the final block) and an iterative reconstruction. Uncertainty estimates are obtained by including a 'Bayesian convolutional layer' (a layer where the weights are stochastic) at the end of each block. Because these layers are stochastic, they will produce a range of possible reconstructions from the deterministic inputs. An alternative approach by Tonolini *et al* [109] could be applicable to medical tomography. A conditional variational auto-encoder is trained to perform reconstructions from corrupted image data. The network is conditioned on the imaging device so that the reconstruction can generalise across different acquisition devices. Variational auto-encoders have a stochastic latent space, so sampling the latent space produces a range of possible reconstructions.

Alternatively, deep-image-prior (DIP) methods use an untrained convolutional network to regularise an iterative reconstruction, thereby using the structure of the convolutional network to favour reconstructions which are visually plausible. Most DIP approaches use MAP estimation. However, Narnhofer *et al* [110] use a Bayesian neural network as a DIP. They ensemble several iterative reconstruction processes, each with a different starting point.

The starting points are generated by taking independent samples from the DIP generator, meaning that pixel-wise mean and uncertainty estimates can be determined from the mean and variance of the numerous final reconstructions. There is a compromise inherent in this proposal; to determine the mean and uncertainty accurately a large number of reconstructions are required, but a large number of reconstructions mean that computation time will increase linearly because each reconstruction process is independent. Thus the ensembling approach becomes computationally expensive when a physically expressive algorithm such as FWI is used.

Neural-network-based methods can produce excellent reconstructions and they have fast uncertainty estimation, but such data-to-model approaches are sensitive to small changes in the data [134]. Further issues with the data-to-model approach include the large memory footprint of neural networks (which could exclude 3-D tomographic reconstruction), a need for large paired training datasets, and poor generalisation across imaging devices. Most concerning is that models designed around neural networks can hallucinate features into the reconstruction [135]. Hallucination seems to be most significant for the DIP approach, suggesting that training can mitigate this problem.

Finally, Monte-Carlo algorithms are a popular class of methods to solve Bayesian inference. Unfortunately, Monte-Carlo has a comparable computational cost to Hessian-based approaches and, therefore, it can be used in Geophysical tomography [78], [99] but not medical tomography, which has stricter time constraints. A potentially cheaper technique is the stochastic gradient Langevin dynamics algorithm, which has been applied to PET reconstruction [136]. Stochastic gradient Langevin dynamics couples stochastic gradient descent to a normally distributed Monte-Carlo sampler. The step-size of the stochastic gradient is reduced as the iterations progress so that the Monte-Carlo sampling term dominates the final set of reconstructions. The mean and uncertainty are calculated from this final set of reconstructions. Unfortunately this algorithm is challenging to implement because it is difficult to specify an appropriate step-size schedule. Thus, it is practically difficult to tune the step-size schedule and the optimal schedule is believed to have slower asymptotic convergence than some MCMC solvers [137].

In contrast, this chapter applies stochastic variational inference directly to tomographic reconstruction algorithm, a scheme called variational tomography (VT). Of particular interest is FWI due to its significant potential in medical imaging (see Chapter 1 for details), so VT is applied to FWI as an illustrative example. VT-FWI consists of Bayesian model that produces a MAP estimator, which is equivalent to the standard FWI estimator, and an additional uncertainty estimate, which can be calculated with little additional computational cost. More sophisticated approximations of the posterior are possible, albeit with greater computational cost. Thus, VT-FWI approach produces two images, an image reconstruction that is comparable to the standard FWI approach and a variance reconstruction which provides a pixel-wise estimate of the variance. The VT approach that is described is not limited to FWI and can be applied to other tomographic reconstruction algorithms or adjoint methods.

In the Theory (Section 3.2), an introduction to Frequentist and Bayesian methods is presented in the context of FWI. Initially, this focuses on the theory of the Frequentist *maximum likelihood estimation* approach. This is the traditional, statistical basis for iterative tomographic reconstruction. It is emphasised that minimising a data-fit function with the form of an L2-norm is statistically equivalent to maximising a likelihood function with a Gaussian form. Then, the Bayesian stochastic variational inference method is described in detail, with emphasis on the choice between the score-function or the reparameterisation approach. Finally, in the Method (Section 3.3), the reparameterisation approach is applied in FWI using a mean-field Gaussian approximation to describe the pixel-wise distribution of the reconstructed image. The application of mean-field SVI to FWI produces an update for the mean which is equivalent to the standard FWI estimator, and an update for the variance which provides uncertainty with no additional computational cost. The results show variance reconstructions for an FWI inversion of the MIDA dataset [126]. The effect of prior information is assessed, although the prior is used implicitly (i.e. as a starting point for the optimisation) rather than explicitly (i.e. as a constraint). Finally, the dependence of the posterior on the hyperparameters of the optimisation algorithm is discussed.

3.2 Theory

This section presents a technical introduction to two statistical methods; *maximum likelihood estimation*, which is the basis of traditional reconstruction, and stochastic variational inference, which is relevant to the development of the method in Section 3.3. In addition, the discussion of MLE and SVI will contextualise their use in tomography and adjoint methods. The link between iterative tomographic reconstruction and MLE has been explored by others [104], [106], but this interpretation is not explicitly discussed in the FWI literature. Furthermore, although iterative reconstruction has become popular across tomography, uncertainty estimation has not achieved widespread use due to the computational cost involved in estimating the Hessian of the likelihood function.

Full-waveform inversion for ultrasound tomography is of specific interest. Ultrasound tomography seeks to invert observed wavefield data to obtain a pixel-wise estimate of the sound-speed of an object. Observed wavefield data is produced by measuring pressure changes on the surface of a transducer, herein referred to as a receiver. The pressure wavefield originates at a source (also a piezo-electric transducer), which produces a pressure disturbance in its surroundings. The disturbance propagates away from the source and is modulated by a sample, or patient, before it arrives at the receiver. In other applications the sources and receivers can use different methods to produce and record wavefield disturbances, this can include recording displacement rather than pressure.

3.2.1 The maximum likelihood interpretation of FWI

FWI is a reconstruction algorithm that uses a model of the physics to simulate a wavefield. A popular choice of physics model is the acoustic wave-equation, whose operator $\Box(c)$ is defined,

$$\Box(c) = \left[c^{-2} \frac{\partial^2}{\partial t^2} - \nabla^2 \right] \tag{3.1}$$

This is also known as the d'Alembert operator. The acoustic wave-equation operator is dependent on a set of acoustic sound speeds c, which are usually time-constant but spatially

inhomogeneous.

In general, the ground-truth physics is modelled using a partial differential equation operator $H(\cdot)$ with model parameters m. Together, the operator, the model parameters, and a source disturbance $\varsigma(x_s, t)$, can be solved to simulate the propagation of a wavefield,

$$H(m)\, u_p(x, t; m, x_s, \varsigma) = \varsigma(x_s, t) \tag{3.2}$$

where the simulated wavefield $u_p(x, t)$ is commonly referred to as the predicted wavefield.

In the case of US-CT, the source disturbance is the time-varying pressure on the surface of the emitting transducer. Clearly, the source disturbance is non-zero only at the source positions x_s. Similarly, the wavefield cannot be measured throughout the imaging region. Instead, predicted wavefield data $d_p(x_r, t)$ are extracted from the wavefield $u_p(x, t)$ at the receiver positions x_r such that,

$$d_p(x_r, t; m, x_s, \varsigma) = u_p(x_r, t; m, x_s, \varsigma) \tag{3.3}$$

Thus Equation 3.2 is a time dependent map from the model parameters to the wavefield data. This means a further model can be defined,

$$L(m; t, x_r, x_s, \varsigma) = d_p(x_r, t; m, x_s, \varsigma) \tag{3.4}$$

where $L(m)$ is a non-linear predictor, which maps the model parameters to the predicted wavefield data for a particular time position, receiver position, source position, and source disturbance. In practice, a numerical simulator (for example, finite-difference, finite-element, etc. [90]) can be used to solve Equations 3.2 and 3.3 so that predicted wavefield data are simulated from a set of model parameters.

The predictor can estimate the observed wavefield data $d(x_r, t, x_s, \varsigma)$ such that,

$$d = L(m) + \gamma \tag{3.5}$$

which differs from the output of the predictor by an error γ. Note that the fixed parameters

have been dropped for compactness. By a simple rearrangement we obtain,

$$\gamma = d - L(m) \tag{3.6}$$

Now, consider an experiment where a pressure disturbance is produced at a source. The disturbance propagates from the source to the receivers, and is modulated by the parameters of the materials between them. The observed wavefield data $d(x_r, t)$ is the measurement of the pressure at the receivers. In this example, the model parameters are fixed, but other uncontrolled variables might alter the measurement in a random manner. Thus, the observed wavefield data d is a random function of the fixed model parameters m of the sample. Conventional FWI assumes the model parameters m are known and that the error follows a Gaussian distribution, so Equation 3.6 can be used to define a probability distribution $p(\mathcal{D}|m)$ with mean 0 and variance Σ^2, where $\mathcal{D}$ is the set of all the measurements of the observed wavefield data $\mathcal{D} = \{d\}$.

A different situation arises when solving the inverse problem. In this case, the model parameters are varied in order to maximise the estimated probability of some fixed observed wavefield data. Thus, the output is no longer a true estimate of the probability of the data because the true model parameters are unknown. However, the probability distribution of the error provides the structure for the relationship between the observed wavefield data and the model parameters. Hence, Equation 3.6 is used to define a deterministic function $\mathcal{L}(\mathcal{D}|m)$ which describes the likelihood of the data given the model parameters. Assuming the true error is Gaussian distributed, then the likelihood function for a particular model parameter estimate is,

$$\mathcal{L}(\mathcal{D}|m) = \det(2\pi\Sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[\gamma^T\ \Sigma^{-2}\ \gamma]\right\} \tag{3.7}$$

and, by inserting the definition of the error from Equation 3.6,

$$\mathcal{L}(\mathcal{D}|m) = \det(2\pi\Sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[d - L(m)]^T\ \Sigma^{-2}\ [d - L(m)]\right\} \tag{3.8}$$

The maximum of the likelihood function is found by optimisation with respect to the

model parameters such that,

$$\hat{m} = \arg\max_{m} \det(2\pi\Sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[d - L(m)]^T \ \Sigma^{-2} \ [d - L(m)]\right\} \tag{3.9}$$

where $\hat{m}$ is optimum set of model parameters, also known as the maximum likelihood estimator.

As the Equation 3.7 has the form $f(\gamma) = exp(-\gamma^T \gamma)$, the magnitude of the function decreases as the magnitude error increases, meaning that an equivalent but numerically simpler process is to maximise the data-fit directly,

$$\hat{m} = \arg\max_{m} -\frac{1}{2}[d - L(m)]^T \ [d - L(m)] \tag{3.10}$$

where it is also assumed that $\Sigma^2 = \mathrm{I}$ and I is the identity matrix.

This data-fit function is a popular choice for FWI [15], [71], [138], and the same process can be followed for any system described by a physical model $L(m)$. Thus, we have demonstrated that many FWI approaches assume the observed wavefield data are Gaussian distributed.

The most important implication of this section is to justify the adoption a Bayesian likelihood with a Gaussian form in Section 3.3.2. We have demonstrated that this choice is consistent with the approximations adopted for traditional iterative reconstruction. Furthermore, a no-cost uncertainty estimator can be found using the Bayesian setting. This is in contrast to previous work, which has used the Hessian of the likelihood function to estimate variances associated with tomography. Obtaining variance via the Hessian is consistent with an MLE interpretation of tomography as these variance estimates are found using the Cramer-Rao lower bound [104], [106]. However, Hessian based approaches are too numerically expensive to achieve wide-spread use.

3.2.2 The variational Bayesian interpretation of FWI

Variational inference is a method to solve Bayes' equation by gradient-based optimisation. Because the MLE approach also uses gradient-based optimisation, variational inference is more comparable to the MLE-based approach than alternative Bayesian methods such as Monte-Carlo sampling. To introduce variational inference, we start by deriving the evidence lower bound (ELBO), which underlies all variational inference. Finally, we discuss the two solutions of the ELBO which encompass stochastic variational inference.

Variational inference is a Bayesian method. In the context of FWI, Bayes' equation can be written as,

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \tag{3.11}$$

where the 4 terms present are: the posterior distribution $p(m|\mathcal{D})$, the prior distribution $p(m)$, the evidence distribution $p(\mathcal{D})$, and a conditional distribution $p(\mathcal{D}|m)$ which can be interpreted as a likelihood function for fixed $\mathcal{D}$.

Previously, we stated that MLE seeks to obtain a point-estimate $\hat{m}$ at the optimum of the likelihood function. In contrast, Bayesian methods seek to obtain a distribution, called the posterior distribution $p(m|\mathcal{D})$, which describes the distribution of a set of model parameters m based on a set of observed measurement data $\mathcal{D} = \{d\}$. Moments (i.e. mean, variance) can be obtained directly from the posterior distribution, which means that, in comparison to MLE, Bayesian methods can obtain additional information from the measurement data without requiring the computationally expensive Hessian. Similar to the previous section, $p(\mathcal{D}|m)$ can be interpreted as a likelihood function for fixed $\mathcal{D}$. Section 3.3 demonstrates that this likelihood function can also use a Gaussian approximation. The prior distribution $p(m)$ describes the expected distribution of the model parameters before any data is observed. When multiplied together using the law of total probability, the prior and likelihood are collectively called the joint distribution $p(m, \mathcal{D})$ such that $p(m, \mathcal{D}) = p(\mathcal{D}|m)p(m)$. The evidence is the marginal of the joint distribution, which is defined symbolically below.

The aim of Bayesian methods is to obtain an estimate of the posterior distribution using known estimates of the likelihood, prior and evidence. This is difficult to achieve for most

inverse problems because the number of model parameters means the evidence is numerically intractable. To illustrate this, consider that the evidence is obtained by marginalising the joint distribution $p(\mathcal{D}, m)$ over the model parameters,

$$p(\mathcal{D}) = \int_m p(\mathcal{D}, m)\, dm \tag{3.12}$$

which can be rearranged using the law of total probability,

$$p(\mathcal{D}) = \int_m p(\mathcal{D}|m)\, p(m)\, dm \tag{3.13}$$

The numerical integration, which is required to solve Equation 3.13, is intractable due to the number of combinations of model parameters, which also means the posterior is intractable. Exceptions occur when the prior and likelihood have analytical expressions which are conjugate (i.e. they come from related distributional families, for example, when both distributions are Gaussian). In this case, analytical solutions to Bayes' equation can be found which are relatively inexpensive to invert, such as by a Gaussian process.

Another class of solvers are Monte-Carlo methods which sample the posterior directly such that,

$$q_{MC}(m|\mathcal{D}) = \frac{1}{C} \sum_i^N p(\mathcal{D}|m^{(i)}) p(m^{(i)}) \tag{3.14}$$

where $q_{MC}(m|\mathcal{D})$ is the Monte-Carlo estimate of the posterior, $p(m^{(i)}) \sim p(m)$ is a sample from the prior distribution, and C is a normalising constant. Thus, samples from the prior distribution $p(m)$ are used to calculate the joint estimate $p(m, \mathcal{D})$, from which the Monte-Carlo approximation to the posterior can be calculated directly by summing over multiple samples. Note that the evidence distribution has been replaced by the normalising constant C, which simply scales the output distribution into the range between 0 and 1. As the number of samples increases, the approximate Monte-Carlo posterior $q_{MC}(m|\mathcal{D})$ will converge to the true posterior $p(m|\mathcal{D})$. Monte-Carlo methods can be effective, but they are fundamentally limited because they need at least tens of samples per parameter. This makes Monte-Carlo methods computationally expensive, and often intractable, for high

dimensional problems such as tomography.

An alternative approach is variational inference, which tries to solve Bayes' equation using gradient-based optimisation methods. Variational inference seeks to compare an approximating distribution $q(m)$ to the posterior distribution $p(m|\mathcal{D})$ by using the Kulback-Leibler (KL) divergence,

$$\mathbb{KL}[q(m)\,||\,p(m|\mathcal{D}))] = \int_m q(m)\,\ln\left[\frac{q(m)}{p(m|\mathcal{D})}\right]\,dm \tag{3.15}$$

At this point the approximating distribution has no specific form, for example it could be parametric or non-parametric, mono-modal or multi-modal. Later, we will make some assumptions about the structure of the distribution to make the problem tractable. For now, we note that the optimal approximate distribution $q^*(m)$ is found at the minimum of the KL divergence,

$$q^*(m) = \min_{q(m)}\ \mathbb{KL}[q(m)\,||\,p(m|\mathcal{D}))] \tag{3.16}$$

This formulation leaves two unsolved problems:

1. The posterior $p(m|\mathcal{D})$, as mentioned above, is intractable.

2. Performing gradient descent with respect to a distribution (i.e. the derivative with respect to the approximating distribution $q(m)$) is not well defined.

Problem I: The evidence lower bound, commonly referred to as the ELBO, provides the solution to problem (i). To find the ELBO, start by using Bayes' equation $p(m|\mathcal{D}) = p(m,\mathcal{D})/p(\mathcal{D})$ to split the posterior in Equation 3.15 such that,

$$\mathbb{KL}[q(m)\,||\,p(m|\mathcal{D}))] = \ln p(\mathcal{D}) + \int_m q(m)\,\ln\left[\frac{q(m)}{p(m,\mathcal{D})}\right]\,dm \tag{3.17}$$

Note that the logarithm of the evidence distribution is always negative because probabilities are bounded from 0 to 1 and that, by definition, the KL-divergence is positive;

$$\mathbb{KL}[q(m)\,||\,p(m|\mathcal{D}))] \geqslant 0 \tag{3.18}$$

Therefore, by applying this definition to Equation 3.17, we have,

$$\ln p(\mathcal{D}) \geqslant -\mathbb{KL}[q(m) \,||\, p(m, \mathcal{D}))] \tag{3.19}$$

which defines the evidence lower bound. Finally, observe from Equation 3.17 that $\ln p(\mathcal{D}) = -\mathbb{KL}[q(m) \,||\, p(m, \mathcal{D})]$ if, and only if, $\mathbb{KL}[q(m) \,||\, p(m|\mathcal{D}))] = 0$. Therefore, the maximum of the evidence lower bound occurs if, and only if, the approximating distribution is equal to the posterior distribution. Thus, an new optimisation problem (which is equivalent to Equation 3.16) is obtained such that,

$$\begin{aligned} q^*(m) &= \max_{q(m)} \mathbb{KL}[q(m) \,||\, p(m, \mathcal{D})] \\ &= \max_{q(m)} \int_m q(m) \, \ln \left[\frac{q(m)}{p(m, \mathcal{D})} \right] dm \end{aligned} \tag{3.20}$$

Because this KL-divergence is taken with respect to the joint distribution $p(m, \mathcal{D})$ rather than the posterior $p(m|\mathcal{D})$, any explicit dependence on the evidence distribution $p(\mathcal{D})$ has been removed from the optimisation problem. Therefore, the marginalisation of the joint distribution (from Equation 3.13) is no longer required and we find that Equation 3.20 can be solved numerically. Section 3.3 deals with the practical aspects of specifying each distribution such that they are tractable for FWI, meaning that we can conclude the discussion of problem (i).

Problem II: There are several approaches to solve problem (ii). Two of these approaches are classed as stochastic variational inference (SVI) and seek to replace the approximating distribution $q(m)$ with a parametric equivalent $q(m; \lambda)$, where λ are a set of parameters describing the distribution. This means that the optimisation problem can be posed to optimise the ELBO with respect to the parameters of the approximating distribution rather than the distribution itself. Therefore, consider that we can choose a parametric approximating distribution $q(m; \lambda)$ (e.g. the binomial distribution, the Gaussian distribution, etc.), and adjust Equation 3.20 such that,

$$\begin{aligned} q^*(m|\lambda) &= \arg\max_{\lambda} \mathbb{KL}[q(m;\lambda)\,||\,p(m,\mathcal{D})] \\ &= \arg\max_{\lambda} \int_m q(m;\lambda)\,\ln\left[\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right] dm \end{aligned} \tag{3.21}$$

The aim is to update the approximating distribution using the gradient of the KL-divergence in Equation 3.21. Both of the SVI methods discussed below will obtain the gradient by Monte-Carlo estimation but, in contrast to direct Monte-Carlo estimation of the posterior (i.e. Equation 3.14), we will use a Monte-Carlo method to estimate the gradient of the ELBO (as defined in Equation 3.21). Monte-Carlo methods sample from a distribution so that the approximate moments of the distribution can be estimated. In this case, the Monte-Carlo method allows the first moment (i.e. the expectation or mean) of the ELBO to be estimated.

To obtain the Monte-Carlo estimate of the gradient, first it is necessary to reinterpret Equation 3.21 as an expectation such that,

$$\mathbb{KL}[q(m;\lambda)\,||\,p(m,\mathcal{D})] = \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] \tag{3.22}$$

Thus, a gradient of the ELBO can be taken with respect to the parameters λ of the optimisation,

$$\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = \nabla_\lambda \int_m q(m;\lambda)\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\} dm \tag{3.23}$$

Superficially, it appears that we can calculate the gradient of the ELBO by taking the expectation of the gradient of the term in the logarithm such that,

$$\begin{aligned} \mathbb{E}_{q(m;\lambda)}\left[\nabla_\lambda \ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] &= \int_m q(m;\lambda)\nabla_\lambda \ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\} dm \\ &\neq \nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] \end{aligned} \tag{3.24}$$

Unfortunately, it is apparent that this operation is not valid because of the dependence

of the expectation on the approximating distribution itself. This means that the estimator on the right hand side of Equation 3.24 is biased relative to the estimator defined in Equation 3.23. In other words, a gradient of the ELBO with respect to the parameters of the approximating distribution ∇_λ cannot be exchanged with the expectation term, because the expectation is implicitly dependent on an integrand which contains the approximating distribution such that $\mathbb{E}_{q(m;\lambda)}\left[\, f(m) \,\right] = \int q(m;\lambda) f(m) dm$.

Alternatively, the score function approach [113] rearranges Equation 3.22 to find an unbiased estimator of the ELBO. Using the chain rule and a method called the log-derivative trick, in which the relationship $\nabla_x f(x) = f(x) \nabla_x \ln[f(x)]$ is used, we obtain an unbiased estimate of the gradient of the ELBO,

$$
\begin{aligned}
&\nabla_\lambda \mathbb{E}_{q(m;\lambda)} \left[\ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} \right] \\
&\quad = \int_m q(m;\lambda)\, \nabla_\lambda \ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} dm \\
&\qquad + \int_m q(m;\lambda)\, \ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} \nabla_\lambda \ln \left\{ q(m;\lambda) \right\} dm \\
&\quad = \mathbb{E}_{q(m;\lambda)} \left[\nabla_\lambda \ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} \right] \\
&\qquad + \mathbb{E}_{q(m;\lambda)} \left[\ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} \nabla_\lambda \ln \left\{ q(m;\lambda) \right\} \right]
\end{aligned}
\tag{3.25}
$$

Because the estimator in Equation 3.25 is unbiased, we can take a Monte-Carlo estimate of the gradient. Thus, the approximating distribution is sampled so $q(m^{(i)};\lambda) \sim q(m;\lambda)$, and the integral in Equation 3.25 can be replaced with an asymptotically equivalent operation, namely the discerete sum over the samples, such that,

$$
\begin{aligned}
&\nabla_\lambda \mathbb{E}_{q(m;\lambda)} \left[\ln \left\{ \frac{q(m;\lambda)}{p(m,\mathcal{D})} \right\} \right] \\
&\quad \approx \frac{1}{N} \sum_i^N \nabla_\lambda \ln \left\{ \frac{q(m^{(i)};\lambda)}{p(m^{(i)},\mathcal{D})} \right\} \\
&\qquad + \frac{1}{N} \sum_i^N \ln \left\{ \frac{q(m^{(i)};\lambda)}{p(m^{(i)},\mathcal{D})} \right\} \nabla_\lambda \ln \left\{ q(m^{(i)};\lambda) \right\}
\end{aligned}
\tag{3.26}
$$

Unfortunately, the score function estimator has high variance because the probabilities are bounded from 0 to 1 and the logarithm tends to infinity when its argument tends to zero, a situation which occurs in the first part of the second expectation in Equation 3.26. This can slow the convergence of the algorithm because the direction of successive gradient estimates becomes unstable.

An alternative approach is to use the reparameterisation trick [112], [114], [139], [140]. During reparameterisation, the parametric approximating distribution $q(m;\lambda)$ is transformed into a deterministic function $g(\cdot)$ of another distribution $q(\epsilon)$ with parameters $\{\lambda\}$ such that,

$$q(m;\lambda) = g[\{\lambda\}, q(\epsilon)] \tag{3.27}$$

As an example, consider a Gaussian distribution where the approximating distribution $\mathcal{N}(\mu,\Sigma)$ can be transformed into the standard normal distribution $\mathcal{N}(0,1)$ using,

$$\begin{aligned} \mathcal{N}(\mu,\Sigma) &= g[\,\{\mu,\Sigma\},\ \mathcal{N}(0,1)\,] \\ &= [\Sigma^T\,\mathcal{N}(0,1)] + \mu \end{aligned} \tag{3.28}$$

where the parameters of the transform λ are the mean vector μ and standard deviation matrix Σ.

The reparameterisation trick means that the gradient of the ELBO can be taken with respect to the parameters of the transform,

$$\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = \nabla_\lambda \int_\epsilon q(\epsilon)\,\ln\left\{\frac{g[\lambda,\epsilon]}{p(g[\lambda,\epsilon],\mathcal{D})}\right\} d\epsilon \tag{3.29}$$

where the shorthand $g[\{\lambda\}, q(\epsilon)] \equiv g[\lambda,\epsilon]$ is used. Splitting the expression into a regularisation term and a likelihood term gives,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] &= \nabla_\lambda \int_\epsilon q(\epsilon)\ \ln\left\{\frac{g[\lambda,\epsilon]}{p(g[\lambda,\epsilon])}\right\}\ d\epsilon \\ &\quad - \nabla_\lambda \int_\epsilon \ln\left\{p(\mathcal{D}\,|\,g[\lambda,\epsilon])\right\}\ d\epsilon\end{aligned} \tag{3.30}$$

By rearranging the expression, taking the derivative through the integrand, and using the chain rule, we find,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] &= \int_\epsilon q(\epsilon)\,\nabla_g \ln\{g[\lambda,\epsilon]\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon \\ &\quad - \int_\epsilon q(\epsilon)\,\nabla_g \ln\{p(g[\lambda,\epsilon])\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon \\ &\quad - \int_\epsilon q(\epsilon)\,\nabla_g \ln\left\{p(\mathcal{D}\,|\,g[\lambda,\epsilon])\right\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon\end{aligned} \tag{3.31}$$

and, finally, taking a Monte-Carlo estimate with respect to the new distribution,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] &= \frac{1}{N}\sum_i^N \nabla_g \ln\{g[\lambda,\epsilon_i]\}\nabla_\lambda\, g[\lambda,\epsilon_i]\,d\epsilon \\ &\quad - \frac{1}{N}\sum_i^N \nabla_g \ln\{p(g[\lambda,\epsilon_i])\}\nabla_\lambda\, g[\lambda,\epsilon_i]\,d\epsilon \\ &\quad - \frac{1}{N}\sum_i^N \nabla_g \ln\left\{p(\mathcal{D}\,|\,g[\lambda,\epsilon_i])\right\}\nabla_\lambda\, g[\lambda,\epsilon_i]\end{aligned} \tag{3.32}$$

This is called a pathwise gradient estimator. It is less flexible than the score function estimator, because it requires a parametric transform from a standard distribution, but in Section 3.3 the pathwise estimator is used because it is faster to converge and has lower variance.

3.3 Method

Variational inference allows us to avoid the computationally intractability of the evidence by arranging Bayes' equation (Equation 3.11) such that it can be solved by specifying only

the prior distribution, approximating distribution, and likelihood function. It is practical to parameterise these distributions in a number of ways, but the simplest statistical model is to assume that all of the distributions have mean-field Gaussian form. This means that the distributions are easy to specify and solve, but also that they won't provide a true estimate of the variance because the mean-field Gaussian ignores correlation between the pixels. The subtleties of this approximation are discussed in Section 4.4.

Below, we present the mean-field Gaussian solution for FWI which can be repurposed to solve other forms of tomography and adjoint methods with limited effort. The mean-field Gaussian produces a mean update, which is similar to standard FWI (for a summary of standard FWI see Section 3.2.1), and a variance update, which has no additional computational cost compared to standard FWI. This proof is the main contribution of this paper. Because VT does not alter the forward or inverse simulators, I expect that VT will generalise to other iterative tomographic reconstruction algorithms, for example which use ray tracing for simulation.

3.3.1 An optimisation perspective

It is common to interpret FWI as a local optimisation process with respect to an objective function, where the objective function includes data-fit terms and excludes regularisation terms [15]. Any correct Bayesian interpretation of FWI must include a consideration of prior information and would use regularisation, for example see Section 7.1. However, correctly incorporating prior information is complex and problem specific [133], which would add additional complexity to this work. Instead we will consider the correct incorporation of prior information in future work. Therefore, it is expedient to consider the estimation of the parameters of the approximating distribution purely as an optimisation process. Rewriting Equation 3.22 such that,

$$\begin{aligned}\mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = &-\int_{\epsilon} q(\epsilon)\,\ln\left\{p(\mathcal{D}\,|\,m)\right\}\,d\epsilon \\ &+\int_{\epsilon} q(\epsilon)\,\ln\left\{\frac{q(m;\lambda)}{p(m)}\right\}\,d\epsilon\end{aligned} \tag{3.33}$$

it is clear [133] that this expression is equivalent to an optimisation with an objective function of the form,

$$\mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = -\mathcal{L}(m) + a\,\mathcal{S}(m) \tag{3.34}$$

where the constant a has been introduced. This constant modulates the importance of the data-fit term $\mathcal{L}$ relative to the regularisation term $\mathcal{S}$. In practice, the constant is adjusted as a hyper-parameter of the optimisation process. In our analysis we choose $a = 0$, which is a common choice for FWI [15]. Therefore, the full gradient expression, shown in Equation 3.32, can be simplified to,

$$\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = -\int_\epsilon q(\epsilon)\nabla_g \ln\{p(\mathcal{D}|g[\lambda,\epsilon])\}\nabla_\lambda g[\lambda,\epsilon]d\epsilon \tag{3.35}$$

In the following, Equation 3.35 is used as the starting point for the calculation of the gradients with respect to the mean and standard deviation. This will produce solutions for the approximate posterior which are strictly incorrect, but these solutions are illustrative of the broader technique and have some useful properties.

3.3.2 The mean-field Gaussian approximation

Starting with the prior distribution $p(m)$, assuming it has a mean-field Gaussian parameterisation, and using Equation 3.28 we find,

$$p(m) = [\Sigma_{prior}^T\ q(\epsilon)] + \mu_{prior} \tag{3.36}$$

The prior information is not used for regularisation and only to start the reconstruction process. The prior distribution is a special case of the approximating distribution. In general the approximating distribution is,

$$q(m;\lambda) = [\Sigma^T\, q(\epsilon)] + \mu \tag{3.37}$$

where $\lambda = \{\mu, \Sigma\}$ are the variational parameters and $q(\epsilon) = \mathcal{N}(0, I)$ is a vector whose elements are samples from the standard normal distribution.

Now, only the likelihood distribution remains to be specified. When describing standard (MLE) FWI, we demonstrated that a data-fit function using the L2-norm is equivalent to a Gaussian likelihood function. Therefore, we will make a comparable choice and use a standard normal Gaussian likelihood function, which leads to the expression,

$$\begin{aligned} &\ln\{p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,)\} = \\ &\quad \ln\left(\text{trace}(2\pi I)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[d - L(m)]^T\ [d - L(m)]\right\}\right) \end{aligned} \tag{3.38}$$

The exponential term in the definition of the Gaussian cancels with the logarithm and we obtain the data-fit,

$$\ln\{p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,)\} = \frac{1}{2}[d - L(g[\lambda,\epsilon])]^T[d - L(g[\lambda,\epsilon])] \tag{3.39}$$

where $q(m;\lambda) = g[\lambda,\epsilon]$ using the reparameterisation trick. The variational inference likelihood is still a function of the model parameters, but the model parameters are specified by the approximating distribution and, in turn, the approximating distribution is a deterministic function of the two variational parameters.

For acoustic FWI, the model parameters are the acoustic sound speed $m = c$ discretised into pixels and, therefore, the approximating distribution is over the acoustic sound-speed of each pixel. In Geophysics it is common to extend acoustic FWI to include anisotropy [119], such that each sound speed "pixel" becomes a vector, or density, as an additional parameter [71]. In the case of density, a promising approach could be to use an empirical relationship with sound speed [141], which is already a well-established for different tissues [29] and could be incorporated into the prior. The inclusion of these parameters can increase the ill-posedness of the problem [15], so the scalar, mono-parameter case is used for this thesis.

Together, these functions describe a statistical model which can be solved by taking pathwise derivatives of the likelihood function with respect to the parameters of the mean-field Gaussian approximating distribution.

Gradient with respect to the mean: Taking Equation 3.35 the update for the mean parameter is,

$$\begin{aligned}\frac{\partial}{\partial \mu} \ln p(\mathcal{D} \,|\, g[\lambda, \epsilon]) &= \frac{\partial}{\partial g} \ln p(\mathcal{D} \,|\, g[\lambda, \epsilon]) \; \frac{\partial}{\partial \mu} g[\lambda, \epsilon] \\ &= \frac{1}{2} \; \frac{\partial}{\partial g} \; [d - L(g)]^T \; [d - L(g)] \; \frac{\partial}{\partial \mu} (\mu + \Sigma^T q(\epsilon))\end{aligned} \tag{3.40}$$

Then, using the numerator form of the multi-variable chain rule $\frac{du^T v}{dm} = u^T \frac{dv}{dm} + v^T \frac{du}{dm}$, and the relationship $A^T B = B^T A$, we obtain,

$$\frac{\partial}{\partial \mu} p(\mathcal{D} \,|\, g[\lambda, \epsilon]) = \left(\frac{\partial L}{\partial g} \right)^T \; [d - L(g)] \tag{3.41}$$

Finally, taking a sample from the approximating distribution $m^i \sim g[\lambda, \epsilon]$ means,

$$\frac{\partial}{\partial \mu} p(\mathcal{D} \,|\, m^i) = \left(\frac{\partial L}{\partial m^i} \right)^T \left[d - L\left(m^i\right) \right] \tag{3.42}$$

Therefore, the iterative process to obtain the mean is,

$$\mu = \mu - \tau \Delta_m \tag{3.43}$$

where $\Delta_m = \frac{\partial}{\partial \mu} p(\mathcal{D} \,|\, m^i)$ and τ is the step length. Section 3.4 shows that this thesis determines step length by a grid-search. However, other strategies for determining the step length are used in numerical optimisation. One strategy is to view the gradient descent as a Newtonian optimisation scheme, which means that the step length represents a scalar approximation of the Hessian (i.e. $\tau \approx \nabla^2 \mathcal{L}(\mathcal{D}|m)$) and has a bound determined by the Lipschitz constant. Further details are available in [142], (pages 42-44).

Note that by assuming $q(\epsilon) = 0$, then $g(\lambda, \epsilon) = \mu$. Thus, the standard, deterministic, FWI algorithm is returned from the mean update when $q(\epsilon) = 0$. In addition, the Monte-

Carlo sample m^i is not a probability (i.e. bounded from 0 to 1) because we are not dealing directly with the probability distribution $q(m;\lambda) = g[\lambda,\epsilon]$. Instead, the function $g[\lambda,\epsilon]$ is used to obtain the sound-speed estimate, which corresponds to a probability $q(\epsilon^i) \sim q(\epsilon)$.

Gradient with respect to the standard deviation: Similarly, we find the standard deviation by taking the derivative with respect to the standard deviation parameter Σ,

$$\begin{aligned}\frac{\partial}{\partial\Sigma}\,\ln\,p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) &= \frac{\partial}{\partial g}\ln\,p(\,\mathcal{D}\,|g[\lambda,\epsilon]\,)\;\frac{\partial}{\partial\Sigma}g[\lambda,\epsilon] \\ &= \frac{1}{2}\;\frac{\partial}{\partial g}[d-L(g)]^T\,[d-L(g)]\;\frac{\partial}{\partial\Sigma}(\mu+\Sigma^T q(\epsilon))\end{aligned} \tag{3.44}$$

Taking the derivatives of the expressions means,

$$\frac{\partial}{\partial\Sigma}\,\ln\,p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) = [d-L(g)]^T\;\left(\frac{\partial L}{\partial g}\right)\;q(\epsilon) \tag{3.45}$$

This expression can be rearranged using the relationship $A^T\,B\,C = C^T\,B^T\,A$, such that,

$$\frac{\partial}{\partial\Sigma}\,\ln\,p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) = q(\epsilon)^T\;\left(\frac{\partial L}{\partial g}\right)^T\;[d-L(g)] \tag{3.46}$$

Finally, sampling from the standard normal distribution $\epsilon^i \sim q(\epsilon)$, we have

$$\frac{\partial}{\partial\Sigma}\,p(\,\mathcal{D}\,|\,m^i\,) = (\epsilon^i)^T\;\left(\frac{\partial L}{\partial m^i}\right)^T\;[d-L(m^i)] \tag{3.47}$$

This gradient provides the update for the standard deviation. Therefore the iterative process to obtain the standard deviation is,

$$\Sigma = \Sigma - (\tau)\,(\epsilon^i)^T\,\Delta_m \tag{3.48}$$

where τ is the step length, which has the same value as in Equation 3.43.

Because uncertainty is equal to the variance of the distribution, we take the square of the standard deviation image (Σ^2) once the iterative process has completed.

In the statistical model above, the mean-field approximation is used to simplify the description. However, researchers often have an interest in the covariance of the model parameters. We will propose a model-based approach to this in Section 4.4.3 and Chapter 7, but it is also possible to specify a full Gaussian for the approximating distribution (Equation 3.37). This is done by enumerating the off-diagonal elements of the variance matrix (also called the covariance matrix), and then the usual derivation can be followed as above.

3.3.3 The adjoint-state method and stochastic variational inference

The Fréchet derivatives ($\partial L/\partial g$) in Equations 3.41 and 3.46 are computationally expensive because a numerical simulation of the wavefield propagation must be run for each model parameter g in turn. In practice, FWI uses an adjoint-state method to make the gradient computationally tractable. Now, we will show how the adjoint-state method can be applied to the standard deviation gradient in Equation 3.46.

Starting with the expression for the forward problem in Equation 3.2,

$$\varsigma(x_s) = H(m)\, u_p(m)$$

take the derivative of both sides with respect to m and rearrange,

$$\frac{\partial u_p(m)}{\partial m}^T = u_p(m)^T \frac{\partial H(m)}{\partial m} \left[H(m)^{-1}\right]^T \tag{3.49}$$

Note that $\frac{\partial L(m)}{\partial m} = \frac{\partial u_p(m,x_r)}{\partial m}$. Then, using the probabilistic notation ($q(m;\lambda) = g[\lambda,\epsilon]$) and inserting Equation 3.49 into the standard deviation update (Equation 3.46) yields,

$$\frac{\partial}{\partial \Sigma}\, p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) = q(\epsilon)\, u_p(g)^T \frac{\partial H(g)}{\partial g} \left[H(g)^{-1}\right]^T [d - L(g)] \tag{3.50}$$

In addition, we can define a new wavefield, called the adjoint wavefield, by rearranging Equation 3.2 such that,

$$u_{adj}(m) = \left[H(m)^{-1}\right]^T \varsigma_{adj}(x_r) \tag{3.51}$$

Using the probabilistic notation, defining an adjoint source $d - L(g) = \varsigma_{adj}(x_r)$, and applying the stochastic operator $H(g)^{-1}$ to the source, the adjoint wavefield $u_{adj}(g)$ can defined as,

$$u_{adj}(g) = \left[H(g)^{-1}\right]^T \left[d - L(g)\right] \tag{3.52}$$

Finally, using this definition it becomes apparent that the last two terms in Equation 3.50 represent a new wavefield propagating from the receivers to the source,

$$\begin{aligned} \frac{\partial}{\partial \Sigma}\, p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) &= q(\epsilon)^T u_p(g)^T\, \frac{\partial H(g)}{\partial g}\, u_{adj}(g) \\ &= q(\epsilon)^T \Delta_m \end{aligned} \tag{3.53}$$

This is called the adjoint wavefield because it has terminal conditions, which are the difference between the observed and predicted wavefield data at the receivers, and is calculated by backwards time-stepping [138]. Therefore the Fréchet derivatives in Equations 3.41 and 3.46 can be calculated by calculating the zero-lag correlation between the predicted wavefield and the adjoint wavefield, and it is computationally efficient to do this at each time-step in a backwards time-stepping scheme. See section 2.3 for a visual description.

A similar analysis finds that the adjoint-state equation for the mean in Equation 3.41 is,

$$\begin{aligned} \frac{\partial}{\partial \mu}\, p(\,\mathcal{D}\,|\,g[\lambda,\epsilon]\,) &= u_p(g)^T\, \frac{\partial H(g)}{\partial g}\, u_{adj}(g) \\ &= \Delta_m \end{aligned} \tag{3.54}$$

This derivation demonstrates that the adjoint-state method calculation, required for the gradient with respect to the mean (Equation 3.54), can be recycled when calculating the gradient with respect to the standard deviation (Equation 3.53). This means that the computational cost of the standard deviation gradient is minimal.

The mean-field Gaussian pathwise derivative estimator is also simple to implement. Algorithm 1 outlines the practical aspects required to implement standard FWI and Algorithm

Algorithm 1 Full-waveform inversion

Input: m, ς, x_s, x_r
Returns: m

1: **while** $n \leqslant n_{max}$ **do**
2: $L(m) = \textbf{forward}(m, \varsigma, x_s, x_r)$ (Equation 3.2)
3: $\Gamma = || \, d - L(m) \, ||_2^2$
4: $\Delta_m = \textbf{adjoint}(m, \varsigma, x_s, x_r, \Gamma)$ (Equation 3.54)
5: $m \leftarrow m + \Delta_m$

Algorithm 2 Full-waveform inversion using Stochastic Variational Inference

Input: $\mu, \Sigma, \varsigma, x_s, x_r$
Returns: μ, Σ

1: **while** $n \leqslant n_{max}$ **do**
2: $\epsilon^i \sim \mathcal{N}(0, I)$
3: $m^i = \mu + \Sigma \cdot \epsilon^i$
4: $L(m) = \textbf{forward}(m^i, \varsigma, x_s, x_r)$ (Equation 3.2)
5: $\Gamma = || \, d - L(m) \, ||_2^2$
6: $\Delta_m = \textbf{adjoint}(m^i, \varsigma, x_s, x_r, \Gamma)$ (Equation 3.54)
7: $\mu \leftarrow \mu + \Delta_m$
8: $\Sigma \leftarrow \Sigma + \epsilon^{(i)} \cdot \Delta_m$

2 outlines the practical aspects required to implement VT-FWI. The major changes between VT-FWI and standard FWI are highlighted in red. It is apparent that VT-FWI requires three simple changes to the standard FWI approach. The important changes are in line 3, where a sample from the distribution over the model parameters in introduced, and line 8, where the update for the variance is implemented.

3.3.4 The MIDA *in-silico* dataset

The inversion was performed using in-house FWI software written in Python, details in Section 2.4. The software used a finite-difference time domain wave equation solver. A 400×530 grid with a grid-spacing of $0.5[mm]$ and a time-step of $0.049[\mu s]$ was used. The source was a 3-cycle tone burst with a centre frequency of $370[kHz]$. For the adjoint problem, this was filtered at $200[kHz]$. The aim was to invert the data using a realistic transducer density, so 240 sources and 240 receivers per source were used. The derivatives of the acoustic wave equation were approximated by finite difference stencils with time-order 2 and space-order 10 using optimised coefficients (see Section 2.2).

Inversions were performed on a slice of the MIDA head phantom [126], using acoustic tissue properties taken from the IT'IS database [29]. The forward and inverse simulations used the same physics, and therefore were susceptible to the inverse crime. However, the objective of these results is to demonstrate key properties of the VT estimator, which are

expected to hold regardless of the physical realism of the forward model. In addition, source normalisation was used. This is a strategy commonly used on real data and, therefore, this will increase the realism of the reconstruction process.

The update for the tomographic mean reconstruction is described by Equation 3.54 and the update for the tomographic standard deviation is described Equation 3.53. These updates are applied iteratively as part of a gradient descent optimisation algorithm. Gradient descent algorithms have several adjustable hyperparameters including the step length, batch size, number of iterations, and starting point. To illustrate the effect these hyperparameters have on the reconstructions, step lengths of 2, 5 & 15 and batch sizes of 6, 12, 24, & 48 sources are investigated. Sources are assigned to each batch stochastically. The number of iterations per epoch varies by batch size; if the batch size 6 then 1 epoch is 40 iterations, if the batch size 12 then 1 epoch is 20 iterations, if the batch size 24 then 1 epoch is 10 iterations, and if the batch size 48 then 1 epoch is 5 iterations. The same step length is used for the mean and standard deviation updates. In all cases 192 iterations were completed and the result from each iteration was visualised. This means itermediate results can be easily evaluated, but only the final result is shown. The effect of starting point was evaluated by assuming the prior information is a flat model representing the sound speed of water, or a complex model which also contains the true location and sound speed of the skull. In all cases the prior variance is $0.25\,[m\,s^{-1}]^2$.

Equation 3.53 describes the standard deviation update. This standard deviation has positive and negative components, so the final result is squared to obtain the variance reconstruction. The quality of the tomographic variance reconstruction is assessed qualitatively against the absolute error between the tomographic mean reconstruction (μ) and the ground truth sound speed model (X_{GT}), where the absolute error (AE) is;

$$AE = \sqrt{(\mu - X_{GT})^2} \tag{3.55}$$

This absolute error is also positive by definition, so it is expected that the variance reconstruction and absolute error should contain qualitatively similar features.

Hardware	Mean update (s)	Standard deviation update (s)
CPU	1600 ± 178	0.0020 ± 0.0002
GPU	82.9 ± 0.1	0.0066 ± 0.0003

Table 3.1: The time taken to compute the mean and standard deviation update on different devices when batch size is 24.

3.4 Results

Stochastic variational inference is often referred to as doubly stochastic gradient descent because stochastic gradient descent is used for optimisation and the variance update (Equation 3.47) contains ϵ_i (i.e. the sample from the standard normal distribution). In machine learning, SVI can produce unstable updates. Therefore, the stability of the variance estimator is assessed as a function of starting sound speed, step length, and batch size. In general, the variance estimates are qualitatively credible compared to the absolute error.

Table 3.1 shows the time required to compute the mean estimator (Equation 3.54) and the standard deviation estimator (Equation 3.53) during the *in-vitro* reconstruction. These updates were calculated on a CPU (Ryzen Threadripper 2950X 16 core, Advanced Micro Devices, CA) and a GPU (Quadro RTX 8000, Nvidia, CA). Compute time is an average of 10 iterations.

In the case of the GPU, the standard deviation update takes 0.007% of the computaion time of the mean update. It is clear that the uncertainty estimator has effectively no computational cost. As expected, the mean update is 19 times slower on the CPU than the GPU, but the standard deviation update is faster on the CPU than the GPU. This is probably because the GPU parallelisation cannot compensate for the overhead of transferring the vector of samples to the GPU. 192 iterations were used for the inversion. However, if 80 iterations were used, the results in Figure 3.1(b) (i.e. the skull prior) do not change.

3.4.1 Starting sound speed

Figure 3.1 shows the dependence of the VT reconstructions on the choice of prior sound speed. In Figure 3.1(a) the prior sound speed is water. In this case the mean reconstruction has developed features including a fast skull and contrast between grey and white matter,

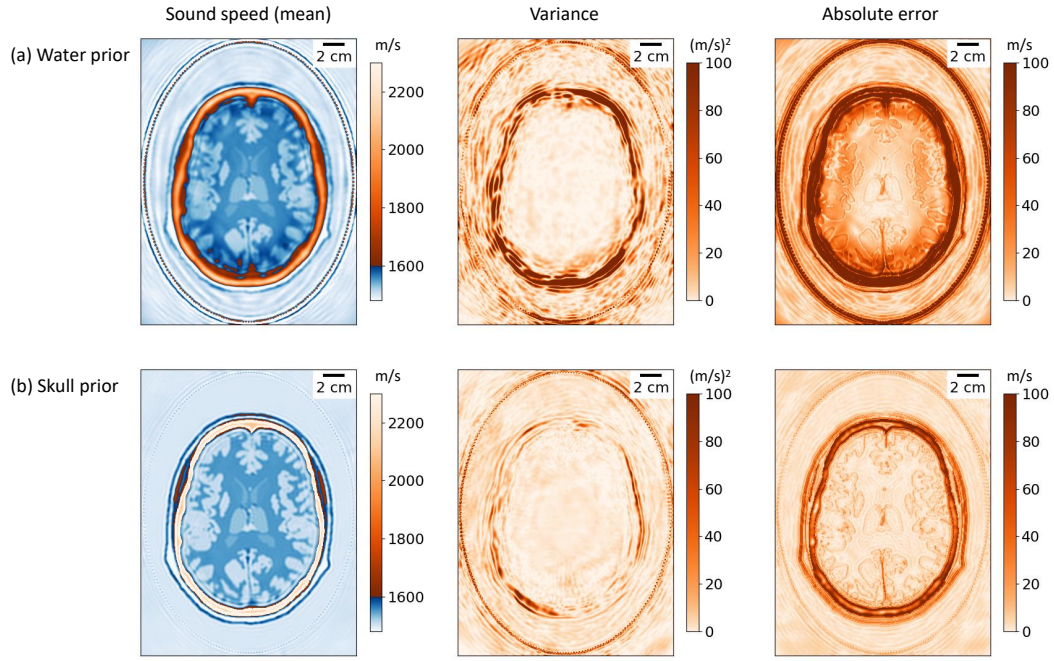


Figure 3.1: (a) The sound speed reconstruction (left), variance reconstruction (centre), and absolute error (right) for a water start. (b) The sound speed reconstruction (left), variance reconstruction (centre), and absolute error (right) for a start where the skull is correct.

but the mean reconstruction is inaccurate compared to the ground truth MIDA model. This is confirmed by the absolute error, which is particularly high in the region of the skull. The tomographic variance reconstruction correctly represents some features of the absolute error. First, the tomographic variance is high in the region of the skull. Second, the tomographic variance is slightly elevated in the brain-skull boundary region, but the correlated noise throughout the tomographic variance reconstruction makes the boundary region difficult to discern.

In Figure 3.1(b) the prior contains accurate values for the sound speed of the skull, which means FWI is able to reconstruct the posterior mean accurately. This is confirmed by the absolute error, which is higher at the boundaries of the skull, and the finding is consistent with previous analysis [13]. There are some indications of increased variance at the boundaries of the skull in tomographic variance reconstruction, for example the right-

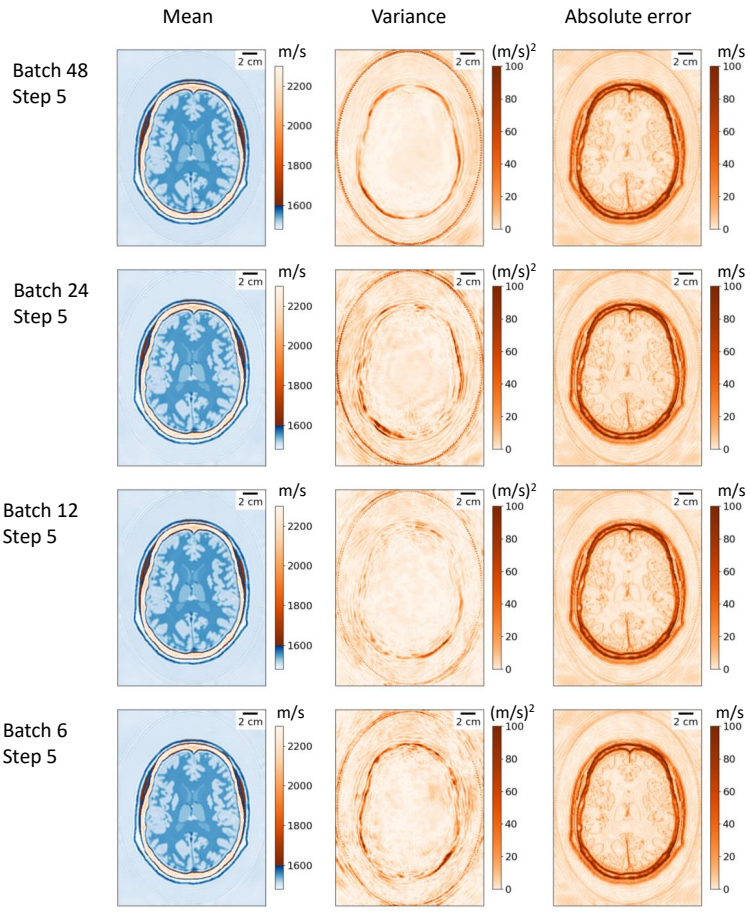


Figure 3.2: A grid search over batch size. Batch size is optimised between 6 and 48 sources, step length is 5.

hand side of the skull, but this is not consistent. The tomographic variance reconstruction is also low in the region of the brain.

3.4.2 Optimisation hyperparameters

Figures 3.2 & 3.3 show a grid search over the hyperparameters of the optimisation algorithm. Specifically, batch size (i.e. the number of sources used to calculate a single gradient) and step length (i.e. the scalar multiplier applied to the gradient) are investigated. It is apparent that step length has more influence on the accuracy of the tomographic variance estimate

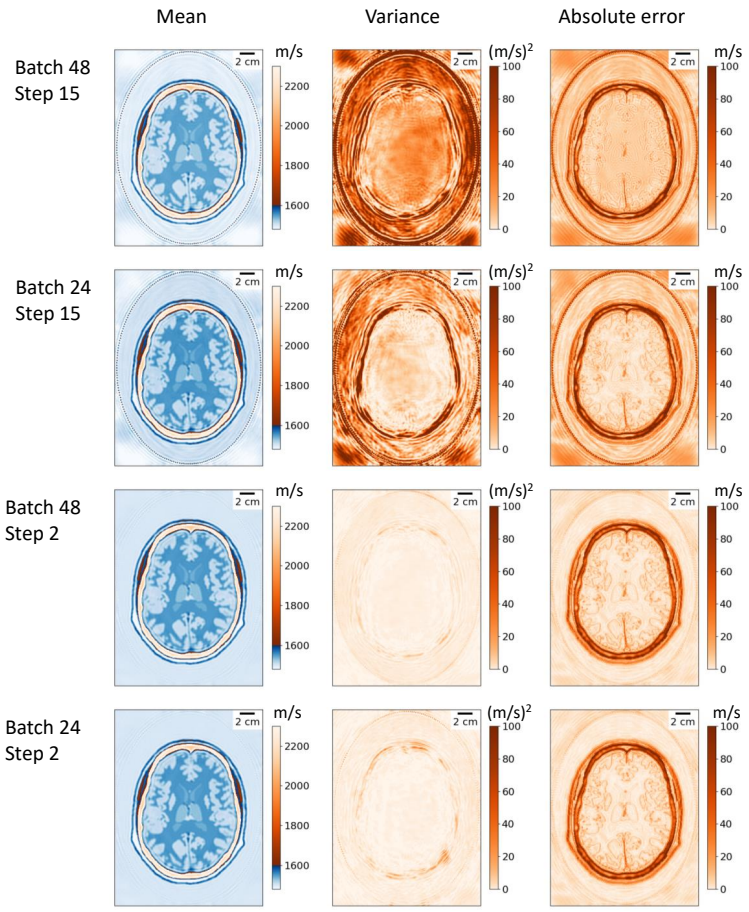


Figure 3.3: A grid search over step length. Batch size is 24 sources, step length is optimised between 2 and 15.

than batch size, but both parameters have an effect.

Batch size. When the step length is 5, batch sizes of $6, 12, 24, \&48$ are investigated. In this case, the tomographic variance reconstruction has consistent magnitude regardless of the batch size. The tomographic variance reconstructions also show that the batch size of 48 (corresponding to 20% of the 240 sources) is more uniform than when the batch size is $6, 12,$ or 24. This minor effect is the result of the 8-fold difference between the batch sizes of 48 and 6 sources per gradient, and is caused by the smaller batch size introducing more stochasticity into the gradient estimate. Similar results are apparent when the step length

is 2; the batch size does not appear to have a significant effect on the tomographic mean reconstruction and the absolute error. The batch of 48 and step of 5 is the closest visual approximation of the absolute error.

However, when the step length is 15, the batch size of 24 appears to produce a superior variance estimate compared to the batch size of 48. This could be associated with the source normalisation when the gradient is calculated; a batch size of 48 is likely to produce a gradient with fewer outliers compared to the batch size of 24 once the gradient for each source has been summed. Using the source normalisation function from Equation 2.21 means that the gradient for batch size 48 is likely to have a larger magnitude because the maximum absolute value is likely to be smaller than when the batch size is 24. Interestingly, the absolute error is also slightly lower when the step length is 15 batch size is 24, which suggests the tomographic mean reconstruction is also more accurate.

Step length. Figure 3.3 also shows the effect of changing step length. It is clear that the posterior variance is highly susceptible to step length. For example, when the step length is 2, the tomographic variance reconstructions contain few meaningful features. In contrast, when the step length is 7, the tomographic variance reconstructions contain too many regions with a high magnitude. However, when the step length is 4, the tomographic variance reconstruction correctly localises the region where the absolute error is highest.

3.5 Discussion

The results show that mean-field Gaussian stochastic variational inference is able to produce numerically stable estimates of the mean for a range of step-lengths and batch sizes. Therefore, the doubly stochastic gradient descent algorithm is not an inherent barrier to tomographic uncertainty estimation. This is not necessarily expected because stochastic variational inference can be unstable without using methods to control the variation of the gradient estimates [113].

The ability to recycle the adjoint method in Equations 3.53 and 3.54 proves that the mean-field Gaussian variance estimator is computationally inexpensive. This is because the additional vector multiplication is computationally efficient. Measurement of the computa-

tion time for the mean and standard deviation gradients support this finding and show that the estimator has a negligible cost compared to solving the FWI adjoint method. Indeed, this shows that the mean-field Gaussian estimator is substantially less computationally expensive than the Hessian calculation or the ensembling technique. It also has a smaller memory footprint than Bayesian neural networks.

Note that, if 80 iterations were used, the results in Figure 3.1(b) (the skull prior) do not change. A greater number of iterations were chosen for two reasons. First, more iterations provide more opportunity for the mean reconstruction to converge from a water prior. Unfortunately, this was not found to be the case. Second, to demonstrate that the variance reconstruction algorithm does not diverge; taking a large number of samples from the sound speed distribution means a greater chance of extreme outliers, which could affect the stability of the gradient descent. Typically, variance control methods (e.g. box constraints on the gradient) might be necessary, but this does not appear to be needed. Thus, VT provides a stable method to estimate the variance of a reconstruction using gradient descent.

3.5.1 Comments on the results

To assess the quality of the variance reconstructions, they are compared to the absolute error. The absolute error produces a similar result to VT because it has a similar mathematical formulation. Indeed, if the ground truth image is replaced by the mean reconstruction, then the standard deviation is obtained (see Section 5.2.1 for details). However, the variational tomography formulation of the standard deviation can have negative directions, so the variance reconstruction is preferred because it shows the magnitude information only. Chapter 5 uses standard deviation magnitude. This is visually similar to variance, but is easier to compare quantitatively to absolute error, a property which is unnecessary for the qualitative comparisons made here.

Starting sound speed The results in Figure 3.1 suggest that the tomographic variance accurately reflects features of the absolute error. When the skull prior is available, the absolute error is lower than when the water prior is used as the starting model. Therefore, the quality of the tomographic mean reconstruction is dependent on prior sound speed. The

variance reconstruction correctly demonstrates this; the variance is lower when the skull prior is used. Furthermore, the variance reconstruction is high in the same locations as the absolute error, suggesting that the variance estimator is able to correctly locate the regions of the tomographic mean reconstruction which are furthest from the ground truth. Figure 3.1(a) shows that the tomographic variance is able to localise the regions with high absolute error when this region is large, but Figure 3.1(b) shows that the tomographic variance localisation becomes patchy when the regions with high absolute error are narrow. This might indicate that the resolution of the tomographic variance reconstruction is lower than the resolution of the tomographic mean reconstruction.

In general, while the variance reconstruction can be compared to the absolute error overall, there is a lack of strong local correspondence. Poor localisation is discussed again in the following chapters, but there are likely to be two causes. First, there is correlated noise in the variance reconstructions. The presence of this correlated noise could indicate a failing of the mean-field Gaussian approximation and suggests that correctly characterising the correlation function between the pixels would improve the estimates. A data driven correlation function is proposed in Chapter 7. Second, the result of the VT tomographic uncertainty estimate is typically more stochastic than the VT mean estimate. This could be because the step length is inappropriate, as discussed below, or it could be the result of the scalar multiplication of the Monte-Carlo sample in the standard deviation gradient (Equation 3.53). The stochastic variation in the variance reconstruction is apparent in Figure 3.2 when the step length is 5 because the stochastic variation is greater than the effect of batch size on the result. This variability is discussed again in Chapter 5 and potential improvements to the sampling algorithm are proposed.

Optimisation hyperparameters To demonstrate the stability of the variance reconstructions, reconstructions were performed for a range of hyperparameter values. These results are shown in Figures 3.2 & 3.3. Clearly, it is possible to determine a set of hyperparameters which produce a tomographic variance estimate that is qualitatively comparable to the absolute error. This is the case for a batch size of 48 and a step length of 5, because these values produce a the tomographic variance which is uniformly high in the region of the

skull and low in the region of the brain. However, the tomographic variance reconstruction has not consistent for different choices of hyperparameter.

The tomographic variance reconstruction is particularly sensitive to step length. However, the same step length is applied to the mean and standard deviation updates, despite the fact that the variance reconstruction is in the range [0, 100], whereas the mean reconstruction in in the range [1480, 1600]. Thus, it is obvious that the tomographic variance reconstruction is more sensitive to step length because the step length is relatively larger for the tomographic variance reconstruction. This result illustrates the classic difficulty in determining the appropriate step length for gradient descent optimisation [142]. Fortunately, it is possible to determine the appropriate step length using a grid search, so this strategy was adopted throughout the Thesis.

When the step length was 15, the mean image was observed to oscillate around the ground truth solution. This is another known problem with optimisation [142] and this oscillatory behaviour suggests that the step length was too high. In addition, the most inaccurate tomographic variance reconstructions correspond to a step length of 15, which demonstrates that the instability of the mean gradient adversely affects the variance reconstruction. Similarly, a step length of 2 may be too small because the variance reconstruction does not seem to differ significantly from the prior variance of $0.25[m\,s^{-1}]^2$. A solution could be to use a quadratic step length calculation [142], which is already used in FWI [15].

The robustness of the variance reconstruction to batch size demonstrates that batch sizes corresponding to 10% of the sources are plausible. This batch size has been proposed for FWI-based US-CT neuroimaging [13], so this result demonstrates that VT can be adopted without adding extra computations to each gradient term. However, it is notable that a batch of 48 (which corresponds to 20% of the sources) produces a smoother result. This might be a desirable algorithm design choice, or remedies such as taking the mean of the final 10 iterations could suffice. Clearly, the batch size should be determined on a problem-specific basis, and this is already the case in Geophysical FWI problems [15]. In addition, the maximum batch size was chosen due to memory constraints in the PyTorch algorithm, so Chapter 5 uses Stride [143] which enables full-batch gradient descent.

3.5.2 Future work

Firstly, these results only show that the tomographic variance estimates are qualitatively comparable with absolute error and have some expected behaviours. They do not discuss repeatability. Obtaining an estimator which is repeatable is important because the aim is to obtain quantitiative uncertainty estimated from VT, and these estimates must be consistent to enable real-time IQA. Therefore, repeatability will be investigated in detail in the subsequent chapters.

Furthermore, there is no quantitative comparison between the variance estimator and absolute error, only visual analysis is used. Chapter 5 contains quantitative analysis of the variance reconstruction. In this chapter, a quantitative comparison is unsuitable because the forward and inverse problems share the same physics and experimental configuration (i.e. source receiver positions, source wavelet, etc.). This means the true variance reconstruction and absolute error should be zero, negating the purpose of a quantitative analysis. In practice, this does not occur because source amplitude normalisation is used, which modulates the contribution of these amplitudes to the gradient in an adverse way. The effect of introducing physics-model and configuration errors into the simulations is investigated in Chapter 5.

Figures 3.2 & 3.3 shows the hyperparameters of the optimisation algorithm being optimised by a grid-search. The results show that step length and batch size can be optimised in this way and it is a conventional method in Deep Learning. Therefore, a grid search is used throughout the rest of the paper. However, this process might not scale well for 3-D models, because grid searches require multiple runs in order to optimise the results, which could be prohibitively expensive. The sensitivity of the outcomes to step length is problematic and adaptive optimisation algorithms, such as Nesterov momentum, might be easier to use because they allow the step length to adapt. However, this scheme introduces an additional hyperparameter which need to be optimised. Another alternative is a quadratic line search. This adapts the step length according to a quadratic approximation of the loss function and it is already popular in Geophysical FWI.

Chapter 4

Variational tomography: Qualitative validation *in-vitro*

Chapter 3 introduced the first novel contribution of the Thesis; a Bayesian model which uses the mean-field Gaussian approximation to obtain a *maximum a-posteriori* estimator and an additional tomographic uncertainty estimator that has no additional computational cost. It has also been demonstrated that the output of this model is numerically stable for FWI, and that the hyperparameters of the optimisation algorithm can be chosen by grid-search. This chapter presents the second novel contribution of the Thesis; that the uncertainty estimator is qualitatively related to some features of image quality *in-vitro*. It is not immediately clear that inversions of real data will be successful because of well-documented complexities applying FWI to real data. The findings are also important to demonstrate that the tomographic uncertainty estimate can be produced for real data and that it has a purpose in tomographic medical imaging. The production of the *in-vitro* dataset is the work of my collaborator Thomas Caradoc Robins, but the inversions and the development of a novel, python-based FWI algorithm are my own work.

The structure of this chapter is as follows. First, the introduction presents research which emphasises the importance of IQA in tomography. In particular, the problem with intuitive image analysis (i.e. clinician lead) is raised, and the potential of automatic image analysis

(i.e. machine learning) to resolve this difficulty is discussed. Second, the performing FWI on real data is complex, so the method introduces a novel numerical simulation framework which was developed for this purpose using PyTorch. Subsequently, *in-vitro* results are presented for a 2.5D poly-vynl acetate phantom and some qualitative relationships between the uncertainty estimates and the image quality of the FWI reconstruction are elucidated.

I have presented these results at the International Ultrasonics Symposium.

O.Bates *et al* (2021) No-cost uncertainty estimation for ultrasound tomography of the human brain. IEEE IUS, Virtual, 2021.

I have also included the results in the second part of a research paper in Inverse Problems.

O.Bates *et al* (2022) A probabilistic approach to tomography and adjoint state methods. Inverse Problems, vol.38, 045008.

4.1 Introduction

Tomographic imaging seeks to render internal anatomical structures visible, a task which is complicated because the quality of a tomographic image is imperfect. Examples of imperfections include blur, artefacts, or a failure to reconstruct altogether [92]. Across the scientific disciplines, uncertainty is the method of choice for quantifying the degree of reliability of a chosen model. This chapter shows that, in the case of tomographic imaging, a pixel-wise estimate of uncertainty is capable of measuring image quality objectively and without relying on reference images. The approach, called variational tomography, is to apply stochastic variational inference, a gradient-based approach for solving Bayes' equation, to full-waveform inversion, a reconstruction algorithm designed for wavefield data such as ultrasound. FWI uses a full numerical simulation of the propagation of a wavefield through the sample, and solves the inverse problem using an adjoint-state method. Therefore, VT is readily applicable to tomography and adjoint-state methods in other fields.

Image quality can be improved by carefully adjusting the settings of a reconstruction. Settings, such as the pass-band of the low-pass filter used in multiscale FWI [75], can cause the reconstruction algorithm to fail when they are not tailored correctly. Calibrating an imaging system currently requires extensive clinical guidance and expert judgement [92]. In addition, line-pair phantoms (e.g. R3L3S1N Negative 1951 USAF Target, ThorLabs, NJ, USA) are often used to measure resolution, but this is misleading because the sharpness, and therefore the resolution, of the reconstruction changes according to the subject in the image. Uncertainty estimation could mitigate these problems [98]; the uncertainty estimator described herein is sensitive to resolution or blur, reconstruction failure, and artefacts. It is also computationally inexpensive, meaning it can be calculated for every reconstruction and that image quality can be continually monitored. In addition, the uncertainty estimates could be used to automatically recalibrate imaging apparatus by seeking to minimise the reconstruction variance.

Pixel-wise uncertainty estimates can help to inform clinicians during image analysis, for example knowing the effective resolution of the reconstruction can inform a clinician of the minimum feature size which is relevant. In addition, the advent of machine-led image analysis increases the need for uncertainty estimates because artificial intelligence algorithms lack human-equivalent perception; they can be easily misled by changes in image quality. Uncertainty estimation might be an effective mitigation strategy because acceptable image quality bounds can be stated in a clear and machine-comprehensible format.

Medical image quality has advanced dramatically in recent decades and high-resolution, high-contrast 3-D imaging is now available. However clinical decision making based on imagery is imperfect, for example Myers *et al* [144] report that mammographic screening programmes have a cumulative false positive rate of between 4% and 10% and that there is significant variation between clinicians. In part, this is because imperfect image analysis can lead to false results. An alternative approach is quantitative imaging, which provides numerical measurements of specific properties of the tissues under observation. Quantitative imaging simplifies image analysis by providing a directly comparative measurement between patients. Unfortunately, modelling physics approximations made during the design of the

reconstruction algorithm can lead to drift between patients. For example, dual energy x-ray absorptiometry (DXA), which measures the response of tissues to x-ray beams with different energy profiles, can drift due to changes in the thickness of the sample [145].

Ultrasound tomography is an emergent 3-D tomographic imaging system which has been studied for imaging the breast [69] and brain [13]. Hopp *et al* [146] explore breast tissue classification according to quantitative sound-speed and absorption measurements. Unfortunately the distributions of fatty, glandular and malignant breast tissues overlap, leading to misclassification if a measurement falls on the wrong side of the decision boundary. This is a common problem with datasets that are not linearly separable, and uncertainty helps by providing a range of values that a pixel can reasonably be expected to take. This means that sound speeds close to the decision boundary can be assigned a probability of belonging in either class, and prioritised for more careful assessment.

This chapter shows that the tomographic uncertainty estimator, which was derived assuming a mean-field Gaussian distribution, is sensitive to changes in the image quality of the tomographic mean estimator. This is demonstrated *in-vitro*, by showing that the tomographic variance reconstruction is sensitive to resolution, reconstruction failure, and artefacts. In detail, the tomographic variance reconstruction reduces as the sharpness or resolution of the tomographic mean reconstruction improves, an effect which I call progressive decoding. The reconstruction variance also increases when the tomographic mean fails to reconstruct when cycle-skipping [15] in introduced. The tomographic variance also identifies an artefact in the tomographic mean reconstruction. These relationships are analysed quantitatively in Chapter 5 and are evidence that the tomographic variance is a useful tool for automatic, no-reference IQA.

Initially, the method describes a Python code for FWI, which is written using the PyTorch library. This code is more simple than typical FWI the algorithms, which are complex as they are developed for Message Passing Interface-based parallel computing. However, the PyTorch code is also fast because the mathematical operations are performed using GPU arrays which native to PyTorch. The details of the *in-vitro* acquisition are also provided. The results contain another practical demonstration of the low computational cost of VT-FWI

and some qualitative applications of the estimator. These applications include sensitivity to resolution, total reconstruction failure, and artefacts. Finally, there is a discussion of some limitations of the results, and how these limitations might be overcome in future work.

4.2 Method

The inversion was performed using in-house FWI software (see Section 2.4) written in Python using PyTorch [131]. The software uses a finite-difference time domain wave equation solver. For the *in-vitro* case, a 890×891 grid with a grid-spacing of $0.2[mm]$ and a time-step of $0.038[\mu s]$ was used.

The dataset contains 384 sources and 1536 receivers. In order to reduce the time taken to invert the dataset, minibatch stochastic gradient descent was adopted with a batch size of 10 (1 epoch = 38 iterations) and a step length of $4[m\,s^{-1}]$. Sources were assigned to each batch stochastically. Previous results suggest that the size of the minibatches does not affect the stability of the inversion and that step length can be optimised by grid-search. Due to a lack of ground-truth, the inversion of the *in-vitro* data was considered to have converged after 76 iterations. The inversions start from a mean-field Gaussian prior with mean $1480[m\,s^{-1}]$ and variance $4[m\,s^{-1}]^2$. The mean is chosen to reflect the sound speed of water and the variance is chosen so that the initial gradient estimate is stable.

4.2.1 The *in-vitro* dataset

Figure 4.1 shows the Poly-Vinyl Acetate (PVA) phantom used for the *in-vitro* experiment. The phantom was produced using moulds, manufactured from the design-image shown in Figure 4.1b. The reconstructions were compared to the design-image using the structural similarity (SSIM) metric [95] to assess their accuracy. The phantom has a width $60.5\,[mm]$, a length $78.8\,[mm]$ and depth $130\,[mm]$. The phantom consists of two layers with different sounds speed [147], which do not vary in depth. Thus the sound-speed varies in the imaging plane but is invariant out of plane. This 2.5D layout prevents out-of-plane effects superimposing on the 2-D reconstruction plane. The transducers rotate in a circle with diameter $150\,[mm]$ centred on the centre of the PVA phantom.

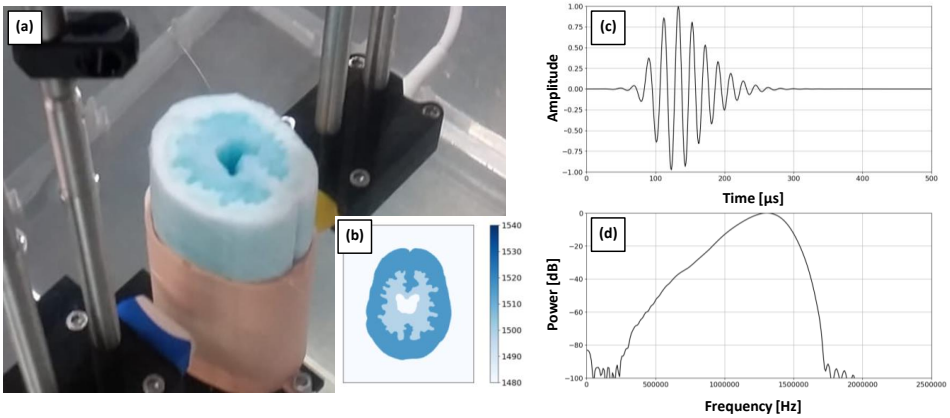


Figure 4.1: (a) The ultrasound tomography apparatus with the PVA tissue-mimicking phantom (centre) surrounded by two P4-1 transducers which are held mechanically in suspension from two rotary motors. (b) The design-image used to produce the moulds for the phantom. The transmitted source wavelet is shown in the time (c) and frequency (d) domains.

The wavefield data were acquired using two P4-1 transducers (ATL, USA), which have 96 elements and were held mechanically in suspension from two rotary motors. Figure 4.1a shows this set-up, with the two transducers on opposite sides of the PVA phantom. Each transducer can be rotated to one of 16 possible positions around the phantom and, during a single acquisition, one of the transducers act as a source while both transducers act as a receiver. The source has a centre frequency of $1.4\,[MHz]$, its excitation signal is shown in Figure 4.1c and its frequency content is shown in Figure 4.1d.

A common cause of reconstruction failure in FWI is cycle skipping. This occurs when the observed and predicted wavefield data arrive more than 90^o out-of-phase and causes the reconstruction to fail. Multiscale methods are a popular way to prevent cycle-skipping [75], and work by low-pass filtering the observed and predicted wavefield data until the pass-band produces filtered data that are in-phase. Therefore, *in-vitro* data are low-pass filtered at $700\,[kHz]$ to prevent cycle-skipping and $1\,[MHz]$ to induce it.

In total there are 384 wavefield data acquisition sequences because 24 sources are used for each of the 16 source positions. A source is a single element on the source transducer. There are 16 possible transducer positions which are located between 0^o and 337.5^o in 22.5^o steps. For a single acquisition sequence, the source transducer is moved to one of the 16

Hardware	Mean update (s)	Standard deviation update (s)
CPU	1200 ± 90	0.00356 ± 0.00004
GPU	122 ± 1	0.014 ± 0.004

Table 4.1: The time taken to compute the mean and standard deviation update on different devices for a batch size of 10.

positions and fixed in place. The receiver transducer is then moved to one of 11 positions. For example, if the the source position is at 0^o, the 11 receive positions are between 45^o and 315^o (in 22.5^o steps). The 4 remaining positions (2 either side of the source transducer) are left unoccupied to prevent the transducers colliding. The source fires the excitation signal and the wavefield data are collected at both the source and receive transducer. Then, the receive transducer is moved to another position and the source is fired again (with data collected at both transducers again). Once the receiver has moved to all 11 positions, the acquisition sequence for that source is complete. Therefore, there are $1,152$ receivers (96 per position, 12 positions when the source transducer is included). Now the source changes, either another of the 24 elements at the current position, or one of the 24 elements at a new position is used. Wavefield data are acquired for each of the 384 sources, at which point the full acquisition process is complete.

This constitutes the data acquisition process for the *in-vitro* data-set. This data-set is used to investigate image quality changes due to progressive decoding (i.e. a resolution improvement), reconstruction failure, and a source position artefact. To investigate the sensitivity of the estimator to image quality issues caused by discrepancies between the true physics and the physics of the inverse problem, an *in-silico* data-set is used.

4.3 Results

To demonstrate our claim of no-cost uncertainty estimation, we measured the computation time of the mean estimator (Equation 3.54) and the standard deviation estimator (Equation 3.53) during the *in-vitro* reconstruction. Table 4.1 shows the time required to execute the mean and standard deviation update on a CPU (Ryzen Threadripper 2950X 16 core, Advanced Mirco Devices, CA) and a GPU (Quadro RTX 8000, Nvidia, CA). Compute time

is an average of 10 iterations.

In the case of the GPU, the standard deviation update takes approximately 0.01% of the computation time of the mean update. This difference is more pronounced in the case of the CPU because the compute time for the mean update is longer, but the standard deviation update is shorter. Again, this is likely because of the time taken transferring data between the CPU and GPU, which negates the parallel computing power of the GPU. It is clear that the uncertainty estimator has effectively no computational cost. Comparing Table 3.1 & 4.1 shows that the CPU mean update is slower in the *in-silico* case, but the GPU mean update is faster.

As previously introduced (in Section 3.3), the mean-field Gaussian uncertainty estimator is not a true representation of the pixel-wise variation of the mean. However, this estimator does show useful behaviours. Specifically, we demonstrate with the following results that the estimator can be used for progressive decoding, to detect reconstruction failures, artefacts, and to identify physical modelling discrepancies.

4.3.1 Progressive decoding

Figure 4.2 shows results for an inversion of the *in-vitro* data-set following low-pass filtering at 700 $[kHz]$. Figure 4.2d shows the reconstruction after 76 iterations (2 epochs). The SSIM was found to be 0.83, indicating a high degree of similarity between the design-image and the final reconstruction. Figure 4.2a demonstrates the stability of the uncertainty estimator by repeating the reconstruction process 20 times and plotting the average of the variance per reconstruction per iteration for the 76 iterations. This shows that for 18 runs, the variance consistently converges to a variance of approximately 2 $[m\,s^{-1}]$. Note the two runs with the highest mean variance (approximately 3 $[m\,s^{-1}]$) contain artefacts which are discussed in Section 4.3.3.

Descriptively, Figure 4.2a shows that the average variance undergoes a rapid increase, followed by a similar decline, and then a consistent improvement until convergence after 76 iterations. The average variance for the reconstruction process presented in Figures 4.2b - 4.2d is indicated in cyan. Figure 4.2b shows that by iteration 10 the mean has not

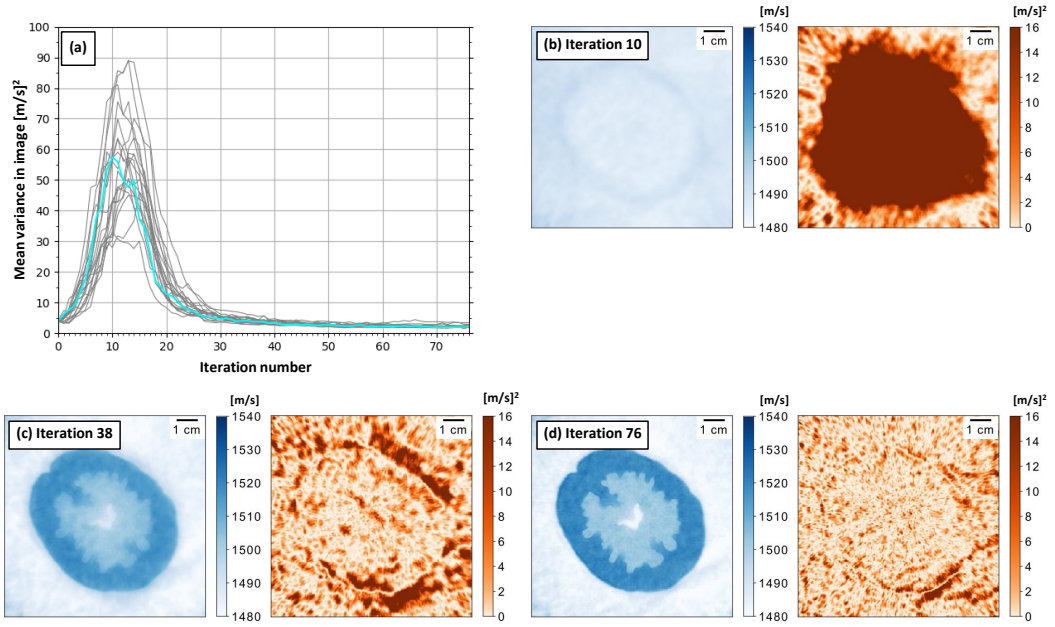


Figure 4.2: Reconstructions using VT-FWI and the *in-vitro* dataset, low-pass filtered at 700 $[kHz]$. (a) The average variance per iteration for 20 different reconstruction processes. (b)-(d) Examples of the mean (left) and variance (right) reconstructions corresponding to the process indicated by the cyan line in (a).

reconstructed to a plausible sound speed. This corresponds to high variance of 63.34 $[m\,s^{-1}]^2$ on average, which is localised where the mean is incorrect. Figure 4.2c shows that by iteration 38 (1 epoch) the mean reconstruction is approximately correct but blurred. The uncertainty estimate is significantly reduced to 3.74 $[m\,s^{-1}]^2$ on average. Figure 4.2d shows the reconstruction by iteration 76 (2 epochs), the mean is much sharper and the average variance has again reduced to 2.07 $[m\,s^{-1}]^2$.

Interestingly, the variance reconstruction in Figure 4.2c contains low-frequency correlated noise and the variance in Figure 4.2d contains high frequency correlated noise. This is related to the covariance between the pixels. Covariance is an extension of the concept of variance; instead of measuring the variation of a single pixel around its mean, covariance measures the common variation between two pixels. The mean-field estimator assumes that the pixels are uncorrelated and therefore have no covariance. The presence of correlated noise indicates a failing in this assumption, and the length-scale of the correlation could indicate the length-scale of the missing correlations.

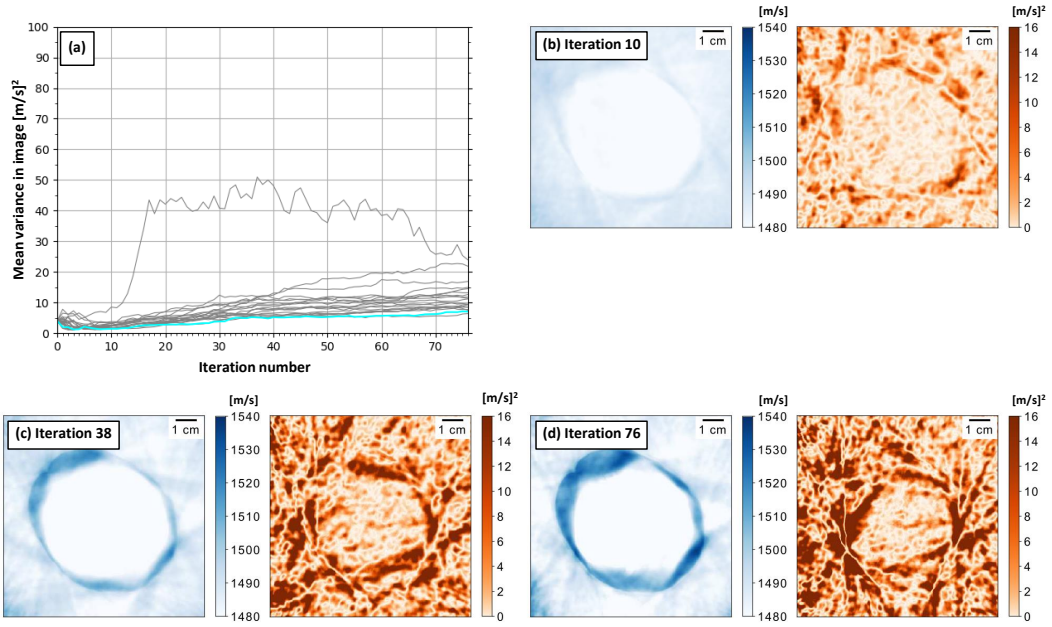


Figure 4.3: Reconstructions using VT-FWI and the *in-vitro* dataset, low-pass filtered at $1\,[MHz]$. (a) The average variance per iteration for 20 different reconstruction processes. (b)-(d) Examples of the mean (left) and variance (right) reconstructions corresponding to the process indicated by the cyan line in (a).

4.3.2 Detecting reconstruction failure

Figure 4.3 shows results for the inversion of the *in-vitro* data-set following low-pass filtering at $1\,[MHz]$. Figure 4.3d shows the reconstruction has failed to reconstruct the expected sound-speed image. The SSIM was found to be 0.41, indicating a low degree of similarity between the design-image and the final reconstruction. In Figure 4.3a the variance starts low, but increases consistently as the reconstruction process progresses. This is significantly different to the $700\,[kHz]$ case. The average variance for the reconstruction process presented in Figures 4.3b - 4.3d is indicated in cyan. Figure 4.3b shows that by iteration 10, the mean estimate of the sound speed has reduced. Figure 4.3c shows that by iteration 38 (1 epoch), the mean estimate remains poor and the variance estimate has increased significantly. Figure 4.3d shows that the variance continues to increase as the number of iterations increases to 76. Note that Figure 4.3 shows reconstructions which were selected from the trend-line with the lowest average variance, the final variance is higher for the other 19 reconstruction

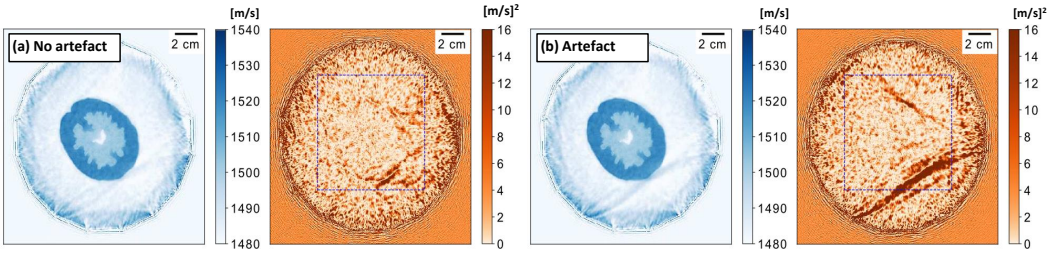


Figure 4.4: (a) An uncropped replica of the result in Figure 4.2(d), where the crop-box is indicated by the blue dotted square. (b) Another example reconstruction produced using the 700$[kHZ]$ filtered data, but in this case the posterior reconstruction contains an artefact due to source positioning error.

processes.

After iteration 76 (2 epochs), the variance of the 20 repeats at 1 $[MHz]$ are consistently higher than the equivalent 700 $[kHz]$ examples. In addition, the mean reconstruction in Figure 4.3d is indicative of cycle-skipping, which is a known cause of reconstruction failure in FWI [15]. Finally, comparing Figure 4.2a with Figure 4.3a shows that the average of the variance is more unstable in the 1 $[MHz]$ case. The average variance in these figures is also trending in different directions; in the 1 $[MHz]$ case the variance increases regularly, while it decreases in the 700 $[kHz]$ case. Therefore, these plots show that uncertainty is consistently higher and increasing when the *in-vitro* data is cycle-skipped.

4.3.3 Artefact detection

In Figure 4.2a two of the reconstruction processes produce a higher average variance than the others. Figure 4.4b shows one of these two runs. Figure 4.4a is a repeat of Figure 4.2d, but now the reconstruction has not been cropped. The blue boxes in Figure 4.4 indicate the crop used in Figures 4.2 and 4.3.

It is clear from Figure 4.4b that the high average variance is caused by a single region. When we compare the mean reconstruction of Figure 4.4b to the mean reconstruction in Figure 4.4a, it is clear that the region with high variance corresponds to an artefact. This artefact is caused by an intermittent source-receiver positioning error (see Section 4.4.1 for details).

4.4 Discussion

Stochastic variational inference was developed to stabilise back-propagation through variational auto-encoder architectures which are based on artificial neural networks [112], [139], [140]. Stochastic variational inference has since been applied to Bayesian Neural Networks (BNN's) [114], where each node in a layer is parameterised by a Gaussian distribution, and Gaussian processes [148], which assume a Gaussian distribution of functions. Similarly, this paper describes a Gaussian process which uses the wave-equation as the link function between the sound speed (i.e. the parameter space) and the wavefields (i.e. the function space). There are also connections between our approach and BNN's, because neural networks and numerical differential equation solvers are connected [115] [149].

4.4.1 Comments on the results

In Chapter 3 the computational cost of the uncertainty update on a GPU was found to add 0.007% to the total computation time. Similarly, in this chapter, the uncertainty update on a GPU was found to add 0.01% to the total computation time. Simularly, in both cases, the uncertainty update was faster on a CPU. Therefore, the results in this chapter confirm that the uncertainty estimator is consistently computationally efficient because it costs less than *one ten-thousandth* of the computation time of the FWI mean update.

A less clear finding was identified when comparing the computation time for mean update *in-silico* and *in-vitro*. Tables 3.1 & 4.1 show that the *in-vitro* mean update is faster on a CPU but slower on a GPU. This is likely to be an artefact of the numerical simulation algorithm. The two most time-consuming operations are the calculation of the Laplacian of the wavefield and the extraction of the wavefield data from the wavefield. It is likely that the interplay between these operations leads to the observed behaviour, but further analysis is required to understand the exact cause. However, because the computational cost of the uncertainty update relative to the mean update is stable, the actual computation time of the mean update is not relevant to the fundamental observation that VT uncertainty estimation is computationally efficient.

Finally, our results demonstrate that the parameters of an optimisation process using

the pathwise derivative can be tailored such that the inversion process is consistently stable using *in-silico* and *in-vitro* observed data. This is an advantage over the stochastic gradient Langevin dynamics method, which requires a step-size schedule that is difficult to tune [137].

Progressive Decoding The results in Figure 4.2 demonstrate that the mean-field variance estimate is a useful metric for measuring the resolution of the reconstructed image during an iterative reconstruction process. It is significant that this result is achieved using an experimental *in-vitro* data-set because it demonstrates the fundamental strength of the approach, namely, stochastic variational inference is immediately applicable to tomographic reconstruction (and adjoint methods) without requiring additional effort to work with real data. It is also unclear whether similar robustness will be achieved by the alternative methods discussed previously, many of which have only been demonstrated on synthetic data.

Detecting reconstruction failure The mean-field Gaussian estimator is a clear indicator of reconstruction failure. When a mean reconstruction is produced from cycle-skipped data, the variance reconstruction is a significantly higher than the case where the mean reconstruction converges correctly.

The results suggest that while average variance is useful, the local variance is less reliable for detecting cycle-skipping. This effect is a result of the mechanics of cycle-skipping. The cycle-skipping appears to cause the reconstruction to diverge, meaning that the mean update (Equation 3.54) consistently reduces the value of the sound speed in the centre of the reconstruction. When the multiplication operation is applied to obtain the variance update (Equation 3.53), approximately half the samples have their direction reversed because sampling the standard normal distribution produces roughly equal quantities of positive and negative samples (the mean of the standard normal distribution is 0). Because the mean updates are consistently negative in the centre of the reconstruction, then roughly half the variance updates are positive and, therefore, the variance update is small. A lack of covariance information might also be hindering the quality of the estimate in this region. Figure 4.3b shows that in the early iterations the sound speed of the entire reconstruction reduces simultaneously, which means that the covariance between pixels would be be high. When the direction of the mean update at the edge of the phantom becomes positive, the magnitude

of the variance update also increases locally. If the covariance between the outer and inner parts of the phantom were initially high, this could cause the variance in the central region to increase.

This line of reasoning also leads to an interesting interpretation of Figure 4.2. Initially the mean update is far from the ground truth, which causes some of the samples from the model distribution to be cycle-skipped while others are not. In turn, this causes the variance to increase rapidly in the region where cycle-skipping occurs, this is observed in Figure 4.2b. The variance will reduce as the mean reconstruction converges, because now most of the samples from the model distribution occur in the locally-convex part of the objective function, this is observed in Figure 4.2c. The general trend shown in Figure 4.2a also follows this pattern.

Artefact detection Our results indicate that locally high variance estimates can indicate artefacts. I expect the detection of artefacts to be consistent regardless of their presentation, which is validated by results in Chapter 5. This is because artefacts are driven by epistemic errors between the gradients produced from different source-receiver pairs. Above, it is explained that high variance is caused by epistemic errors in the data (e.g. cycle-skipping) so, where these gradients conflict, high variance is inevitable.

4.4.2 Usefulness

In many applications, MCMC is considered to be the most effective method to estimate uncertainty. However, in high dimensional problems, MCMC algorithms are often numerically intractable. We have developed an uncertainty estimator which assumes a mean-field Gaussian distribution over the model parameters. This approach is simple, very computationally efficient, and will be applicable to problems where uncertainty is not usually estimated.

It is appropriate to note that the mean-field Gaussian approximation of the posterior relies on a simplistic approximation, but this approximation means that the posterior can be calculated quickly and efficiently. Accurate-but-intractable estimators such as MCMC are also flawed, but in a different way. Furthermore, Sections 3.2.1 and 3.3 show that the approximations associated with the mean-field Gaussian estimator are consistent with the

approximations used by many gradient-based solvers. Therefore, the mean-field Gaussian uncertainty estimates may be accurate for a variety of different problems. Thus, this thesis proposes that uncertainty estimates should be generated where possible and evaluated based on the information that they can provide to a particular problem.

To illustrate the potential of the tomographic uncertainty estimator, some applications to image quality assessment in FWI tomography were demonstrated. Comparisons with other methods are avoided because different approximations are appropriate for different problems and their effects are difficult to anticipate. In future, a comparative study will be performed because is be important to quantify the accuracy of the estimator for specific applications. However, findings of such a study will not hold generally, meaning an application specific quantitative study might be misleading.

Progressive decoding A measurement of the resolution (or sharpness) of the reconstruction has a number of important applications. For example, FWI is a computationally expensive reconstruction algorithm which is slow to converge but, with the variance estimator, it is possible to adapt the stopping point after achieving sufficiently high reconstructed resolution or, equivalently, sufficiently low average variance. This is the simplest way to implement a dynamic reconstruction algorithm. Another application is to provide preliminary reconstructions alongside an estimate of the resolution of the current image. This is useful because a clinician can be presented with an intermediate result which has a low resolution and they can identify larger features in the image. Then, the final result can be presented when available.

During their treatment, a patient may be monitored using a variety of different imaging systems that achieve different image qualities. Variance estimates can simplify comparisons between reconstructions by quantifying resolution and identifying artefacts. For example, an objective measure of resolution makes it possible to make volume measurements with specific error bounds and the pixel-wise estimate means variation in resolution across the image can be taken into account.

Detecting reconstruction failure The capability to automatically identify reconstruction failure enables several applications. First, many imaging systems have settings which

need to be optimised in order to improve the system performance. The pass-band of the low-pass filter for multiscale FWI is one such example. With the variance estimate, it is possible for a non-expert to set the pass-band frequency without relying on subjective judgement. Second, our method also enables error detection for each reconstruction. Unexpected issues, such as the presence of metallic implants, might cause cycle-skipping or obstruction in part of the reconstruction. With our method, the pass-band can be tailored until the reconstruction performs correctly, or the reconstruction can be conclusively identified as having failed. We expect the detection of cycle-skipping to be particularly useful in Geophysics.

Artefact detection Artefacts are an important problem in medical imaging. In tomography, classic examples of artefacts are metallic streaking artefacts in x-ray CT, and movement artefacts in MRI. Figure 4.4 illustrates an intermittent source position artefact in USCT. Artefacts obstruct parts of the image, meaning that they reduce diagnostic capability by reducing the visibility of important features. Artefacts can be subtle, meaning that they can mislead clinicians and lead to false positives or measurement errors (i.e. they can distort an image by changing the magnitude, shape, or size of the feature of interest). Worse, automatic diagnosis by artificial intelligence methods can be severely misled by artefacts, and artefacts are challenging to detect by pure feature detection because they have a wide variation in their presentation. The robust detection of artefacts will become increasingly important as automatic diagnosis becomes more common in medical imaging.

4.4.3 Future work

In addition to the comparative study already discussed, this paper has a number of important extensions. Most prominent, is to perform a quantitative analysis which will confirm these qualitative observations regarding image quality. The *in-vitro* results, presented in this chapter, indicate that tomographic uncertainty is sensitive to resolution, reconstruction failure, and artefacts. Therefore, Chapter 5 presents a quantitative *in-silico* analysis, where tomographic uncertainty is compared to the absolute error between the tomographic mean reconstruction and ground truth reference. In addition, a new IQA metric (mean standard deviation magnitude) is compared to the mean absolute error. The quantitative

analysis is performed for in the context of 2 important limitations, assumptions regarding the reconstruction physics and transducer misplacement, which induce artefacts and degrade resolution and, therefore, reflect the *in-vitro* results shown previously.

Another important direction is to investigate the use of prior information in the reconstruction problem. Interestingly, this complements another topic, which is to incorporate covariance information into the reconstruction. One approach to estimate correlation is to assume a parametric relationship between the pixels of the reconstructed image (e.g. assuming a full Gaussian distribution). However, this approach is likely to reduce the quality of the variance estimate by enforcing nonphysical correlations between pixels. For example, a fully-specified Gaussian tends to prevent sharp changes in contrast such as the change between the skull and the brain. An alternative approach is to use a non-parametric method and learn the correlation function during the reconstruction process. The archetypal non-parametric method is a convolutional artificial neural network, suggesting that combining our pathwise derivative approach with the deep-image-prior could be interesting. Prior information is naturally incorporated into a model-based framework because the correlations must be learnt from a database of example reconstructions *a priori*. Thus we will study the data-based priors to correctly solve Bayes' equation and to regularise ultrasound tomography of the brain.

Finally, the variance reconstruction is available per-iteration, meaning that the estimates could be used to produce adaptive or dynamic reconstruction algorithms, where the optimal choice of data is identified for the next gradient calculation. In FWI, this would consist of choosing the best source-receiver pairs in order to adjust the model parameters where the variance is highest. In MRI, the very low cost of the variance estimates means online reconstruction might be possible. Hence, the variance could be used to guide the data acquisition process. A particularly useful implementation could be to combine these estimators with reinforcement learning in order to achieve these aims.

Chapter 5

Variational tomography: Quantitative validation *in-silico*

Chapter 3 presented my first novel contribution; a Bayesian model capable of no-cost uncertainty estimation. Chapter 4 presented my second novel contribution; evidence of a qualitative relationship between uncertainty and image quality *in-vitro*. This chapter presents my third contribution; evidence of a quantitative relationship between the VT uncertainty estimates and absolute error. To do this the mean standard deviation magnitude (MSDM), a new quantity determined from the VT uncertainty estimate, is compared to the mean absolute error (MAE), a well established full-reference image quality metric. The quantitative relationship between MSDM and MAE is important evidence that uncertainty can be used for routine objective IQA in US-CT. In addition, the quantitative relationship validates the qualitative observations which were presented in the previous chapter. Once a quantitative relationship between tomographic uncertainty and IQA is realised, then the applications such as mitigating confounding during image analysis and adaptive reconstruction become possible.

I have presented these results at the International Ultrasonics Symposium.

O.Bates *et al* (2022) IEEE IUS, Venice, 2022.

5.1 Introduction

The quality of medical images is difficult to determine objectively. In part, this is because high image quality is not equivalent to high diagnostic accuracy. For example, beamformed ultrasound is known to have very poor image quality for lung imaging, but is still a useful diagnostic tool [150]. However, image quality assessment remains desirable because it can be used to evaluate improvements in the imaging apparatus or reconstruction methods. Furthermore, high image quality is desirable to increase clinical confidence in a diagnosis or treatment plan.

Artefacts are a critical problem for medical imaging. They degrade image quality and, because they are often localised, they can have a severe effect on diagnostic accuracy [92]. Artefacts arise due to epistemic errors between the modelled and real-world behaviour of the apparatus, reconstruction algorithms, and patient. Apparatus and reconstruction related errors are similar because they are caused by the poor characterisation of the imaging system or reconstruction physics. Examples include magnetic field inhomogeneity in MRI [151], and beam hardening or dead pixels in CT [152]. Similarly, patient related errors are caused by poor characterisation of the behaviour of the patient; for example, patient movement during the data acquisition [151].

A more general image quality problem is to determine the resolution, contrast, and noise of the reconstructed image [153]. It is desirable that images have the highest possible resolution because the resolution determines the minimum size of feature that the system is capable of resolving; for example, if the system has insufficient resolution, then it is possible for a clinician to miss features of the image associated with disease. Typically, resolution would be determined by the point spread function of the imaging system [154]; for example, the point spread function in x-ray imaging systems is determined using line pair phantoms (e.g. R3L3S1N Negative 1951 USAF Target, ThorLabs, NJ, USA). A related problem is image contrast. Poor contrast occurs when the relative intensity of the diseased feature is not significantly higher than the background intensity. A classic example is the difficulty

with using mammography to detect tumours in women with dense breasts [36], [37].

A variety of methods have been developed for image quality analysis (IQA). Chow *et al* [92] classify IQA methods as subjective or objective, depending on whether they are based on human judgement or use only mathematical analyses. These categories are further subdivided as full-reference or no-reference methods. According to this classification system, the variance or standard deviation reconstruction from VT is an objective, no-reference image quality measure.

Subjective methods use human perception to evaluate image quality. One example is the Double-Stimulus Continuous Quality Scale, originally developed for assessment of televisual quality [92], but repurposed for ultrasound imaging [155]. The technique works by presenting an assessor with two images; a reference image and an impaired image. The assessor is not told which image is impaired and they are required to assess the quality of each image on a continuous scale. Thus, the quality of the impaired image can be determined in comparison to the value of the reference image. In one sense, subjective methods are the best possible approach to IQA because human judgement is usually required to evaluate images and form a diagnosis. The drawback of this approach is that subjective assessment requires expertise, is time consuming, and inconvenient. In addition, humans have strong perceptual biases, an effect which has also been verified in studies of individual variations in diagnostic ability [36]. These biases are a crucial factor motivating the development of machine-led image analysis [70], and are a strong motivation to develop routine objective IQA.

Objective methods do not involve human judgement in the IQA process. Instead, algorithmic methods are used to determine image quality. Almost all objective medical IQA involves ground truth reference images [92]. The simplest algorithm is mean square error. This takes the difference between the distorted and reference image pixel-wise and squares these values to obtain a magnitude. Then the value of each pixel is summed and the final values obtained by division with the number of pixels. Similarly, mean absolute error takes the sum of the magnitude of the difference between the distorted and reference images. Unfortunately, mean squared error and mean absolute error are limited in their fidelity to the human visual system [93], [94]. These problems motivated the development of more sophist-

icated methods including structural similarity [95] and the perceptual difference model [96]. Both methods are more effective than the mean square error and mean absolute error, but they are limted by a requirement for a ground truth reference image. In clinical practice, reference images are never available.

The inaccessibility of reference images led to the development of no reference image quality assessment, although few examples exist for medical imaging [92]. Machine learning is prominent among modern IQA methods. For example, a deep convolutional neural network has been demonstrated to detect motion artefacts on MRI scans [91], while similar work on natural images suggests this technique is useful in general [97]. The main difficulty of this approach is the use of a classification output layer, where each node represents a IQA metric. This introduces an explainability problem, which is common in deep learning and means the algorithm is unable to localise sources of failure in a user interpretable format. Other approaches include classifying potentially impaired images using pre-trained support vector machines [156], and a channelized Hotelling observer [157], which processes the potentially impaired image using filters with frequencies that mimic human perception and determines image quality using statistical analyses [158].

Finally, statistical methods have been proposed to for objective no reference IQA. Dutta *et al* [98] investigate three methods; a Monte-Carlo method, a deterministic expectation maximisation algorithm [105], and a fixed point analysis based on the Hessian [106]. These estimators are able to produce pixel-wise variance estimates and associate them with image quality, but the computational cost of these techniques continues to impede their use.

This chapter demonstrates that tomographic uncertainty estimates from the VT algorithm are suitable for objective, no reference IQA. In a sense, this is similar to the statistical methods for IQA discussed above. However, the VT uncertainty estimator makes different assumptions about the distribution of the posterior, which means it is unclear whether the VT uncertainty estimator is valid. A quantitative analysis compares a novel VT-derived quantity, mean standard deviation magnitude (MSDM), to mean absolute error (MAE), an objective full-reference IQA metric which is well-established. This analysis shows that MSDM and MAE are linearly related, which means that MSDM is suitable to

replace MAE in tomographic applications. In addition, MSDM is more useful than MAE for tomographic applications because it does not require a reference image, which is never available in clinical practice.

In the Method (Section 5.2) the MSDM metric is derived from the VT standard deviation update, which is described by Equation 3.53. Then, MAE is derived based on a method of moments definition of standard deviation. This result shows that MSDM and MAE seek to measure the same quantity. In addition, the details of the simulations are described. The Results (Section 5.3) show that the theoretical relationship between MSDM and MAE is also quantitative relationship, which can be observed in simulated results. The relationship between MSDM and MAE is evaluated in the presence of artefacts, which are induced by physics approximations and epistemic errors in the transducer geometry, and reductions in resolution, which are induced by aleatoric errors in the transducer geometry. The chapter finishes with a discussion including some comments on the results, their usefulness, and some directions for future research.

5.2 Method

The objective of this chapter is to assess the usefulness of the mean standard deviation magnitude for quantitative IQA, by comparing it to the mean absolute error. These comparisons are made for reconstructions which are degraded by artefacts or have different resolution. The study is performed *in-silico* for two reasons. First, it is difficult to synthesise phantoms with the physical properties required for the study. For example, the study uses circular phantoms, which have different densities but the same sound speed. This behaviour is challenging to design *in-vitro*. Second, an *in-vitro* acquisition always contains errors relative to simulated predicted data. For example, errors in the characterisation of the transducers position are impossible to correct [132]. The strategy used in this chapter is to build an "inverse crime" model, where the observed data (in this case obtained by simulation) is produced using the same numerical simulator and physics model as the predicted data, and then to introduce specific errors into the simulation of the observed or predicted data. Thus, the effect that the individual errors have on the mean and standard deviation reconstructions

can be evaluated. This analysis is not possible *in-vitro*.

5.2.1 Image quality analysis metrics

Equation 3.53 shows the standard deviation update in the VT reconstruction algorithm. Iteratively applying this update produces a standard deviation reconstruction when the iterative process terminates. Chapters 3 & 4 use a tomographic variance reconstruction for qualitative analysis. The tomographic variance reconstruction is the pixel-wise square of the standard deviation reconstruction, such that,

$$\Sigma^2 = \left[\sum_{(l)} \left(\epsilon^{(i)}\right)^T_{(l)} \left(\Delta_m\right)_{(l)}\right]^2 \tag{5.1}$$

where Σ^2 is the tomographic variance, (l) is the iteration number, $\epsilon^{(i)}$ is a vector of samples from the standard normal distribution, and Δ_m is the FWI gradient shown in Equation 3.54. Thus the variance reconstruction is found by the sum of squares of the standard deviation once the iterative process has finished.

The standard deviation magnitude is a better quantitative image quality estimator because it has an intuitive meaning; the standard deviation is the width of the Gaussian distribution at half its maximum value. Therefore, the standard deviation magnitude reconstruction (plus or minus the mean reconstruction) represents range in which 68% of sound speed values are expected to reside, assuming the mean-field Gaussian distribution is correct. Mathematically the standard deviation magnitude reconstruction is,

$$\Sigma = \sqrt{\left[\sum_{(l)} \left(\epsilon^{(i)}\right)^T_{(l)} \left(\Delta_m\right)_{(l)}\right]^2} \tag{5.2}$$

and the mean standard deviation magnitude (MSDM) is found by taking the mean of the standard deviation magnitude reconstruction over all the pixels in a chosen region.

The mean absolute error (MAE) is a direct analogue of the mean standard deviation magnitude. This becomes clear by considering the definition of the standard deviation of several mean reconstructions,

$$\Sigma = \sqrt{\sum_{l} \left[(X_l - \mu)^2 \right]} \tag{5.3}$$

where X_l is an instance of the tomographic mean reconstruction, the sum is taken over l tomographic mean reconstructions, and the μ is defined,

$$\mu = \frac{1}{n} \sum_{l} X_l \tag{5.4}$$

where n is the total number of mean reconstructions. By taking a single mean reconstruction and replacing μ with the ground truth image X_{GT}, the standard deviation becomes,

$$\Sigma = \sqrt{(X_l - X_{GT})^2} \tag{5.5}$$

which is the absolute error. The MAE is found by taking the mean of the absolute error for the pixels in a chosen region. Therefore, MAE is directly analogous to MSDM.

5.2.2 The circle *in-silico* dataset

Stride [143] is chosen as the reconstruction engine because it implements variable density and absorption physics. In addition, Stride allows efficient implementations of US-CT codes across distributed systems, with on-chip optimisation provided by Devito [159]. Acoustic simulations are governed by the equation,

$$\frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} = \rho \nabla \cdot \left(\frac{1}{\rho} \nabla p \right) + \eta \frac{\partial}{\partial t} \left(-\nabla^2 \right)^{\frac{y}{2}} p \tag{5.6}$$

where c is sound speed, p is pressure, ρ is the density, $\eta = -2\alpha_0 c^{y-1}$ is a parameter containing the absorption coefficient a_0, and y controls power of the frequency dependence. In this chapter, the inverse problem is always solved using the constant density, zero absorption approximation, which means that Equation 5.6 reduces to d'Alemberts operator given in Equation 3.1. In contrast, the forward problem is solved with perturbations in density, absorption, or source position.

Similar to Gauthier *et al* [74], Figure 5.1 shows a circular phantom, 5$[cm]$ in diameter.

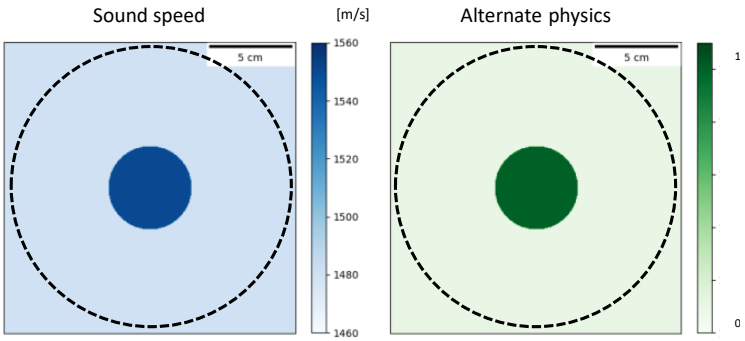


Figure 5.1: A circlular phantom is used for the quantitative evaluation which has (a) a sound speed of 1550$[ms^{-1}]$. (b) The circle can have variable alternate physics, for example density or attenuation. The black circle indicates the location of the transducers.

Figure 5.1(a) shows a finite-difference grid of 350×350 pixels, with a $0.5[mm]$ grid spacing. The sound speed of the circle is approximately that of soft tissue at $1550[ms^{-1}]$ and the background was approximately water at $1480[ms^{-1}]$. The sound speed was the same for all forward simulations. The adjoint method assumes that sound speed varies but that other physical parameters (e.g. density, absorption coefficient) are constant.

Figure 5.1(b) shows the alternative physical parameter. The alternative physical parameter is variable in the forward model but assumed to be constant in the adjoint model. This introduces an epistemic error into the inverse problem because the assumption of constant alternative physical parameter is not correct. This epistemic error causes artefacts, which are common throughout medical imaging. Two alternative physical parameters are evaluated; density and absorption coefficient. When density varies, then the absorption coefficient is $0[Np\ m^{-1}]$. When the absorption coefficient varies, density is constant at $1010[kg\,m^{-3}]$. The alternative physical parameter of the circle is varied systematically for a range of values, where each value represents an approximation of water, soft tissue, trabecular bone, and cortical bone. The values were chosen based on the IT'IS database of tissue properties [29].

For the density the values are,

$$\rho = 1010, 1100, 1200, 1300, 1500, 1700, 2100\ [kg\,m^{-3}]$$

For the absorption coefficient the values are,

$$\eta = 0, 4, 8, 10, 15, 20, 27 \; [dB \; cm^{-1} \, @ \, 500 \, kHz]$$

Erroneous transducer position and patient motion are also common sources of error. These are related because, in US-CT, patient motion presents as a relative shift in the positions of the sources and receivers. Therefore, if the uncertainty estimate is sensitive to transducer position errors, it is expected to also be sensitive to patient motion. In this case, the transducer error was applied to the adjoint problem.

The transducer position error was chosen epistemically or aleatorically at runtime. In the epistemic case, transducer ID's 0-6 were systematically shifted up 1[mm] vertically, while transducers 168-174 were systematically shifted down 1[mm] vertically. In the aleatoric case, each transducer was moved in a randomly in the horizontal and vertical directions using samples from a Gaussian distribution. To measure the quantitative effect of the position error on the variational tomography reconstructions, a range of aleatoric sampling distributions were used. The values of the standard deviation of the aleatoric sampling distribution are,

$$x_{\varsigma} = 0, 0.1, 0.2, 0.3, 0.4 \; [mm]$$

The remaining simulation parameters were chosen to closely reflect an in-vitro acquisition. There were 256 sources distributed evenly on a ring of diameter 16 $[cm]$ and each source has 256 receivers. The source wavelet was a 3-cycle tone-burst with a centre frequency $f_c = 500 \; [kHz]$. The numerical simulation used 2500 time-steps with a time-spacing $dt = 0.08[\mu s]$. A multiscale method was not used for the adjoint model, but a single low-pass filter was applied to the source wavelets and observed data, where the pass-band was set at 500 $[kHz]$ to prevent cycle-skipping. Typically, each gradient was calculated for a batch of 128 sources (1 epoch = 2 iterations), but occasionally 16 sources (1 epoch = 16 iterations) were used. Sources were assigned to each batch stochastically. When 16 source batches are used, the results clearly indicate this. For the mean and variance reconstructions, the effect of changing the step-length was evaluated for steps of $2, 4, \text{or} \, 7$. As usual, source

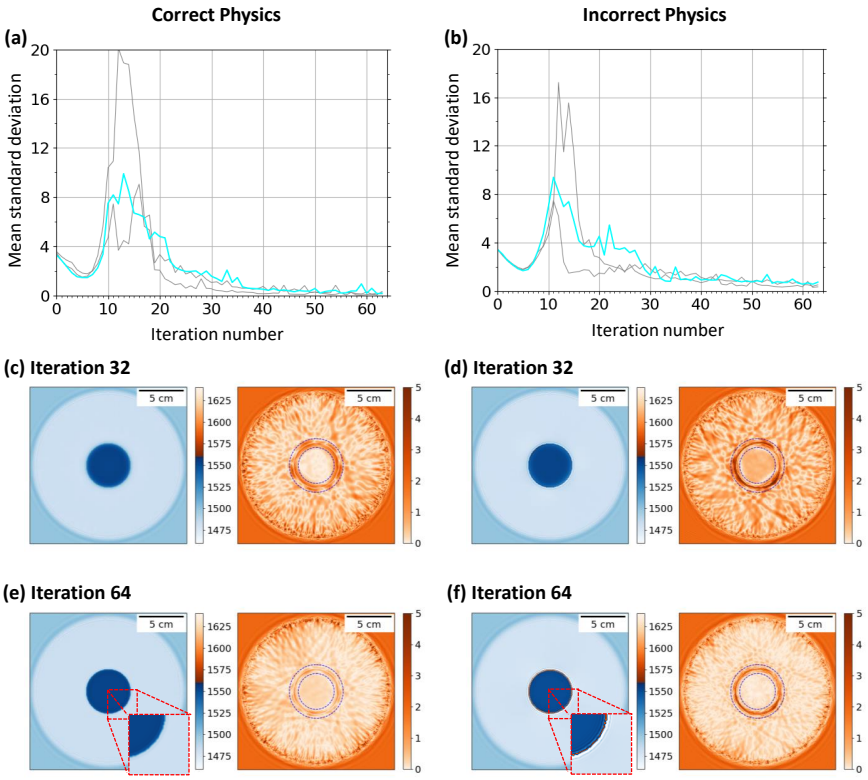


Figure 5.2: VT-FWI reconstructions of the circle dataset. MSDM per iteration is shown for (a) constant density of 1010 $[kg\,m^{-3}]$ and (b) variable density where the circle phantom is 1700 $[kg\,m^{-3}]$ for 3 different reconstruction processes. (c-f) reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue ring, indicating the region of interest.

normalisation is applied for all the adjoint source calculations.

5.3 Results

5.3.1 Density

The FWI adjoint method (i.e. the reconstruction process) assumes homogeneous density. However, Figure 5.2 shows two cases. First, when the physics of the FWI adjoint method is correct; in other words, the reconstruction truly has homogeneous density and therefore the circle has the same density as the surrounding water. Second, when the physics of the FWI adjoint method is incorrect and, therefore, the density of the reconstruction region is

inhomogeneous. In the inhomogeneous case, the circle has a density of 1700 $[kg\,m^{-3}]$ but the surrounding water has a density of 1010 $[kg\,m^{-3}]$.

The MSDM per iteration plots (Figure 5.2a and 5.2b) show behaviour characteristic of an accurate reconstruction process; the standard deviation rises, falls, and settles to a low value once all the data from all the sources have been used. However, comparing Figure 5.2(a) and Figure 5.2(b) shows that the MSDM is slightly higher after 64 iterations when there is a mismatch between the physics of the FWI adjoint method and that of the observed data. The relative increase in the standard deviation in Figure 5.2(b) is because the change in density increases the reflections and reduces the transmissions from the circle compared to the homogeneous case assumed for FWI.

Figures 5.2(c) & (d) show that it is not possible to distinguish the effect of density on the standard deviation reconstruction after the first 32 iterations. In Figure 5.2(f), the adjoint method assumes constant density, but the forward problem has variable density. In this case, the inhomogeneous density increases reflection and reduces transmission through the circle, which FWI explains by increasing the sound speed at the boundary of the circle in the mean reconstruction. This effect also creates a ring of high standard deviation in the boundary region. In contrast, Figure 5.2(e) shows the reconstructions when the adjoint method is true to the physics of the observed data. Now, it is apparent that the sound speed artefact no longer present and the standard deviation in this region is much less prominent.

Additional results (not shown) indicate that after 128 iterations the variable density artefact to continues to increase in magnitude, but the magnitude of the standard deviation decreases without completely disappearing. This happens because the high sound speed at the boundary fully explains the mismatch between the physics of the inversion and the data.

So far, Figure 5.2 shows qualitative results. Figure 5.3 presents the quantitative relationship between MSDM, MAE, and density. Figure 5.3(a) shows the MSDM within the region of interest, for the final 5 iterations, as a function of density. It is apparent that as density increases, so does the MSDM, but that the quantitative relationship is dependent on the hyperparameters. These results are supported by Figure 5.3(b), which shows the MAE, within the region of interest, for the final 5 iterations, as a function of density. It is clear

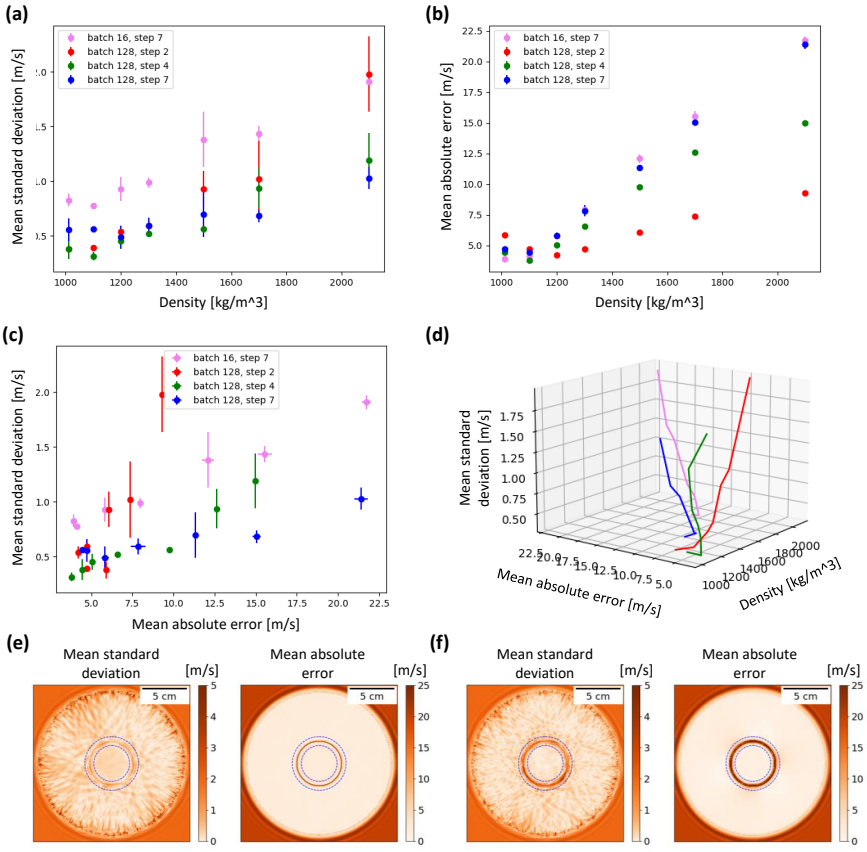


Figure 5.3: The quantitative relationship between mean standard deviation magnitude and mean absolute error where density incudes an artefact. (a) MSDM vs density, (b) MAE vs density, (c) MSDM vs MAE, and (d) MSDM and MAE vs density for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue ring, indicating the region of interest.

that the MAE increases as the density increases. Because the MAE is measured against the ground truth image, this result shows that the artefact formed at the boundary becomes larger as the density increases. This result is also dependent on the hyperparameters, indicating that the artefact formed is actually dependent on the optimisation hyperparameters, as suggested by the MSDM.

Figure 5.3(c) shows that the relationship between the MSDM and MAE is also hyperparameter dependent, and Figure 5.3(d) shows that this dependence is further complicated in relation to density. However, there are some clear observations. First, the MSDM increases

as the MAE increases. This indicates that the standard deviation can be used to quantitatively determine the magnitude of the density induced artefact. Second, the relationship between MSDM and MAE is approximately linear, except in the case where the step length is 2. Third, the hyperparameters have particular effects. For example, batch size creates an offset, while the step-length changes the gradient of the curve. This might imply that the effect of these hyperparameters on the reconstruction can be determined and mitigated. Fourth, Figure 5.3(d) shows a curved relationship between MSDM, MAE, and density when the batch size is 128 and the step length is 4 (i.e. for the green curve). However, Figure 5.3(c) shows that the error bars are large for this case and so a linear relationship cannot be excluded. Finally, Figures 5.3(e) & (f) show that the tomographic standard deviation magnitude reconstruction is correctly located compared to the absolute error. This further supports the hypothesis that the tomographic standard deviation magnitude reconstruction can be used to locate artefacts and quantify their magnitude.

5.3.2 Absorption

The FWI adjoint method assumes there is no absorption in the circle. Similar to the case of variable density, Figure 5.4 shows examples where this assumption is met and where this assumption is incorrect. In the case where the forward and adjoint physics match, the absorption coefficient is zero. In the case where the physics do not match, the forward model uses an absorption coefficient is $10[dB\,cm^{-1}]$ but the adjoint model assumes this is zero.

In Figure 5.4(a) the MSDM is characteristic of a converging reconstruction process. However, Figure 5.4(b) shows that the MSDM does not rise as steeply or achieve the same maximum when the circle is absorbing. In addition, in the absorption case the MSDM settles at approximately $2[m\,s^{-2}]$, but it tends towards zero when there is no absorption.

Figure 5.4(c) shows a reconstruction process that is proceeding as expected for the case where the absorption coefficient is zero. However, Figure 5.4(d) shows the case when the circle has an absorption of $10[dB\,cm^{-1}]$. Now, the sound speed estimate is reduced throughout the circle and the tomographic standard deviation magnitude reconstruction has a high magnitude in this region. These findings are supported by Figures 5.4(e) & (f). In Figure

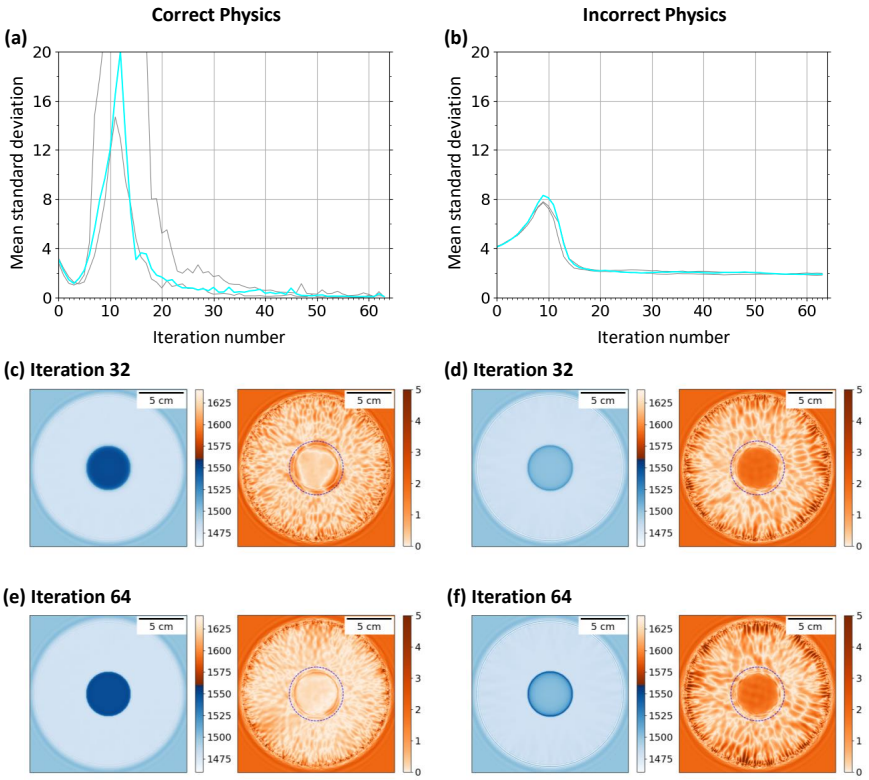


Figure 5.4: VT-FWI reconstructions of the circle dataset. MSDM per iteration is shown for (a) no absorption and (b) variable absorption where the circle phantom is $10[dB\,cm^{-1}]$ for 3 different reconstruction processes. (c-f) Reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue circle, indicating the region of interest.

5.4(e) the mean reconstruction matches the ground truth and the standard deviation reconstruction is low. In Figure 5.4(f) the mean reconstruction continues to under-estimate the sound speed and the standard deviation magnitude reconstruction remains high where the sound speed is under-estimated.

Figure 5.5 explores the quantitative relationship between MSDM, MAE, and absorption. Figure 5.5(a) shows that initially MSDM increases as the absorption coefficient increases, before levelling off. Interestingly, this behaviour is not dependent on the choice of step-length. Figure 5.5(b) shows the same behaviour; the MAE increases and levels off. This means that the magnitude of the under-estimate of the sound speed increases as the ab-

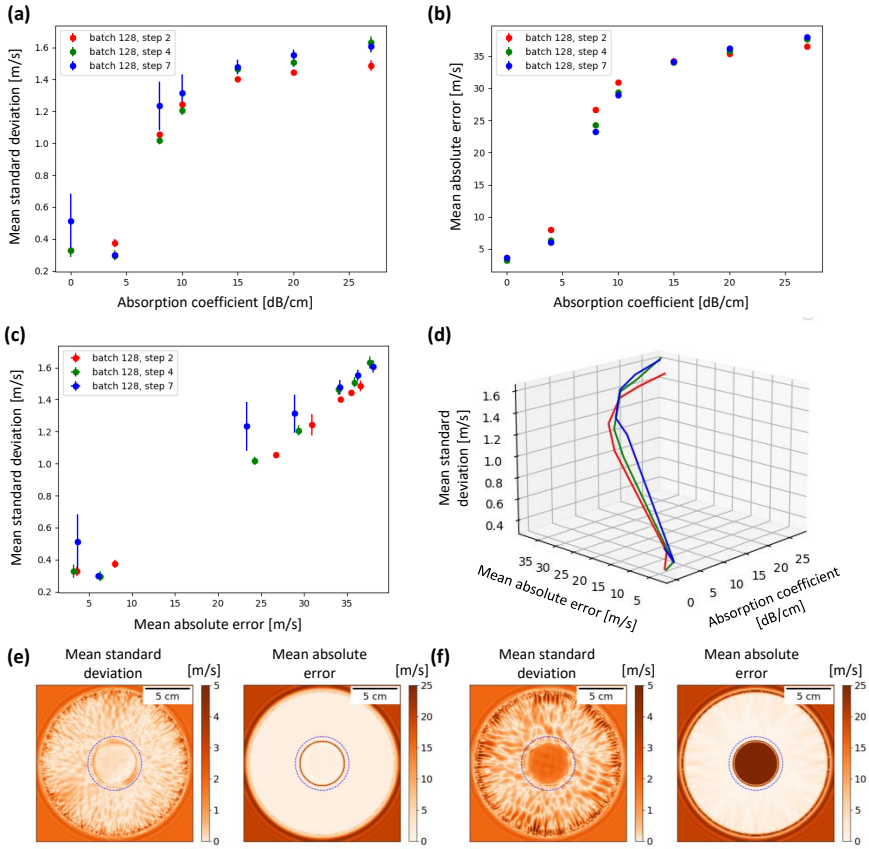


Figure 5.5: The quantitative relationship between mean standard deviation magnitude and mean absolute error where absorption includes an artefact. (a) MSDM vs absorption, (b) MAE vs absorption, (c) MSDM vs MAE, and (d) MSDM and MAE vs density for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue circle, indicating the region of interest.

sorption coefficient increases, but also that the magnitude of the under-estimate begins to saturate once the absorption coefficient reaches $10[dB\,cm^{-1}]$.

Figure 5.5(c) shows a linear relationship between MSDM and MAE. This relationship is also independent of the hyperparameter choice and confirms the observation that MSDM and MAE curve similarly in the presence of absorption artefacts. Finally, Figure 5.5(d) confirms the independence of the results and the choice of step length.

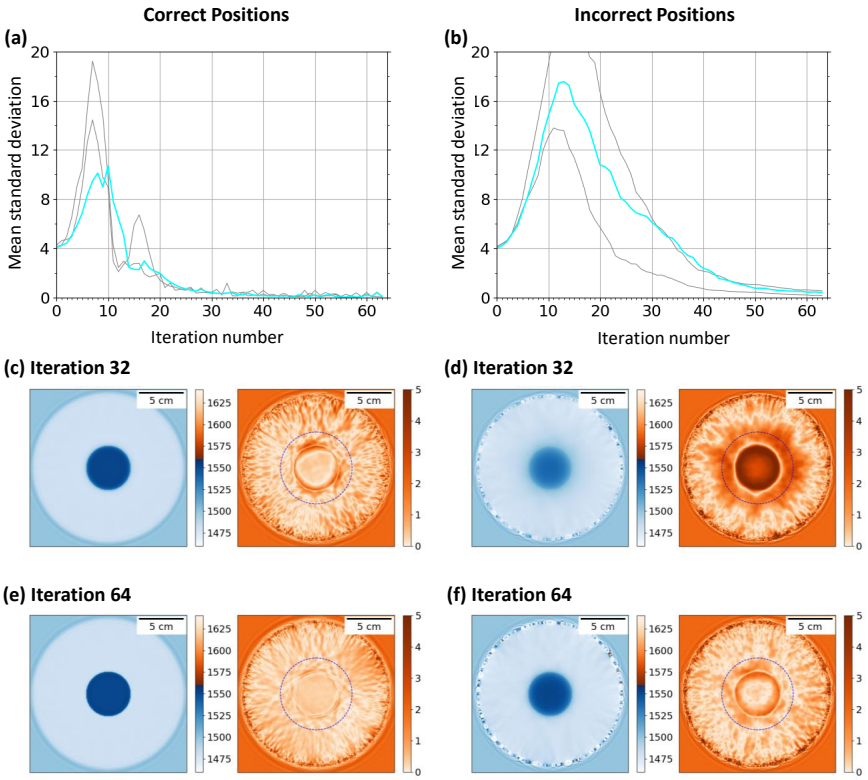


Figure 5.6: VT-FWI recosntructions of the circle dataset. MSDM per iteration is shown for (a) correct transducer positions and (b) random transducer perturbations sampled from 0.2[mm] wide Gaussian for 3 different reconstruction processes. (c-f) Reconstructions of mean (blue) and standard deviation magnitude (orange), taken from process indicated by the cyan line in (a) & (b) respectively. Note the blue circle, indicating the region of interest.

5.3.3 Position errors

In Figures 5.6(a), (c) , & (e) the transducer positions are the same for the forward and adjoint problem. This is a result that has already been described; the MSDM initially rises, before declining as the mean reconstruction converges. Thus, the figures show the expected behaviour when the positions are accurate. In contrast, Figure 5.6(b) shows the case where the transducer positions are correct in the forward problem, but have been perturbed in the adjoint problem. The applied perturbation is a sample from a 2-D Gaussian distribution with a standard deviation of 0.2[mm]. This figure shows that the MSDM rises to a higher maximum and a slightly higher minimum than in the case where the positions are correct.

Comparing Figure 5.6(c) & (d) shows that the standard deviation magnitude reconstruction is significantly higher when the transducer positions have been perturbed. In addition, the region with the highest standard deviation magnitude corresponds to the region where the mean reconstruction remains blurred. This finding is maintained with further iterations; Figure 5.6(f) shows that after 64 iterations the circle has reconstructed, but remains blurred in comparison to Figure 5.6(e) . The blur is evident at the edge of the circle, and the higher standard deviation magnitude is localised to this region.

Figure 5.7 shows a clear quantitative relationship between aleatoric position errors, MSDM, and MAE. In Figure 5.7(a), the MSDM increases as transducer perturbation increases. In addition, MSDM falls slightly when the step length is 4 or 7, and the perturbation is below 0.3$[mm]$. Similar to density, the MSDM is dependent on the step length. These findings are replicated in Figure 5.7(b), where the MAE also falls slightly before rising, and which is also dependent on step length. Because the MAE increases as the step length decreases, this is evidence that the step length changes the rate of reconstruction and, if the step length is too small, then the reconstruction does not converge. Figure 5.7(c) shows a clear linear relationship, which is independent of hyperparameter choice and demonstrates the MSDM and MAE are quantitatively related.

Figure 5.7(d) shows the relationship between the 3 parameters. In the case of aleatoric position errors, the relationship between MSDM and MAE is independent of hyperparameter choice, but the relationship between the metrics and position is not. Finally, note the magnitude of the error bars is significantly greater than in Figures 5.3 and 5.5. This is because the tomographic standard deviation magnitude reconstruction varies significantly across the 3 runs. This effect can also be seen in Figure 5.6 (b).

Figure 5.8 shows the epistemic position error case, where 5 transducers on the bottom of the right are moved up by 1 $[mm]$ and 5 transducers on the left are moved down by 1 $[mm]$. Similar to the aleatoric error, this perturbation is applied to the adjoint problem. The set-up of the epistemic error problem is different to the other results, because the perturbation applies only to a subset of receivers rather than the whole batch. This means the perturbed sources need to regularly appear in the batch, or they need to arrive in a particular order.

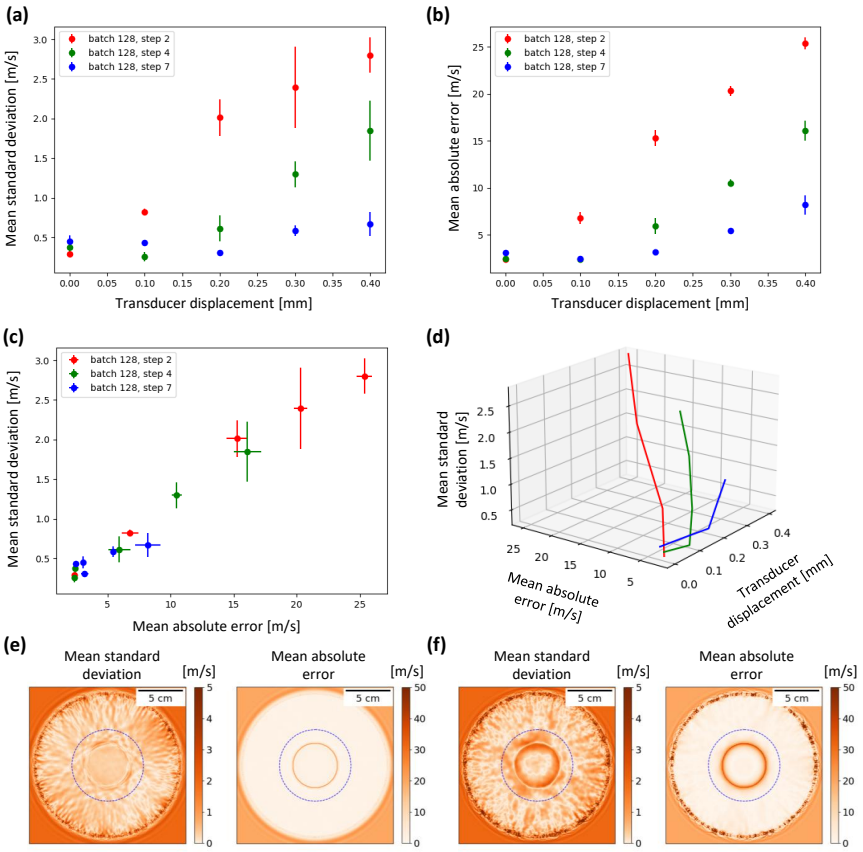


Figure 5.7: The quantitative relationship between mean standard deviation magnitude and mean absolute error where resolution is reduced by transducer misplacement. (a) MSDM vs position error, (b) MAE vs position error, (c) MSDM vs MAE, and (d) MSDM and MAE vs position error for different hyperparameters. (e) & (f) show the standard deviation magnitude and absolute error reconstructions. Note the blue circle, indicating the region of interest.

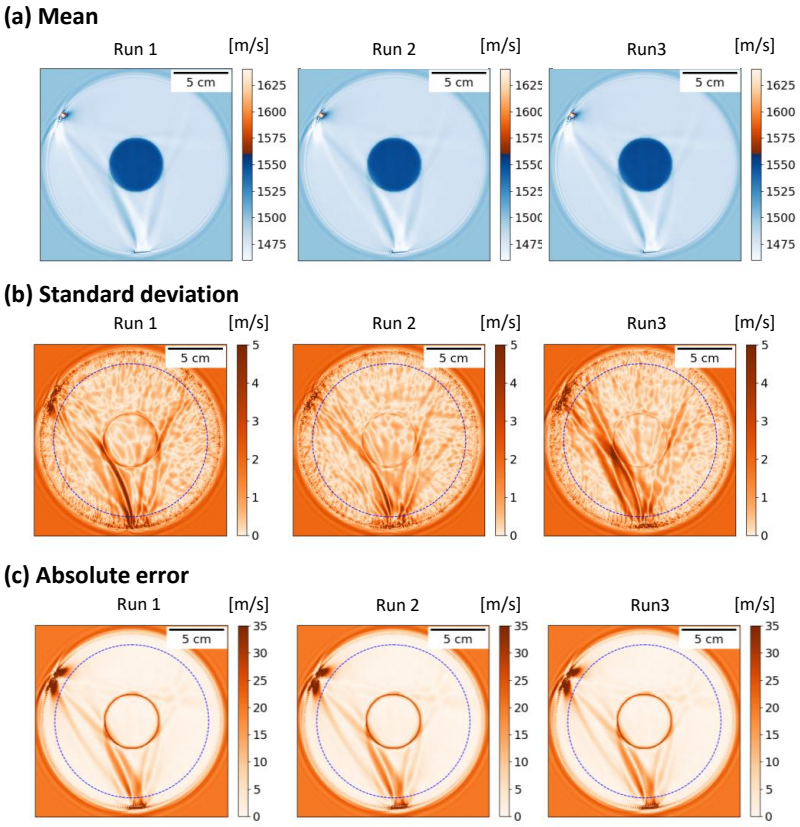


Figure 5.8: VT-FWI reconstructions of the circle dataset, where 10 transducers are systematically displaced 1 $[mm]$ vertically. Reconstructions of (a) mean, (b) standard deviation magnitude, and (c) the absolute error with 3 repetitions.

In addition, more iterations were necessary to invert this dataset. Thus, for this result, the data were low-pass filtered in 2 bands, with the stop bands at $500\,\&\,700[kHz]$ and 64 iterations per band. The use of multiscale techniques is standard practice in FWI, but the number of iterations per band is not typical.

Figure 5.8(a) shows that the effect of the epistemic perturbation is to generate an artefact in the mean reconstruction, which is similar to the *in-vitro* artefact in Figure 4.4(b). Figure 5.8(c) confirms the location and magnitude of the artefact by showing the absolute error compared to the ground truth. The 3 repeats show that the absolute error and mean reconstructions are stable, indicating that the batch gradient descent is not affecting this outcome.

Interpreting Figure 5.8(b) is complicated. The region with the strongest absolute error (at the bottom of the artefact) is consistently identified but the extent and magnitude of the artefact varies between the 3 repeats of the tomographic standard deviation reconstruction. In addition, artefactual regions where the absolute error is weaker are hard to discern in the tomographic standard deviation reconstruction because there is a consistent background noise from which the signal cannot be separated.

5.4 Discussion

5.4.1 Comments on the results

In the previous section, MAE is used as an objective method to determine image quality. The results show that, on average, MSDM is quantitatively related to image quality because the MSDM increases as image quality worsens. The MAE and MSDM results also make intuitive sense because they increase when there is an increase in the relative difference between the forward and inverse problems, which would be expected to degrade image quality. In addition, there is correspondence between artefactual regions identified by reconstructions of tomographic standard deviation magnitude and absolute error. This suggests that the tomographic standard deviation magnitude is also able to locate artefactual regions of the image. However, it is more challenging to conclude that there is a pixel-wise correspondence between the tomographic standard deviation magnitude and the absolute error. This is because, in the tomographic standard deviation reconstruction, there is often high background noise and correlated noise prevents long-range features forming. Both of these issues could be resolved by a data-driven prior. For example, an iterative reconstruction in the latent space of a variational auto-encoder might improve the quality of local standard deviation estimates because the mean-field Gaussian approximation is more appropriate in this space and the encoder/decoder contains an implicit representation of correlations between pixels.

Density The tomographic standard deviation estimator is sensitive to perturbations in the density which aren't explained by the FWI modelling physics. The MSDM is consistently higher when the simulated data contains the effects of variable density and the

standard deviation magnitude reconstruction shows that the region of high MSDM is localised to the regions where the sound speed is erroneous (i.e. the high sound speed artefact at the boundary). In addition, it is expected that this result would translate to *in-vivo* imaging. For example, the range of magnitudes of the density perturbation was chosen to be representative of human tissue; values chosen are typical of cerebrospinal fluid, brain tissues, trabecular bone, and cortical bone [29]. However, the MSDM result is dependent on the choice of hyperparameters. This is not necessarily a problem - if the hyperparameters are consistent, then comparisons are still possible between reconstructions - but it will need to be carefully considered during implementation. In addition, density is challenging to invert using FWI because the sound speed boundary artefacts, which are characteristic of density, can fully explain the effect of density in the loss function. Again, prior information about the correlation between sound speed and density could mitigate this problem, because it can constrain the reconstruction and prevent unphysical artefacts being generated. This information is also readily available from the literature [29].

Absorption The tomographic standard deviation estimator can fully quantify artefacts caused by perturbations in the absorption coefficient which aren't explained by the FWI modelling physics. The MSDM produces a consistent estimate, which is linearly related to MAE, although the absorption coefficients are larger in magnitude than those of human tissues [29]. The relationships between MSDM, MAE and the absorption coefficient are also independent of the hyperparameters, meaning it is simple to define a function to map between MSDM and MAE for these artefacts. The tomographic standard deviation reconstruction is also able to localise the absorption artefact in comparison to the absolute error. However, the pixel-wise correspondence is limited by background noise and correlated noise as discussed previously.

Position The tomographic standard deviation estimator can fully quantify degradation in the resolution of the mean reconstruction, caused by aleatoric perturbations in the transducers. This is validated the qualitative findings from the previous section, demonstrating that MSDM is quantitative of resolution. Furthermore, the blurred region of the reconstruction is correctly located by the tomographic standard deviation reconsturction at the edge of

the circle phantom. The magnitude of the aleatoric perturbation is also realistic for calibartion algorithms, which means the MSDM is sensitive to perturbations of $0.2[mm]$. The tomographic standard deviation estimator is less capable of quantifying artefacts induced by epistemic position artefacts. Comparing the absolute error and the standard deviation magnitude reconstruction shows that regions of the mean reconstruction with the highest artefact magnitude are consistently located. However, smaller artefacts are less consistently located and the problem with pixel-wise correspondence is particularly evident from this example.

5.4.2 Usefulness

This chapter demonstrates that VT is a useful tool for IQA in several realistic scenarios. For example, the tomographic standard deviation magnitude is already capable of locating and quantifying absorption artefacts and blur. Previous studies show that absorption strongly modulates the amplitude of acoustic wavefields brain tissue [31], [32], which means that if absorption is not correctly accounted for in the inverse problem, then significant artefacts can be expected to occur in the sound speed (i.e. tomographic mean) reconstruction. Similarly, calculating the positions and pressure output of ultrasound transducers in an array is difficult, and errors in these parameters reduce resolution and generate artefacts [132]. The tomographic standard deviation magnitude has also been demonstrated to correctly quantify resolution, locate regions of blur, and partially locate position artefacts.

Two applications are immediately clear. First, the tomographic standard deviation magnitude can be optimised to improve the parameters of the reconstruction. For example, hyperparameters such as batch size and step length could be chosen to minimise the MSDM and computation time for a fixed number of iterations. In this scheme, the step length would be as high as possible without causing the reconstruction to diverge, while the batch size would be reduced to speed up computation time without reducing image quality.

Second, the tomographic standard deviation magnitude can be used to automatically identify reconstructions which have poor IQA. This means the data can be automatically re-acquired while the patient is still in position, rather than the patient having to return

for a separate study. This scheme could go even further because the MSDM could be used to identify the sets of the data which are causing the problems. For example, if a patient moved at the end of a study, then the high redundancy in the US-CT data means that these sources could be automatically identified and removed from the reconstruction process with minor changes to image quality. In addition, if more sources are affected, then the position of the entire transducer array could be adjusted until the tomographic standard deviation magnitude is optimal. This would work because of reciprocity between the position of the patient and the position of the transducer array.

5.4.3 Future work

It is important to understand the origins of the image quality problems identified throughout this chapter. The effects of density and position on the mean (sound speed) reconstruction are straight-forward. Inhomogeneous density increases impedance contrast at the boundary of the circle, which changes the relative amplitude of the transmissions and reflections. This effect can be replicated by the high sound speed artefact seen in Figure 5.2. Position errors adjust the phase of the observed and predicted wavefield data, which blurs the position of the boundary of the circle and is seen in Figure 5.6.

Figure 5.9 shows that the effects of absorption are more complicated. Figure 5.9 (a) shows the wavefield data without absorption. The most prominent feature of the acute azimuthal angle transducers is the direct water wave, but these transducers also contain two reflections; first, from the outer surface of the circle and, second, from the inner surface furthest from the source. The most prominent feature of the oblique, near-180^o azimuthal angle transducers is the transmitted pressure, but a diffracted wave is also present and arrives just after the transmissions. Figures 5.9 (b-d) show that as absorption increases, the amplitude of the reflections increase and the amplitude of the transmissions reduce. The amplitude of the diffracted wave remains similar. In addition, the acute angle reflections (Figure 5.9 (c)) arrive early and the oblique angle transmissions (Figure 5.9 (d)) arrive late.

Figure 5.4 shows that when data including attenuation is inverted, the sound speed is under-estimated in the centre of the circle. This will have a similar effect to that of

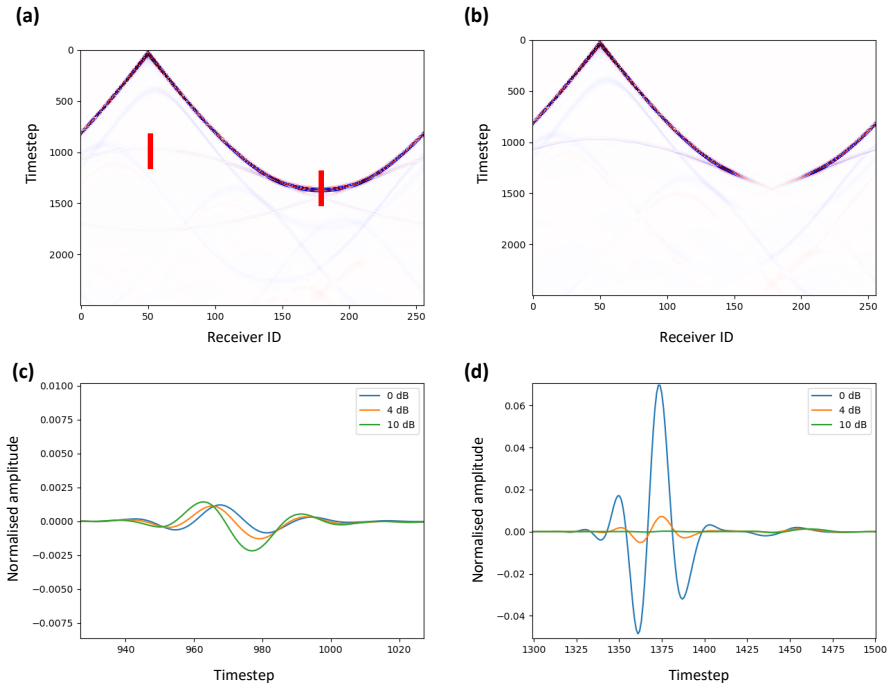


Figure 5.9: Wavefield data are shown when there is (a) no absorption and (b) absorption of 10 [dB]. Plots of the (c) reflected component (receiver ID 50) and (d) transmitted component (receiver ID 178) of the wavefield for different absorption coefficients, the approximate location of the plots are indicated by the red lines in (a).

absorption. First, the reflected pressure on the acute azimuthal angle transducers will increase because the gradient within the circle (i.e. from high to low sound speeds) reflects the wavefield as it propagates. Second, the arrival of the transmitted pressure will be delayed as the average sound speed of the circle reduces. Finally, the diffracted pressure will be maintained because the impedance contrast of the boundary of the circle remains the same.

The results are relevant to US-CT of the breast and brain, but the physics of wave propagation in the brain is also expected to include elastic effects due to the skull. Therefore, an important direction is to evaluate the ability of MSDM to monitor the effect of elastic wave propagation. Elastic wave propagation modulates the phase and amplitude of the wavefield, which wavefields assuming the acoustic approximation cannot replicate. It is expected that elastic effects will induce artefacts into the sound speed or mean reconstruction, and that

these artefacts can be quantified and located like the absorption and position results, which modulate the amplitude and phase of the wavefield data respectively. This result requires further work because it is particularly difficult to invert elastic effects in wavefield data.

There are also research directions which could improve the quality of the tomographic standard deviation magnitude estimate. First, note that the relationship between MSDM and MAE is hyperparameter dependent in the presence of a density perturbation. Then, note that Figure 5.5 shows that choosing a step length of 2 produces a large MSDM error-bar for low absorption coefficients and Figure 5.7 shows that a step length of 2 produces a large MSDM error-bar for high transducer displacements. These results imply that the intermediate tomographic standard deviation magnitude reconstructions are variable and that the scale of the variation depends on the step length. Finally, note that the error bar on the equivalent MAE plot is consistently small, which implies the tomographic mean reconstruction follows a consistent path. The only difference between the mean update (Equation 3.54) and the standard deviation update (Equation 3.53) is a Monte-Carlo sample from the standard normal distribution. This implies that the Monte-Carlo sampling procedure is the significant source of variation in the intermediate tomographic standard deviation magnitude reconstructions.

Together these results suggest that principled strategies for optimising the step length should be explored; an obvious choice would be a line search [142], which is already employed in Geophysical FWI problems [15]. The final result suggests that variance reduction methods would be appropriate to improve the consistency of the intermediate results. For example, these methods could include clipping outlying samples from the standard normal or removing gradients which are dissimilar from the per-iteration average. A more principled approach to variance reduction might involve replacing the Monte-Carlo samplers with a more consistent mechanism. For example, a uniform sample or latin hypercube sample might be suitable alternatives. Finally, some tomographic standard deviation reconstructions show excellent fidelity to the absolute error. If the sampling strategy behind these reconstructions can be replicated, then reducing the variation in the standard deviation update algorithm might improve the quality of the final reconstructions.

Chapter 6

Conclusions

The work presented in my thesis demonstrates the feasibility of a novel, low-cost tomographic uncertainty estimator, which enables real-time image quality analysis. This estimator is stable and sensitive to image quality for *in-vitro* experimental data, and can accurately quantify image quality for *in-silico* experimental data. The conclusion is supported by 3 specific contributions.

My first contribution (Chapter 3) constructs a Bayesian model for FWI. The model contains a *maximum a posteriori* estimator that is similar to the traditional FWI estimator, and an uncertainty estimator which is available by an additional vector multiplication. This scheme, referred to as variational tomography, assumes that the likelihood distribution and the tomographic images are mean-field Gaussian distributed. First, it is demonstrated that a mean-field Gaussian likelihood distribution is equivalent to using the popular L2-norm objective function in traditional FWI. Then, the mean-field Gaussian distribution over the pixels of the tomographic image is justified because 1) the mean estimator is similar to the traditional FWI estimator and 2) the uncertainty estimator is available without significant additional computations. An analysis of the estimator using an *in-silico* dataset shows that the uncertainty estimator is stable, and that the uncertainty reconstructions have qualitative similarities with the absolute error between the mean reconstruction and the ground truth image.

My second contribution (Chapter 4) demonstrates that the estimator is stable using *in-vitro* data, and that the uncertainty estimates are qualitatively related to the image quality of the mean reconstruction. Specifically, it was demonstrated that uncertainty estimator was sensitive to resolution, reconstruction failure, and an artefact. The results also show that the uncertainty estimates are available with very little additional computation time.

My final contribution (Chapter 5) extends the previous qualitative analysis by demonstrating that the uncertainty estimates can quantify image quality. Specifically, it was demonstrated that a new quantity, the mean standard deviation magnitude, is proportional to the mean absolute error, which is an existing and popular image quality analysis metric. The quantitative relationship between the mean standard deviation magnitude and the mean absolute error is shown in 4 contexts; artefacts induced by density perturbations, artefacts induced by absorption perturbations, reductions in resolution caused by incorrect transducer placement, and artefacts caused by incorrect transducer placement.

Variational tomography still has limitations. First, prior information has not been integrated into the reconstruction algorithm. Second, a method of hyperparameters selection for the gradient descent algorithm is not immediately obvious. Finally, the uncertainty estimates contain correlated background noise. These limitations will be addressed in future work, which is already being undertaken.

Chapter 7

Future work

At the end of Chapters 3, 4 & 5, directions for future work have been discussed in detail. The most important outstanding problem is the incorporation of prior information. Existing work on specifying a prior is detailed in Section 7.1 and this will conclude the remaining unmet motivation of this thesis; the incorporation of statistical prior information without additional computational cost.

Other work will focus on improving the existing results. For example, the uncertainty estimator associated with VT demonstrates significant stochastic variation in the results. This is evident in Chapter 5 because the mean standard deviation magnitude is often associated with large error bars. I believe this is caused by the scalar multiplier contained in the standard deviation update (Equation 3.48) which is drawn from the standard normal distribution using the Monte-Carlo method. The Monte-Carlo method is an appropriate sampling algorithm, but it is also likely be the cause of the variation (i.e. non-repeatability) in the standard deviation reconstructions. The evidence in favour is, first, that the mean reconstruction has lower variation and, second, that the other source of stochastic behaviour (i.e. the batch size) is very large for the quantitative reconstructions. Other sampling methods, such as uniform sample or latin hypercube sample, are also appropriate and might lead to a more stable update for the uncertainty term.

The reconstruction process is also dependent on the chosen hyper-parameters, which is

made clear in the results section of Chapter 3. This thesis typically uses a grid-search method to optimise these hyper-parameters. The results in Chapters 4 & 5 are clear evidence that a grid search is effective and repeatable on a variety of datasets. However, currently the hyper-parameter optimisation is chosen per-dataset, which might prevent the use of uncertainty as a metric for optimising acquisition system settings (the 3$^{\text{rd}}$ motivation, Section 1.5). It is a fact that solutions from optimisation algorithms are dependent on their hyper-parametersand there is no guaranteed solution. However, there are principled methods which might be more appropriate than a grid-search; for example the line search algorithm [15], [142]. Further studies on the effect of hyper-parameters on the VT posterior are important.

Resolving this issue will permit the use uncertainty as a metric for optimising acquisition system settings. This is an exciting direction because the VT uncertainty estimate is available per iteration. Therefore, the uncertainty estimator could be used to guide an adaptive or dynamic reconstruction process, where the next state of the optimisation algorithm is influenced by the current state. For example, source-receiver pairs could be chosen to improve the quality of the gradient estimate and therefore speed-up the reconstruction process.

For FWI imaging of the human head, it is also important to evaluate the ability of the uncertainty estimator to monitor the artefacts generated by elastic wave propagation. This is a promising area of further study because elastic wave-propagation is present in bone tissues. I expect that the VT uncertainty estimator will be more sensitive to the discrepancies of this type because elastic effects modulate the phase of the wavefield data, FWI is more sensitive to phase changes, and, unlike density, elastic effects cannot be consistently explained by adjusting the sound speed. It will also be important to assess whether the estimator is equally effective for 3-D imaging because most modern tomographic reconstruction algorithms produce 3-D posteriors.

7.1 Prior information

I am currently investigating the use of prior information in the reconstruction problem, which will meet the primary objective of incorporating statistical prior information without significant additional computational cost. This work will compliment another topic, which is to

incorporate covariance information into the reconstruction. In Chapter 4, I speculated that a lack of covariance information was affecting the quality of the uncertainty reconstruction (see Section 4.4.1).

Prior information is introduced in Chapter 3, where the Bayesian interpretation of FWI produces a gradient-based update, given by Equation 3.31. This equation contains terms for the likelihood and prior distributions such that,

$$\begin{aligned}\nabla_\lambda \mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = &\int_\epsilon q(\epsilon)\,\nabla_g \ln\{g[\lambda,\epsilon]\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon \\ &- \int_\epsilon q(\epsilon)\,\nabla_g \ln\{p(g[\lambda,\epsilon])\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon \\ &- \int_\epsilon q(\epsilon)\,\nabla_g \ln\left\{p(\mathcal{D}\,|\,g[\lambda,\epsilon])\right\}\nabla_\lambda\, g[\lambda,\epsilon]\,d\epsilon\end{aligned} \tag{7.1}$$

Then, VT was developed in Section 3.3.1 by considering only the likelihood component of the update. This is strictly incorrect, but it is a useful simplification and it is common practice in FWI. However, the initial motivation for VT was to develop a framework where prior information can be encoded statistically. This contrasts with traditional FWI, where a point estimate (the "starting model") is used to initiate the optimisation algorithm but is not used as a constraint. If we choose to include a prior in the optimisation, then Equation 7.1 can be arranged such that,

$$\begin{aligned}\mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = &-\int_\epsilon q(\epsilon)\,\ln\left\{p(\mathcal{D}\,|\,m)\right\}\,d\epsilon \\ &+ \int_\epsilon q(\epsilon)\,\ln\left\{\frac{q(m;\lambda)}{p(m)}\right\}\,d\epsilon\end{aligned} \tag{7.2}$$

which is the same as Equation 3.33. The second term in this expression is a KL-divergence between the prior and the approximating distribution. This term will constrain the approximate posterior distribution (i.e. $q(m;\lambda)$) to be similar to the prior, but will also give the approximate posterior more freedom where the uncertainty is higher.

This formulation raises the important question about what form the prior should take. Priors can come in two forms; hand-crafted and data-driven [133]. In addition, Section 4.4.1

introduces the concept of parametric and non-parametric. Typically, hand-crafted priors would assume a parametric distribution, for example a Gaussian distribution, which would enforce a specific correlation function between the pixels. This works when the correlation function can be determined analytically, but that is not usually the case. In reality, a hand-crafted prior could limit the solution space in an unrealistic way. For example, a fully-specified Gaussian distribution would prevent sharp changes in contrast but sharp changes in contrast are present in the human head, for example, the sound speed boundary between the skull and the brain.

Regularisation will affect the quantitative standard deviation magnitude estimates investigated in the previous chapter. Equation 7.2 includes a regularisation term and estimates the posterior variance. Section 3.3.1 states that the KL-divergence term can be seen as a regularisation term such that,

$$\mathbb{E}_{q(m;\lambda)}\left[\ln\left\{\frac{q(m;\lambda)}{p(m,\mathcal{D})}\right\}\right] = -\mathcal{L}(m) + a\,\mathcal{S}(m) \tag{7.3}$$

In Chapter 3 the weight a is set to zero. Figure 7.1 shows the effect of L2-regularisation by introducing weights $a = [0.5, 0.7, 1]$. Figure 7.1 (a) shows the effect of L2 regularisation on the mean reconstruction; the mean reconstruction tends towards the mean of the prior as the magnitude of the regularisation weight increases. Figure 7.1 (b) shows that the standard deviation of the likelihood (i.e. the unregularised standard deviation) maintains it's relationship with qualitative perceived image quality. In contrast, Figure 7.1 (c) shows that the standard deviation of the posterior (i.e. the regularised standard deviation), tends towards the standard deviation of the prior as the regularisation weight increases.

7.2 Learnt statistical priors

In practice, the simplest choice of prior is to assume a mean-field Gaussian distribution. Section 7.1 shows a mean-field Gaussian prior which has a uniform mean and is approximately the sound speed of water. This could be extended to a head prior by taking the

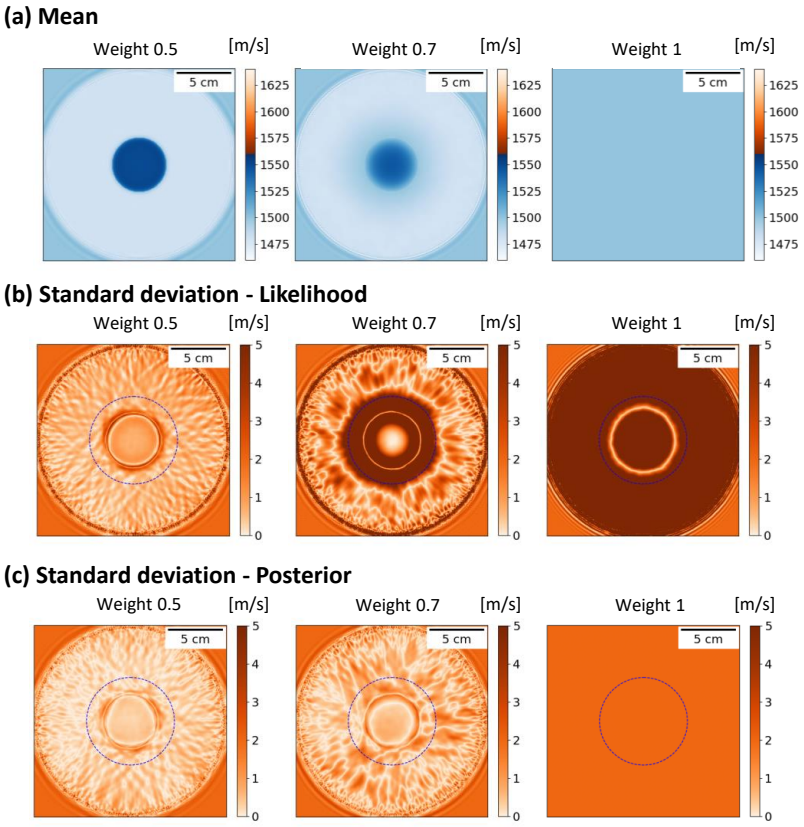


Figure 7.1: L2 regularisation is introduced progressively by adjusting the weight term. (a) shows the effect of regularisation on the mean, (b) shows the unregualarised standard deviation of the likelihood, and (c) shows the regularised standard deviation of the posterior.

mean and variance of a sample of sound speed images of human heads. The problem with a mean-field Gaussian prior is that there is significant variation in the size and shape of different individual skull bones. Figure 7.2 shows CT images for 9 patients in the RSNA 2019 Brain CT Hemorrhage Challenge dataset [160]. It is clear that taking a pixel-wise mean and standard deviation on the processed images (Figure 7.2(j-r)) would significantly overestimate the variability at the boundary of the skull because this procedure does not appropriately consider variation the shape and size. Therefore, a mean-field Gaussian prior is inappropriate.

A non-parametric, data-driven prior is preferable as constraint for FWI. This is because a non-parametric model can be specified which has sufficient flexibility to learn the correl-

ation function of the human head images. Ideally, the model would bias the reconstruction algorithm towards posterior distributions which are close to the prior distribution, without being overly restrictive. However this requires that a sufficiently large dataset of head images is available to train the model.

The key to solving this problem is to determine a suitable choice of non-parametric model. A long standing framework in computational imaging are statistical shape and appearance models [161], [162]. A traditional statistical shape model works using labelled data [161]. Images are labelled with landmark points, for example, an image of a face might be labelled with landmarks at the around the boundary of the mouth, nose, and eyes. Then image of different individual faces are compared by measuring the the distance between corresponding landmarks on the different faces. Finally the statistical variation in these distances is determined by principle components analysis. However, a labelled approach is unsuitable for 3-D datasets because labelling dozens of 3-D scans is too time consuming. Instead, I will use a statistical deformation model [163], where the 3-D scans are registered to one-another using an elastic or non-rigid registration algorithm. In this case, the principle components analysis is applied to the deformation fields generated by the non-rigid registration algorithm, which means that it is possible to determine the principle components of variation between individuals. Deformation models of this type are already used in neuroimaging analysis [164], suggesting that they are a suitable choice form encoding prior information.

Clearly, a shape and deformation model is incapable of encoding variation in intensity between individuals. This is problematic because the sound speed of the skull is known to vary between individuals [118], [165], meaning that an average value would be unsuitable. The solution is a statistical appearance model [162], which performs principle components analysis directly on the residual pixel intensity between the individual images and the shape model prediction. Thus, the statistics of intensity variation are determined by principle components analysis after the shape model component has been removed. This idea can be taken to the extreme using a variational auto-encoder (VAE) [112]. A limiting case of a VAE is a single linear layer between the decoded image and the encoding space, which corresponds

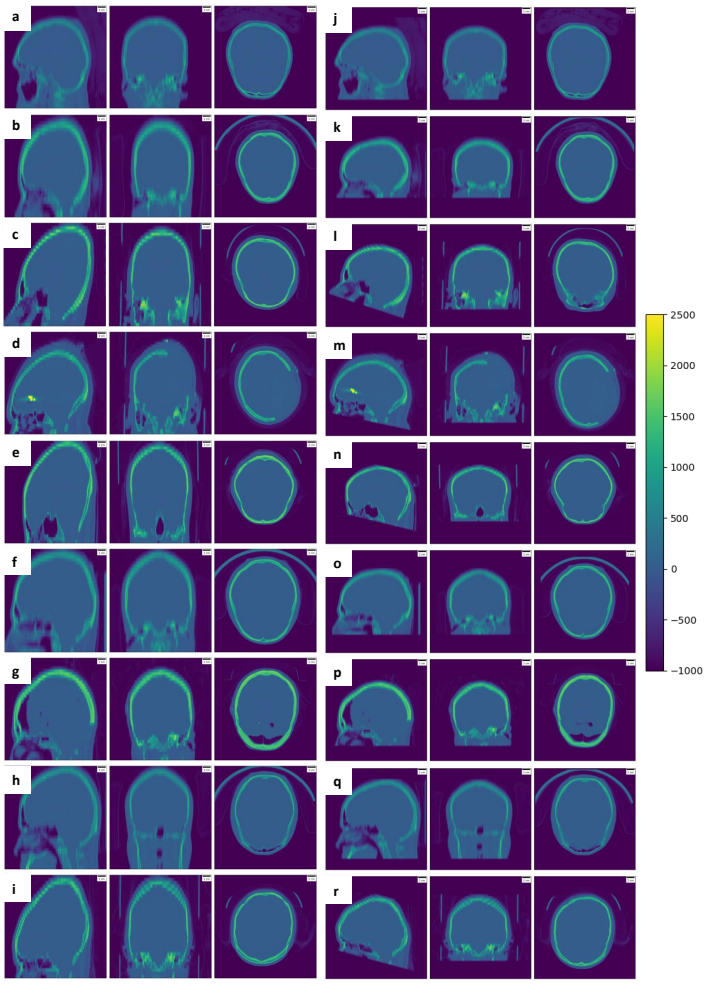


Figure 7.2: Examples from the RSNA dataset (a-i) before preprocessing and (j-r) after preprocessing. Sub-figures show saggital (left), coronal (centre), and axial (right) slices.

directly to principle components analysis. The advantage of the VAE is that a non-linear function can be used between the image space and the encoded latent vector. This means that VAE's are capable of capturing complex correlation functions [166] and might negate the need for a shape model. An additional advantage is that the latent vector of a VAE is mean-field Gaussian distributed by definition, potentially alleviating the inaccuracy of the mean-field Gaussian assumption when it is made in image space.

A dataset of example reconstructions is critical for data-driven prior. In the case of a VAE this must be thousands of images in order to fully represent the significant variation

between individuals. Sound speed images of the human head are not available, so the example images must be derived from other modalities. Open-source datasets are available from CT [160] and MRI [167]. Because FWI struggles to reconstruct sound speed images without prior knowledge of the skull, the most important property of the prior distibution is to have good fidelity to the morphology and properties of the skull. CT reconstructions are generally superior to MRI in hard tissue and the relationship between the CT Housfield unit and the acoustic properties of bone is well understood [118], [165]. Therefore, the CT datasets are preferred.

The RSNA 2019 Brain CT Hemorrhage Challenge dataset [160] is a dataset of $25,262$ volumetric CT scans of $22,456$ unique patients, drawn from Stanford University (Palo Alto, Calif), Universidade Federal de São Paulo (São Paulo, Brazil) and Thomas Jefferson University Hospital (Philadelphia, Pa). The dataset contains single and repeat scans; $20,698$ patients have a single study taken, while $1,758$ patients have 2 or more studies. This means it might be possible to infer temporal changes in the condition of the patient but, as temporal labels are not available, this will be challenging. The scale of the dataset means that it is suitable for training deep generative models such as VAEs and $21,774$ studies have 5 labels describing the haemorrhage status of the patient. This means that the shape or appearance models can be conditioned on these labels and will allow the prior to be used as a diagnostic tool.

Unfortunately the dataset is not immediately usable for training 3-D models because various pre-processing tasks are required to organise the raw data. The dataset comes as single slices, so the individual slices are sorted by the Patient Unique Identifier and Series Instance Unique Identifier. Figure 7.2(a-i) shows examples of data once the sorting algorithm has been applied, but before further preprocessing. Note that the aspect ratio of these images is unequal to compensate for the unequal shape of the voxels. Once the dataset was sorted, further processing is required to resolve the following problems. First, some volumes have intensity values outside of the the $[-1000, 2500]$ range appropriate for Hounsfield units. Second, the volumes have unequal spacing, which is true within individual scans and between different scans. Within a typical volume, there is an horizontal spacing

of $0.5\,[mm]$ but a vertical spacing of $4\,[mm]$. There is also variation between patients, for example the vertical slice spacing differs from $1-5\,[mm]$ throughout the dataset. Third, approximately $11,000$ volumes suffer from Gantry Tilt, where the rotating gantry of the CT scanner is not perpendicular to the head. This causes distortion to the unprocessed data, for example in Figures 7.2(c-e, i), and causes a rotated orientation in the corrected data, for example Figures 7.2(l-n, r). Fourth, approximately $5,000$ scans have unequal slice spacing, where the slices have different spacing depending on the vertical position (relative to the patient) of the rotating gantry. To make the dataset suitable for 3-D generative modelling, the volumes were rescaled into the [-1000, 2500] range, the gantry tilt was corrected, and the volumes were resampled onto an equispaced grid. The resampling is performed using a bilinear interpolation algorithm. Scans with unequal slice spacing were removed from the analysis. Figure 7.2(j-r) shows examples from the dataset after these corrections have been applied.

Regarding full development of the prior, the first step will be to implement a VAE capable of modelling the correlation function of the human head. This will be trained using the RSNA dataset once 2 further corrections have been implemented. First, the processed data will be aligned using a rigid registration algorithm. This means that the VAE will not need to account for significant variations in patient position, which is unnecessary and will slow down training. Second, the processed data will be downsampled into a grid of 64^3 pixels, with a slice spacing of $4\,[mm]$. This is necessary to permit model training on two GPU's (Quadro RTX 8000, Nvidia, CA). The smaller model will provide a proof-of-concept for a larger more sophisticated model, which will need to be trained on a more capable system.

Bibliography

[1] M. Lustig, D. L. Donoho, J. M. Santos and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 72–82, Mar. 2008.

[2] D. J. Brenner and E. J. Hall, "Computed tomography — an increasing source of radiation exposure," *N. Engl. J. Med.*, vol. 357, no. 22, pp. 2277–2284, 29th Nov. 2007.

[3] K. Wang, T. Matthews, F. Anis, C. Li, N. Duric and M. A. Anastasio, "Waveform inversion with source encoding for breast sound speed reconstruction in ultrasound computed tomography," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 62, no. 3, pp. 475–493, Mar. 2015.

[4] G. Y. Sandhu, C. Li, O. Roy, S. Schmidt and N. Duric, "Frequency domain ultrasound waveform tomography: Breast imaging using a ring transducer," *Phys. Med. Biol.*, vol. 60, no. 14, pp. 5381–5398, Jun. 2015.

[5] O. Calderon Agudo, L. Guasch, P. Huthwaite and M. Warner, "3d imaging of the breast using full-waveform inversion," 2018.

[6] M. Perez-Liva, J. L. Herraiz, J. M. Udias, E. Miller, B. T. Cox and B. E. Treeby, "Time domain reconstruction of sound speed and attenuation in ultrasound computed tomography using full wave inversion," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. 1595–1604, Mar. 2017.

[7] F. Lucka, M. Pérez-Liva, B. E. Treeby and B. T. Cox, "High resolution 3d ultrasonic breast imaging by time-domain full waveform inversion," *Inverse Problems*, vol. 38, no. 2, p. 025 008, Dec. 2021.

[8] B. Arnal, G. Pinton, P. Garapon, M. Pernot, M. Fink and M. Tanter, "Global approach for transient shear wave inversion based on the adjoint method: A comprehensive 2d simulation study," *Phys. Med. Biol.*, vol. 58, no. 19, Sep. 2013.

[9] J. W. Wiskin, D. T. Borup, E. Iuanow, J. Klock and M. W. Lenox, "3-d nonlinear acoustic inverse scattering: Algorithm and quantitative results," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 64, no. 8, pp. 1161–1174, Aug. 2017.

[10] S. Bernard, V. Monteiller, D. Komatitsch and P. Lasaygues, "Ultrasonic computed tomography based on full-waveform inversion for bone quantitative imaging," *Phys. Med. Biol.*, vol. 62, no. 17, pp. 7011–7035, Aug. 2017.

[11] J. Wiskin, B. Malik, D. Borup, N. Pirshafiey and J. Klock, "Full wave 3d inverse scattering transmission ultrasound tomography in the presence of high contrast," *Sci. Rep.*, vol. 10, no. 1, p. 20 166, 19th Nov. 2020.

[12] Q. Shi, Y. Li, Y. Liu *et al.*, "Index-rotated fast ultrasound imaging of cortical bone based on predicted velocity model," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 5, pp. 1582–1595, May 2022.

[13] L. Guasch, O. Calderón Agudo, M.-X. Tang, P. Nachev and M. Warner, "Full-waveform inversion imaging of the human brain," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–12, 6th Mar. 2020.

[14] U. Taskin, K. Solberg Eikrem, G. Nævdal, M. Jakobsen, D. J. Verschuur and K. W. A. van Dongen, "Ultrasound imaging of the brain using full-waveform inversion," in *2020 IEEE International Ultrasonics Symposium (IUS)*, Sep. 2020, pp. 1–4.

[15] J. Virieux and S. Operto, "An overview of full-waveform inversion in exploration geophysics," *GEOPHYSICS*, vol. 74, no. 6, WCC1–WCC26, 1st Nov. 2009.

[16] V. L. Feigin, E. Nichols, T. Alam *et al.*, "Global, regional, and national burden of neurological disorders, 1990–2016: A systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 5, pp. 459–480, 1st May 2019.

[17] M. P. Lindsay, B. Norrving, R. L. Sacco *et al.*, "World stroke organization (WSO): Global stroke fact sheet 2019," *Int. J. Stroke*, vol. 14, no. 8, pp. 806–817, 1st Oct. 2019.

[18] P. A. Barber, A. M. Demchuk, J. Zhang and A. M. Buchan, "Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy," *The Lancet*, vol. 355, no. 9216, pp. 1670–1674, 13th May 2000.

[19] "Good is not good enough: The benchmark stroke door-to-needle time should be 30 minutes," *Can. J. Neurol. Sci.*, vol. 41, no. 6, pp. 694–696, Nov. 2014.

[20] W. Johnson, O. Onuma, M. Owolabi and S. Sachdev, "Stroke: A global response is needed," *Bull. World Health Organ.*, vol. 94, no. 9, 634–634A, 1st Sep. 2016.

[21] V. L. Feigin, C. M. Lawes, D. A. Bennett, S. L. Barker-Collo and V. Parag, "Worldwide stroke incidence and early case fatality reported in 56 population-based studies: A systematic review," *The Lancet Neurology*, vol. 8, no. 4, pp. 355–369, 1st Apr. 2009.

[22] S. W. Smith, N. M. Ivancevich, B. D. Lindsey *et al.*, "The ultrasound brain helmet: Feasibility study of multiple simultaneous 3d scans of cerebral vasculature," *Ultrasound in Medicine & Biology*, vol. 35, no. 2, pp. 329–338, 1st Feb. 2009.

[23] B. D. Lindsey and S. W. Smith, "Refraction correction in 3d transcranial ultrasound imaging," *Ultrasonic Imaging*, vol. 36, no. 1, pp. 35–54, 1st Jan. 2014.

[24] Smith S W, von Ramm O T, Kisslo J A and Thurstone F L, "Real time ultrasound tomography of the adult brain.," *Stroke*, vol. 9, no. 2, pp. 117–122, 1st Mar. 1978.

[25] F. J. Fry, "Transkull transmission of an intense focused ultrasonic beam," *Ultrasound Med. Biol.*, vol. 3, no. 2, pp. 179–184, 1st Jan. 1977.

[26] J. Ylitalo, J. Koivukangas and J. Oksman, "Ultrasonic reflection mode computed tomography through a skullbone," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 11, pp. 1059–1066, Nov. 1990.

[27] M. Hajian, R. Gaspar and R. Maev, "Accurate 3-d profile extraction of skull bone using an ultrasound matrix array," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2858–2871, 20th Nov. 2017.

[28] W.-D. Niesen, M. Rosenkranz and C. Weiller, "Bedsided transcranial sonographic monitoring for expansion and progression of subdural hematoma compared to computed tomography," *Front. Neurol.*, vol. 9, 2018.

[29] P. Hasgall, E. Neufeld, M. Gosselin, A. Klingenbock and N. Kuster, *IT'IS database for thermal and electromagnetic parameters of biological tissues*, type: dataset, 2015.

[30] K. A. Dines, F. J. Fry, J. T. Patrick and R. L. Gilmor, "Computerized ultrasound tomography of the human head: Experimental results," *Ultrasonic Imaging*, vol. 3, no. 4, pp. 342–351, 1st Oct. 1981.

[31] B. Rajagopalan, J. F. Greenleaf, P. J. Thomas and L. D. Harris, "Ultrasonic characterization and display of human brain in vitro, 5th int. symp," *Ultrasonic Imaging and Tissue Characterization*, vol. 1, 1980.

[32] C. Labuda, W. R. Newman, B. K. Hoffmeister and C. K. M. Chambliss, "Two-dimensional mapping of the ultrasonic attenuation and speed of sound in brain," *Ultrasonics*, vol. 124, p. 106 742, 1st Aug. 2022.

[33] A. Smiley, M. Howell, G. T. Clement and A. J. Fleischman, "Toward transcranial ultrasound tomography: Design of a 456-element low profile conformal array," *Biomed. Phys. Eng. Express*, vol. 5, no. 2, p. 025 025, Jan. 2019.

[34] G. T. Clement, "A projection based approach to diffraction tomography on curved boundaries," *Inverse Problems*, vol. 30, no. 12, p. 125 010, 19th Nov. 2014.

[35] J. Cudeiro-Blanco, C. Cueto, O. Bates *et al.*, "Design and construction of a low-frequency ultrasound acquisition device for 2-d brain imaging using full-waveform inversion," *Ultrasound Med. Biol.*, 25th Jul. 2022.

[36] M. Løberg, M. L. Lousdal, M. Bretthauer and M. Kalager, "Benefits and harms of mammography screening," *Breast Cancer Res.*, vol. 17, no. 1, p. 63, 1st May 2015.

[37] I. G. Murphy, M. F. Dillon, A. O. Doherty *et al.*, "Analysis of patients with false negative mammography and symptomatic breast carcinoma," *J. Surg. Oncol.*, vol. 96, no. 6, pp. 457–463, 2007.

[38] "Radiation risk with digital mammography in breast screening," GOV.UK. (), [Online]. Available: https://www.gov.uk/government/publications/breast-screening-radiation-risk-with-digital-mammography/radiation-risk-with-digital-mammography-in-breast-screening (visited on 30/07/2022).

[39] P. J. Littrup, N. Duric, M. Sak *et al.*, "Multicenter study of whole breast stiffness imaging by ultrasound tomography (SoftVue) for characterization of breast tissues and masses," *J. Clin. Med.*, vol. 10, no. 23, p. 5528, Jan. 2021.

[40] "Premarket approval (PMA)." (), [Online]. Available: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma.cfm?id=P200040 (visited on 30/07/2022).

[41] N. Duric, P. Littrup, M. Sak *et al.*, "A novel marker, based on ultrasound tomography, for monitoring early response to neoadjuvant chemotherapy," *J. Breast Imaging*, vol. 2, no. 6, pp. 569–576, 21st Nov. 2020.

[42] J. F. Greenleaf, S. A. Johnson, W. F. Samayoa and F. A. Duck, "Algebraic reconstruction of spatial distributions of acoustic velocities in tissue from their time-of-flight profiles," in *Acoustical Holography: Volume 6*, Boston, MA: Springer US, 1975, pp. 71–90.

[43] K. A. Dines and A. C. Kak, "Ultrasonic attenuation tomography of soft tissues," *Ultrasonic Imaging*, vol. 1, no. 1, pp. 16–33, 1st Jan. 1979.

[44] G. Glover and J. Sharp, "Reconstruction of ultrasound propagation speed distributions in soft tissue: Time-of-flight tomography," *IEEE Trans. Sonics Ultrason.*, vol. 24, no. 4, pp. 229–234, Jul. 1977.

[45] N. Ozmen, R. Dapp, M. Zapf, H. Gemmeke, N. V. Ruiter and K. W. A. van Dongen, "Comparing different ultrasound imaging methods for breast cancer detection," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 62, no. 4, pp. 637–646, Apr. 2015.

[46] K. W. van Dongen, L. Demi and M. D. Verweij, "Numerical schemes for the iterative nonlinear contrast source method," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1918–1918, Sep. 2012.

[47] R. Jirik, I. Peterlik, N. Ruiter *et al.*, "Sound-speed image reconstruction in sparse-aperture 3-d ultrasound transmission tomography," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 59, no. 2, pp. 254–264, Feb. 2012.

[48] M. André, J. Wiskin, D. Borup, S. Johnson, H. Ojeda-Fournier and L. Olson, "Quantitative volumetric breast imaging with 3d inverse scatter computed tomography," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 1110–1113.

[49] J. Wiskin, D. T. Borup, S. A. Johnson and M. Berggren, "Non-linear inverse scattering: High resolution quantitative breast tissue tomography," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 3802–3813, May 2012.

[50] P. Huthwaite and F. Simonetti, "High-resolution imaging without iteration: A fast and robust method for breast ultrasound tomography," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1721–1734, Sep. 2011.

[51] N. Duric, P. Littrup, P. Chandiwala-Mody *et al.*, "In-vivo imaging results with ultrasound tomography: Report on an ongoing study at the karmanos cancer institute," in *Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy*, vol. 7629, SPIE, 12th Mar. 2010, pp. 209–217.

[52] J. Wiskin, D. Borup, S. Johnson *et al.*, "Inverse scattering and refraction corrected reflection for breast cancer imaging," in *Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy*, vol. 7629, SPIE, 12th Mar. 2010, pp. 189–200.

[53] C. Li, N. Duric, P. Littrup and L. Huang, "In vivo breast sound-speed imaging with ultrasound tomography," *Ultrasound Med. Biol.*, vol. 35, no. 10, pp. 1615–1628, 1st Oct. 2009.

[54] R. J. Lavarello and M. L. Oelze, "Density imaging using inverse scattering," *J. Acoust. Soc. Am.*, vol. 125, no. 2, pp. 793–802, Feb. 2009.

[55] D. P. Duncan, J. P. Astheimer and R. C. Waag, "Scattering calculation and image reconstruction using elevation-focused beams," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3101–3119, May 2009.

[56] R. Lavarello and M. Oelze, "A study on the reconstruction of moderate contrast targets using the distorted born iterative method," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 55, no. 1, pp. 112–124, Jan. 2008.

[57] N. Duric, P. Littrup, L. Poulo *et al.*, "Detection of breast cancer with ultrasound tomography: First results with the computed ultrasound risk evaluation (CURE) prototype," *Med. Phys.*, vol. 34, no. 2, pp. 773–785, 2007.

[58] R. C. Waag, F. Lin, T. K. Varslot and J. P. Astheimer, "An eigenfunction method for reconstruction of large-scale and high-contrast objects," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 54, no. 7, pp. 1316–1332, Jul. 2007.

[59] R. G. Pratt, L. Huang, N. Duric and P. Littrup, "Sound-speed and attenuation imaging of breast tissue using waveform tomography of transmission ultrasound data," in *Medical Imaging 2007: Physics of Medical Imaging*, vol. 6510, SPIE, 19th Mar. 2007, pp. 1523–1534.

[60] K. W. A. van Dongen and W. M. D. Wright, "A forward model and conjugate gradient inversion technique for low-frequency ultrasonic imaging," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2086–2095, Oct. 2006.

[61] M. L. Oelze and W. D. O'Brien, "Application of three scattering models to characterization of solid tumors in mice," *Ultrasonic Imaging*, vol. 28, no. 2, pp. 83–96, 1st Apr. 2006.

[62] N. Duric, P. Littrup, A. Babkin *et al.*, "Development of ultrasound tomography for breast imaging: Technical assessment," *Medical Physics*, vol. 32, no. 5, pp. 1375–1386, 2005.

[63] T. Hopp, F. Zuch, M. Zapf, H. Gemmeke and N. V. Ruiter, "Experimental analysis of ray-based sound speed reconstruction algorithms for phase aberration corrected USCT SAFT imaging," in *Medical Imaging 2019: Ultrasonic Imaging and Tomo-*

graphy, vol. 10955, International Society for Optics and Photonics, 15th Mar. 2019, 109550J.

[64] M. P. André, H. S. Janée, P. J. Martin, G. P. Otto, B. A. Spivey and D. A. Palmer, "High-speed data acquisition in a diffraction tomography system employing large-scale toroidal arrays," *Int. J. Imaging Syst. Technol.*, vol. 8, no. 1, pp. 137–147, 1997.

[65] P. L. Carson, A. L. Scherzinger, C. R. Meyer, W. Jobe, B. Samuels and D. D. Adler, "Lesion detectability in ultrasonic computed tomography of symptomatic breast patients," *Investigative Radiology*, vol. 23, no. 6, pp. 421–427, Jun. 1988.

[66] M. L. Oristaglio, "A geometric approach to the filtered backpropagation algorithm," *Ultrasonic Imaging*, vol. 5, no. 1, pp. 30–37, 1st Jan. 1983.

[67] G. Y. Sandhu, E. West, C. Li, O. Roy and N. Duric, "3d frequency-domain ultrasound waveform tomography breast imaging," in *Medical Imaging 2017: Ultrasonic Imaging and Tomography*, vol. 10139, SPIE, 13th Mar. 2017, pp. 56–69.

[68] A. Sadeghi-Naini, L. Sannachi, H. Tadayyon *et al.*, "Chemotherapy-response monitoring of breast cancer patients using quantitative ultrasound-based intra-tumour heterogeneities," *Sci. Rep.*, vol. 7, no. 1, p. 10 352, 4th Sep. 2017.

[69] H. Gemmeke, T. Hopp, M. Zapf, C. Kaiser and N. V. Ruiter, "3d ultrasound computer tomography: Hardware setup, reconstruction methods and first clinical results," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 873, pp. 59–65, 2017.

[70] S. Nagel, D. Sinha, D. Day *et al.*, "E-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients," *Int. J. Stroke*, vol. 12, no. 6, pp. 615–622, 1st Aug. 2017.

[71] A. Tarantola, "Inversion of seismic reflection data in the acoustic approximation," *GEOPHYSICS*, vol. 49, no. 8, pp. 1259–1266, 1st Aug. 1984.

[72] P. Lailly and J. Bednar, "The seismic inverse problem as a sequence of before stack migrations," 1983.

[73] J. F. Claerbout, "Toward a unified theory of reflector mapping," *GEOPHYSICS*, vol. 36, no. 3, pp. 467–481, 1st Jun. 1971.

[74] O. Gauthier, J. Virieux and A. Tarantola, "Two-dimensional nonlinear inversion of seismic waveforms: Numerical results," *GEOPHYSICS*, vol. 51, no. 7, pp. 1387–1403, Jul. 1986.

[75] C. Bunks, F. M. Saleck, S. Zaleski and G. Chavent, "Multiscale seismic waveform inversion," *GEOPHYSICS*, vol. 60, no. 5, pp. 1457–1473, 1st Oct. 1995.

[76] Z. Koren, K. Mosegaard, E. Landa, P. Thore and A. Tarantola, "Monte carlo estimation and resolution analysis of seismic background velocities," *J. Geophys. Res. Solid Earth*, vol. 96, pp. 20 289–20 299, B12 1991.

[77] S. Jin and R. Madariaga, "Background velocity inversion with a genetic algorithm," *Geophys. Res. Lett.*, vol. 20, no. 2, pp. 93–96, 1993.

[78] K. Mosegaard and A. Tarantola, "Monte carlo sampling of solutions to inverse problems," *J. Geophys. Res. Solid Earth*, vol. 100, pp. 12 431–12 447, B7 1995.

[79] M. Sambridge and G. Drijkoningen, "Genetic algorithms in seismic waveform inversion," *Geophys. J. Int.*, vol. 109, no. 2, pp. 323–342, 1st May 1992.

[80] P. Mora, "Nonlinear two-dimensional elastic inversion of multioffset seismic data," *GEOPHYSICS*, vol. 52, no. 9, pp. 1211–1228, Sep. 1987.

[81] P. Mora, "Elastic wave-field inversion of reflection and transmission data," *GEOPHYSICS*, vol. 53, no. 6, pp. 750–759,

[82] R. G. Pratt and M. H. Worthington, "Inverse theory applied to multi-source cross-hole tomography.," *Geophys. Prospect.*, vol. 38, no. 3, pp. 287–310, 1990.

[83] R. G. Pratt, "Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model," *GEOPHYSICS*, vol. 64, no. 3, pp. 888–901, May 1999.

[84] L. Sirgue, "The importance of low frequency and large offset in waveform inversion," presented at the 68th EAGE Conference and Exhibition incorporating SPE EUROPEC 2006, European Association of Geoscientists & Engineers, 12th Jun. 2006, cp.

[85] J. Virieux, "P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method," *GEOPHYSICS*, vol. 51, no. 4, pp. 889–901, Apr. 1986.

[86] D.-J. Min, C. Shin, R. G. Pratt and H. S. Yoo, "Weighted-averaging finite-element method for 2d elastic wave equations in the frequency domain," *Bull. Seismol. Soc. Am.*, vol. 93, no. 2, pp. 904–921, 1st Apr. 2003.

[87] P. Danecek and G. Seriani, "An efficient parallel chebyshev pseudo-spectral method for large scale 3d seismic forward modelling," presented at the 70th EAGE Conference and Exhibition incorporating SPE EUROPEC 2008, European Association of Geoscientists & Engineers, 9th Jun. 2008, cp.

[88] R. Brossier, J. Virieux and S. Operto, "Parsimonious finite-volume frequency-domain method for 2-d p–SV-wave modelling," *Geophys. J. Int.*, vol. 175, no. 2, pp. 541–559, 1st Nov. 2008.

[89] A. Fichtner, H. Igel, H.-P. Bunge and B. L. N. Kennett, "Simulation and inversion of seismic wave propagation on continental scales based on a spectral-element method," *JNAIAM*, vol. 4, no. 1, pp. 11–22, 2009.

[90] H. Igel, *Computational Seismology: A Practical Introduction*. Oxford University Press, 2017, ISBN: 978-0-19-871740-9.

[91] T. Küstner, A. Liebgott, L. Mauch *et al.*, "Automated reference-free detection of motion artifacts in magnetic resonance images," *Magn. Reson. Mater. Phy.*, vol. 31, no. 2, pp. 243–256, 1st Apr. 2018.

[92] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomed. Signal Process. Control*, vol. 27, pp. 145–154, 1st May 2016.

[93] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2002, pp. IV–3313–IV–3316.

[94] A. Eskicioglu and P. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[95] Z. Wang, A. Bovik, H. Sheikh and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[96] J. Miao, D. Huo and D. L. Wilson, "Quantitative image quality evaluation of MR images using perceptual difference models," *Med. Phys.*, vol. 35, no. 6, pp. 2541–2553, 2008.

[97] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[98] J. Dutta, S. Ahn and Q. Li, "Quantitative statistical methods for image quality assessment," *Theranostics*, vol. 3, no. 10, pp. 741–756, 4th Oct. 2013.

[99] L. Gebraad, C. Boehm and A. Fichtner, "Bayesian elastic full-waveform inversion using hamiltonian monte carlo," *J. Geophys. Res. Solid Earth*, vol. 125, no. 3, e2019JB018428, 2020.

[100] A. Fichtner and T. v. Leeuwen, "Resolution analysis by random probing," *J. Geophys. Res. Solid Earth*, vol. 120, no. 8, pp. 5549–5573, 2015.

[101] Q. Liu, S. Beller, W. Lei, D. Peter and J. Tromp, "Pre-conditioned BFGS-based uncertainty quantification in elastic full-waveform inversion," *Geophys. J. Int.*, vol. 228, no. 2, pp. 796–815, 1st Feb. 2022.

[102] X. Zhang and A. Curtis, "Variational full-waveform inversion," *Geophys. J. Int.*, vol. 222, no. 1, pp. 406–411, 1st Jul. 2020.

[103] X. Zhang and A. Curtis, "Seismic tomography using variational inference methods," *J. Geophys. Res. Solid Earth*, vol. 125, no. 4, e2019JB018589, 2020.

[104] J. Fessler and A. Hero, "Cramer-rao lower bounds for biased image reconstruction," in *Proceedings of 36th Midwest Symposium on Circuits and Systems*, Aug. 1993, 253–256 vol.1.

[105] H. H. Barrett, D. W. Wilson and B. M. W. Tsui, "Noise properties of the EM algorithm. i. theory," *Phys. Med. Biol.*, vol. 39, no. 5, pp. 833–846, May 1994.

[106] J. Fessler, "Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. Image Process.*, vol. 5, no. 3, pp. 493–506, Mar. 1996.

[107] J. Qi, "A unified noise analysis for iterative image estimation," *Phys. Med. Biol.*, vol. 48, no. 21, pp. 3505–3519, Oct. 2003.

[108] R. Barbano, C. Zhang, S. Arridge and B. Jin, "Quantifying model uncertainty in inverse problems via bayesian deep gradient descent," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 1392–1399.

[109] F. Tonolini, J. Radford, A. Turpin, D. Faccio and R. Murray-Smith, "Variational inference for computational imaging inverse problems," *J. Mach. Learn. Res.*, vol. 21, no. 179, pp. 1–46, 2020.

[110] D. Narnhofer, A. Effland, E. Kobler, K. Hammernik, F. Knoll and T. Pock, "Bayesian uncertainty estimation of learned variational MRI reconstruction," *arXiv:2102.06665 [cs, eess]*, 12th Feb. 2021. arXiv: 2102.06665.

[111] M. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal and A. Srivastava, "Fast and scalable bayesian deep learning by weight-perturbation in adam," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, 3rd Jul. 2018, pp. 2611–2620.

[112] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114 [cs, stat]*, 1st May 2014. arXiv: 1312.6114.

[113] R. Ranganath, S. Gerrish and D. Blei, "Black box variational inference," in *Artificial Intelligence and Statistics*, PMLR, 2nd Apr. 2014, pp. 814–822.

[114] C. Blundell, J. Cornebise, K. Kavukcuoglu and D. Wierstra, "Weight uncertainty in neural networks," *arXiv:1505.05424 [cs, stat]*, 21st May 2015. arXiv: 1505.05424.

[115] R. T. Q. Chen, Y. Rubanova, J. Bettencourt and D. Duvenaud, "Neural ordinary differential equations," *arXiv:1806.07366 [cs, stat]*, 19th Jun. 2018. arXiv: 1806.07366.

[116] R. P. Feynman, *The Feynman lectures on physics Vols. 1-2*, 6th ed. Perseus Books, 1977, ISBN: 0-201-02010-6-H.

[117] G. Clement, P. White and K. Hynynen, "Enhanced ultrasound transmission through the human skull using shear mode conversion," *J. Acoust. Soc. Am.*, vol. 115, no. 3, pp. 1356–1364, 2004.

[118] S. Pichardo, C. Moreno-Hernández, R. A. Drainville, V. Sin, L. Curiel and K. Hynynen, "A viscoelastic model for the prediction of transcranial ultrasound propagation: Application for the estimation of shear acoustic properties in the human skull," *Phys. Med. Biol.*, vol. 62, no. 17, pp. 6938–6962, Aug. 2017.

[119] M. Warner, A. Ratcliffe, T. Nangoo *et al.*, "Anisotropic 3d full-waveform inversion," *GEOPHYSICS*, vol. 78, no. 2, R59–R80, 1st Mar. 2013.

[120] J.-F. Aubry, M. Tanter, M. Pernot, J.-L. Thomas and M. Fink, "Experimental demonstration of noninvasive transskull adaptive focusing based on prior computed tomography scans," *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 84–93, 1st Jan. 2003.

[121] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2007, ISBN: 978-0-89871-629-0.

[122] Z. Miao and J. Zhang, "Reducing error accumulation of optimized finite-difference scheme using the minimum norm," *GEOPHYSICS*, vol. 85, no. 5, T275–T291, 21st Jun. 2020.

[123] Z. He, J. Zhang and Z. Yao, "Determining the optimal coefficients of the explicit finite-difference scheme using the remez exchange algorithm," *GEOPHYSICS*, vol. 84, no. 3, S137–S147, 24th Jan. 2019, ISSN: 0016-8033.

[124] K. Kumari, R. Bhattacharya and D. A. Donzis, "A unified approach for deriving optimal finite differences," *Journal of Computational Physics*, vol. 399, p. 108 957, 15th Dec. 2019.

[125] J.-H. Zhang and Z.-X. Yao, "Optimized finite-difference operator for broadband seismic wave modeling," *GEOPHYSICS*, vol. 78, no. 1, A13–A18, 21st Dec. 2012.

[126] M. I. Iacono, E. Neufeld, E. Akinnagbe *et al.*, "MIDA: A multimodal imaging-based detailed anatomical model of the human head and neck," *PLOS ONE*, vol. 10, no. 4, e0124126, 22nd Apr. 2015.

[127] R. G. Pratt, C. Shin and G. J. Hick, "Gauss–newton and full newton methods in frequency–space seismic waveform inversion," *Geophysical Journal International*, vol. 133, no. 2, pp. 341–362, May 1998.

[128] R. Cockett, S. Kang, L. J. Heagy, A. Pidlisecky and D. W. Oldenburg, "SimPEG: An open source framework for simulation and gradient based parameter estimation in geophysical applications," *Comput. Geosci.*, vol. 85, pp. 142–154, 1st Dec. 2015.

[129] R. J. Hewett and L. Demanet, "{PySIT}: {Python} seismic imaging toolbox," *Zenodo*, Jan. 2020.

[130] B. Zehner, O. Hellwig, M. Linke, I. Görz and S. Buske, "Rasterizing geological models for parallel finite difference simulation using seismic simulation as an example," *Comput. Geosci.*, vol. 86, pp. 83–91, 1st Jan. 2016.

[131] A. Paszke, S. Gross, F. Massa *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[132] C. Cueto, L. Guasch, J. Cudeiro *et al.*, "Spatial response identification enables robust experimental ultrasound computed tomography," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 1, pp. 27–37, Jan. 2022.

[133] S. Arridge, P. Maass, O. Öktem and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica*, vol. 28, pp. 1–174, May 2019.

[134] C. Szegedy, W. Zaremba, I. Sutskever *et al.*, "Intriguing properties of neural networks," *arXiv:1312.6199 [cs]*, 19th Feb. 2014. arXiv: 1312.6199.

[135] S. Bhadra, V. A. Kelkar, F. J. Brooks and M. A. Anastasio, "On hallucinations in tomographic image reconstruction," *arXiv:2012.00646 [cs, eess]*, 1st Apr. 2021. arXiv: 2012.00646.

[136] H. Carrillo, M. Millardet, T. Carlier and D. Mateus, "Low-count PET image reconstruction with bayesian inference over a deep prior," in *Medical Imaging 2021: Image Processing*, vol. 11596, International Society for Optics and Photonics, 15th Feb. 2021, p. 115960V.

[137] Y. W. Teh, A. H. Thiery and S. J. Vollmer, "Consistency and fluctuations for stochastic gradient langevin dynamics," *J. Mach. Learn. Res.*, vol. 17, no. 7, pp. 1–33, 2016, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v17/teh16a.html.

[138] A. Fichtner, H. .-. Bunge and H. Igel, "The adjoint method in seismology: I. theory," *Phys. Earth Planet. Inter.*, vol. 157, no. 1, pp. 86–104, 14th Aug. 2006.

[139] D. J. Rezende, S. Mohamed and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning*, PMLR, 18th Jun. 2014, pp. 1278–1286.

[140] T. Salimans and D. A. Knowles, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, Dec. 2013.

[141] K. Chrapkiewicz, M. Paulatto, B. A. Heath *et al.*, "Magma chamber detected beneath an arc volcano with full-waveform inversion of active-source seismic data," *Geochemistry, Geophysics, Geosystems*, vol. 23, no. 11, e2022GC010475, 2022.

[142] J. Nocedal, *Numerical optimization* (Series: Springer series in operations research.), 2nd ed. Springer, 2006, ISBN: 0-387-40065-6.

[143] C. Cueto, O. Bates, G. Strong *et al.*, "Stride: A flexible software platform for high-performance ultrasound computed tomography," *Comput. Methods Programs Biomed.*, vol. 221, p. 106 855, 1st Jun. 2022.

[144] E. R. Myers, P. Moorman, J. M. Gierisch *et al.*, "Benefits and harms of breast cancer screening: A systematic review," *JAMA*, vol. 314, no. 15, pp. 1615–1634, 20th Oct. 2015.

[145] A. Pietrobelli, C. Formica, Z. Wang and S. B. Heymsfield, "Dual-energy x-ray absorptiometry body composition model: Review of physical concepts," *Am. J. Physiol. Endocrinol. Metab.*, vol. 271, no. 6, E941–E951, 1st Dec. 1996.

[146] T. Hopp, A. Stromboni, N. Duric and N. V. Ruiter, "Evaluation of breast tissue characterization by ultrasound computer tomography using a 2d/3d image registration with mammograms," in *2013 IEEE International Ultrasonics Symposium (IUS)*, Jul. 2013, pp. 647–650.

[147] A. J. Y. Chee, C. K. Ho, B. Y. S. Yiu and A. C. H. Yu, "Walled carotid bifurcation phantoms for imaging investigations of vessel wall motion and blood flow dynamics," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 63, no. 11, pp. 1852–1864, Nov. 2016.

[148] D. Tran, R. Ranganath and D. M. Blei, "The variational gaussian process," *arXiv:1511.06499 [cs, stat]*, 17th Apr. 2016. arXiv: 1511.06499.

[149] L. Ruthotto and E. Haber, "Deep neural networks motivated by partial differential equations," *arXiv:1804.04272 [cs, math, stat]*, 10th Dec. 2018. arXiv: 1804.04272.

[150] D. A. Lichtenstein, "Lung ultrasound in the critically ill," *Ann. Intensive Care*, vol. 4, no. 1, p. 1, 9th Jan. 2014.

[151] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Pol. J. Radiol.*, vol. 80, pp. 93–106, 23rd Feb. 2015.

[152] P. Hermanek, J. S. Rathore, V. Aloisi and S. Carmignato, "Principles of x-ray computed tomography," in *Industrial X-Ray Computed Tomography*, Cham: Springer International Publishing, 2018, pp. 25–67, ISBN: 978-3-319-59573-3.

[153] J. P. De Wilde, J. A. Lunt and K. Straughan, "Information in magnetic resonance images: Evaluation of signal, noise and contrast," *Med. Biol. Eng. Comput.*, vol. 35, no. 3, pp. 259–265, 1st May 1997.

[154] M. Fuderer, "The information content of MR images," *IEEE Trans. Med. Imaging*, vol. 7, no. 4, pp. 368–380, Dec. 1988.

[155] M. C. Hemmsen, M. M. Petersen, S. I. Nikolov, M. B. Nielsen and J. A. Jensen, "Ultrasound image quality assessment: A framework for evaluation of clinical image quality," in *Medical Imaging 2010: Ultrasonic Imaging, Tomography, and Therapy*, vol. 7629, SPIE, 12th Mar. 2010, pp. 105–116.

[156] R. A. Pizarro, X. Cheng, A. Barnett *et al.*, "Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm," *Front. Neuroinf.*, vol. 10, 2016.

[157] B. L. Eck, R. Fahmi, K. M. Brown *et al.*, "Computational and human observer image quality evaluation of low dose, knowledge-based CT iterative reconstruction," *Med. Phys.*, vol. 42, no. 10, pp. 6098–6111, 2015.

[158] H. H. Barrett, J. Yao, J. P. Rolland and K. J. Myers, "Model observers for assessment of image quality.," *Proc. Natl. Acad. Sci.*, vol. 90, no. 21, pp. 9758–9765, Nov. 1993.

[159] F. Luporini, M. Louboutin, M. Lange *et al.*, "Architecture and performance of devito, a system for automated stencil computation," *ACM Trans. Math. Softw.*, vol. 46, no. 1, 6:1–6:28, 26th Apr. 2020.

[160] A. E. Flanders, L. M. Prevedello, G. Shih *et al.*, "Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge," *Radiology Artificial Intelligence*, vol. 2, no. 3, e190211, May 2020.

[161] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1st Jan. 1995.

[162] T. Cootes, G. Edwards and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[163] D. Rueckert, A. Frangi and J. Schnabel, "Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 22, no. 8, pp. 1014–1025, Aug. 2003.

[164] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

[165] S. Pichardo, V. W. Sin and K. Hynynen, "Multi-frequency characterization of the speed of sound and attenuation coefficient for longitudinal transmission of freshly excised human skulls," *Phys. Med. Biol.*, vol. 56, no. 1, pp. 219–250, Dec. 2010.

[166] R. Child, "Very deep VAEs generalize autoregressive models and can outperform them on images," presented at the International Conference on Learning Representations, 10th Feb. 2022.

[167] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub and K. Ugurbil, "The WU-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 15th Oct. 2013.