Generative Models for Preprocessing of Hospital Brain Scans

by

Mikael Brudfors

Prepared under the supervision of
Prof. John Ashburner and Prof. Parashkev Nachev

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Wellcome Centre for Human Neuroimaging
Institute of Neurology
University College London
London
UK

February 2020
I, Mikael Brudfors, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.
Abstract

I will in this thesis present novel computational methods for processing routine clinical brain scans. Such scans were originally acquired for qualitative assessment by trained radiologists, and present a number of difficulties for computational models, such as those within common neuroimaging analysis software. The overarching objective of this work is to enable efficient and fully automated analysis of large neuroimaging datasets, of the type currently present in many hospitals worldwide. The methods presented are based on probabilistic, generative models of the observed imaging data, and therefore rely on informative priors and realistic forward models.

The first part of the thesis will present a model for image quality improvement, whose key component is a novel prior for multimodal datasets. I will demonstrate its effectiveness for super-resolving thick-sliced clinical MR scans and for denoising CT images and MR-based, multi-parametric mapping acquisitions. I will then show how the same prior can be used for within-subject, intermodal image registration, for more robustly registering large numbers of clinical scans.

The second part of the thesis focusses on improved, automatic segmentation and spatial normalisation of routine clinical brain scans. I propose two extensions to a widely used segmentation technique. First, a method for this model to handle missing data, which allows me to predict entirely missing modalities from one, or a few, MR contrasts. Second, a principled way of combining the strengths of probabilistic, generative models with the unprecedented discriminative capability of deep learning. By introducing a convolutional neural network as a Markov random field prior, I can model nonlinear class interactions and learn these using backpropagation. I show that this model is robust to sequence and scanner variability. Finally, I show examples of fitting a population-level, generative model to various neuroimaging data, which can model, e.g., CT scans with haemorrhagic lesions.
Impact Statement

The automatic analysis of brain scans of large populations of subjects has the potential of uncovering patterns that closely map to underlying disease mechanisms and cognitive behaviour. Such analyses can be based on extracting tissue segmentations from the subjects’ images and aligning them to a common mean (or template), by the means of nonlinear registration techniques. This voxel-by-voxel alignment among subjects enables inter-subject comparisons of brains of different shapes and topography. This methodology has been used for decades, in hundreds of studies, to shed light on the structural correlates of neurological and psychiatric disorders.

The vast majority of these studies are performed on imaging data obtained in a controlled research environment. These images are of good quality and of uniform contrasts, making them fairly straightforward to segment and register with neuroimaging analysis packages, such as the statistical parametric mapping (SPM) software. Acquiring such images, however, is costly and therefore limits study size to, at maximum, a few hundred subjects. Hospital databases of images, on the other hand, exhibit a diverse range of patients and pathologies, where huge numbers of scans are available at cost neutrality. In the age of ‘Big Data’, small scale research studies, performed in controlled environments, could only dream of approaching the results that could be achieved on such very large datasets. The caveat is, routine clinical neuroimaging data is much more challenging to process due to its very large variability.

The aim of the work conducted in this thesis has been to facilitate segmentation and registration of routine clinical brain scans. The proposed models could, furthermore, enable complex inferential models of functional anatomy, which require large scale data to become computationally tractable. Software implementations of the generative models will be integrated into the SPM software package, which is used by many thousands of researchers all around the globe. Although the applications were focused on brain scans, the proposed models are general and could be applied
to other types of medical imaging data. The work carried out in this thesis could therefore result in further research in other domains of medical imaging.
Acknowledgements

When I arrived UCL, at the start of my MRes, my intention was to do research in computer-assisted interventions (CAI). That is what I had already been working on for a few years – a PhD focusing on CAI therefore seemed like a natural next step. This plan, however, was for various reasons not realised and I found myself in search for a new PhD project. Meeting with different teams of supervisors, I came across John Ashburner and Parashkev Nachev. After just a brief chat I realised that I would be fortunate if I could do my PhD with them – improving the segmentation and registration of clinical brain scans. That is what I ended up doing, therefore changing fields from CAI to medical image computing (MIC).

Working with John and Parashkev has been a great experience and I would like to thank them both for being so helpful, always available to quickly answer any question or query that I might have. John, my primary supervisor, is a warm, kind and ingenious person whom, without his guidance, I would have been utterly lost battling probabilistic models, variational Bayes and (supposedly) monotonically increasing lower bounds. I will truly miss the occasional code snippet, containing some brilliant idea John had coded up in an afternoon (or less). I was also very fortunate to see the arrival of a new post-doc in John’s group, Yaël Balbastre, while at the FIL. Yaël has been incredibly helpful to me during my PhD, as well as infinitely patient answering many of my, sometimes, stupid questions. I am amazed about how much he knows about the techniques that we are working on, and I owe him a huge thank you for all the help he has given me.

I would like to additionally thank all the wonderful people at the FIL, as well as everyone in John’s and Parashkev’s groups, many whom have made it very easy for me to get some good procrastinating in, whenever I was in need of a break in writing or coding. I am also forever grateful for the support of my family and friends in Sweden, as well as the ones I have made during my time in London. Finally, I would like to thank Mikaela for putting me on a new path, Lina for pushing me out
into the world and showing me that true kindness exists, and Amy for being there for me, through many ups and many downs.
# Contents

<table>
<thead>
<tr>
<th>List of Abbreviations</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>17</td>
</tr>
<tr>
<td>List of Tables</td>
<td>21</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>23</td>
</tr>
<tr>
<td>1.1 Motivations and Aims</td>
<td>23</td>
</tr>
<tr>
<td>1.2 ‘Big Data’ Neuroimage Analysis</td>
<td>24</td>
</tr>
<tr>
<td>1.3 Research vs Clinical Neuroimaging Data</td>
<td>27</td>
</tr>
<tr>
<td>1.4 Some Neuroimaging Modalities</td>
<td>29</td>
</tr>
<tr>
<td>1.4.1 Computed Tomography</td>
<td>30</td>
</tr>
<tr>
<td>1.4.2 Magnetic Resonance Imaging</td>
<td>33</td>
</tr>
<tr>
<td>1.4.3 Positron-Emission Tomography</td>
<td>37</td>
</tr>
<tr>
<td>1.5 A Modelling Framework for Medical Image Computing</td>
<td>38</td>
</tr>
<tr>
<td>1.5.1 Probabilistic Modelling</td>
<td>39</td>
</tr>
<tr>
<td>1.5.2 Inference in Probabilistic Models</td>
<td>41</td>
</tr>
<tr>
<td>1.5.3 Discriminative vs Generative Models</td>
<td>45</td>
</tr>
<tr>
<td>1.6 Publicly Available Neuroimaging Data</td>
<td>47</td>
</tr>
<tr>
<td>1.7 Thesis Overview</td>
<td>51</td>
</tr>
<tr>
<td>2 Resolution Recovery in Clinical Neuroimaging Data</td>
<td>57</td>
</tr>
</tbody>
</table>
Contents

4.2.1 Intensity Distribution ........................................ 137
4.2.2 Tissue Distribution ........................................ 138
4.2.3 Model Optimisation ........................................ 140
4.3 Population-Level Modelling of Images ......................... 141
  4.3.1 Learning the Intensity Model ............................. 142
  4.3.2 Learning the Template Model ............................. 147
4.4 A Model for Medical Image Translation ....................... 151
  4.4.1 Methods .................................................. 152
  4.4.2 Evaluation ............................................... 155
  4.4.3 Discussion ............................................... 161
4.5 A Convolutional Neural Network MRF ......................... 163
  4.5.1 Methods .................................................. 167
  4.5.2 Evaluation ............................................... 171
  4.5.3 Discussion ............................................... 175
4.6 Examples of Learned Models .................................. 176
  4.6.1 Predicting Fluid Intelligence ............................. 176
  4.6.2 Normalising Haemorrhagic Images ....................... 177
  4.6.3 Modelling Brain and Spine Data ......................... 181

5 Discussion .......................................................... 187
  5.1 Contributions ................................................ 187
  5.2 Future Directions ............................................ 189
    5.2.1 Unified Super-Resolution ............................... 189
    5.2.2 Improved MTV Registration ............................. 191
    5.2.3 Unified Generative Models ............................. 192

Bibliography .......................................................... 195
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMM</td>
<td>Alternating direction method of multipliers</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the curve</td>
</tr>
<tr>
<td>BS</td>
<td>B-spline</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional neural network</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional random field</td>
</tr>
<tr>
<td>CSF</td>
<td>Cerebrospinal fluid</td>
</tr>
<tr>
<td>CT</td>
<td>Computed tomography</td>
</tr>
<tr>
<td>DWI</td>
<td>Diffusion-weighted imaging</td>
</tr>
<tr>
<td>ECC</td>
<td>Entropy correlation coefficient</td>
</tr>
<tr>
<td>ELBO</td>
<td>Evidence lower bound</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation maximisation</td>
</tr>
<tr>
<td>FOT</td>
<td>First-order Tikhonov</td>
</tr>
<tr>
<td>FOV</td>
<td>Fields of view</td>
</tr>
<tr>
<td>FWHM</td>
<td>Full width at half maximum</td>
</tr>
<tr>
<td>GLM</td>
<td>General linear model</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>GM</td>
<td>Grey matter</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics processing unit</td>
</tr>
<tr>
<td>HR</td>
<td>High-resolution</td>
</tr>
<tr>
<td>HU</td>
<td>Hounsfield unit</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler</td>
</tr>
<tr>
<td>LR</td>
<td>Low-resolution</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>MTV</td>
<td>Multi-channel total variation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MTw</td>
<td>Magnetic-transfer weighted</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum a posteriori</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual information</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov random field</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>NCC</td>
<td>Normalised cross correlation</td>
</tr>
<tr>
<td>NMI</td>
<td>Normalised mutual information</td>
</tr>
<tr>
<td>NMTV</td>
<td>Normalised multi-channel total variation</td>
</tr>
<tr>
<td>PDw</td>
<td>Proton-density weighted</td>
</tr>
<tr>
<td>PET</td>
<td>Positron-emission tomography</td>
</tr>
<tr>
<td>PVE</td>
<td>Partial volume effect</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak signal-to-noise ratio</td>
</tr>
<tr>
<td>RF</td>
<td>Radiofrequency</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root-mean-square error</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operator curve</td>
</tr>
<tr>
<td>qMRI</td>
<td>Quantitative MRI</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SPM</td>
<td>Statistical parametric mapping</td>
</tr>
<tr>
<td>T1w</td>
<td>T1-weighted</td>
</tr>
<tr>
<td>T2w</td>
<td>T2-weighted</td>
</tr>
<tr>
<td>TV</td>
<td>Total variation</td>
</tr>
<tr>
<td>VB</td>
<td>Variational Bayes</td>
</tr>
<tr>
<td>VBM</td>
<td>Voxel based morphometry</td>
</tr>
<tr>
<td>VEM</td>
<td>Variational expectation-maximisation</td>
</tr>
<tr>
<td>VTV</td>
<td>Vectorial total variation</td>
</tr>
<tr>
<td>WM</td>
<td>White matter</td>
</tr>
</tbody>
</table>
List of Figures

1.1 VBM data preprocessing ........................................... 25
1.2 Research vs clinical MRI ........................................... 30
1.3 Windowed CT images .............................................. 32
1.4 Instrumental variability found in CT scans ......................... 32
1.5 Examples of different MR contrasts. ................................ 34
1.6 Bias field corrupted MR scan ...................................... 36
1.7 Quantitative maps created with the hMRI-toolbox ............... 37
1.8 MRI and PET scans of patient with dementia ..................... 38
1.9 Three common medical image computations ........................ 39
1.10 Public neuroimaging data used in this thesis ..................... 48

2.1 Example of patient MR images ..................................... 60
2.2 Super-resolution to combine multiple LR images ............... 61
2.3 Graphical representation of the super-resolution model .......... 65
2.4 Histogram of MR image gradients ................................ 68
2.5 Super-resolution forward model ................................... 70
2.6 Rician mixture fits to MR image intensity histograms ........... 73
2.7 Line fit for estimating the regularisation parameter ............. 74
2.8 Super-resolution validation on simulated data .................... 80
2.9 Validation for estimating the regularisation parameter .......... 83
2.10 Validation for super-resolving different slice-thickness images 84
2.11 Validation for super-resolving with increasing number of LR images 85
2.12 Conjugate gradients vs approximate Newton method . . . . . . . . 86
2.13 Baseline vs super-resolution on hospital MRI . . . . . . . . . . . . 87
2.14 Evaluation of the super-resolution method on clinical data . . . . . 88
2.15 Super-resolution: Age distribution and prediction results . . . . . 91
2.16 Denoising clinical CT scans . . . . . . . . . . . . . . . . . . . . . 92
2.17 Denoising: Age distribution and prediction results . . . . . . . . . 95
2.18 CT image, before and after denoising . . . . . . . . . . . . . . . . 96
2.19 qMRI map creation . . . . . . . . . . . . . . . . . . . . . . . . . . 97
2.20 Results for improving SNR in qMRI maps . . . . . . . . . . . . . . 98
3.1 Patient scans before and after image registration . . . . . . . . . . . 105
3.2 Pairwise vs groupwise image registration . . . . . . . . . . . . . . 106
3.3 Registration of bias field corrupted MR image . . . . . . . . . . . . 107
3.4 Modulating the NMTV cost function . . . . . . . . . . . . . . . . . 114
3.5 CT and MRI intensity histograms . . . . . . . . . . . . . . . . . . . 117
3.6 Simulation process on BrainWeb data . . . . . . . . . . . . . . . . . 119
3.7 NMTV cost as a function of translation . . . . . . . . . . . . . . . . 119
3.8 Simulated BrainWeb images, before and after NMTV registration . . 122
3.9 Voxel-wise contribution to the NMTV cost function . . . . . . . . . 122
3.10 Results from 2,000 registration simulations . . . . . . . . . . . . . . 124
3.11 Comparing registration error between NMTV and MI . . . . . . . . 125
3.12 RIRE MRI images, before and after NMTV registration . . . . . . . 127
3.13 RIRE images, before and after NMTV registration . . . . . . . . . . 128
4.1 Modelling gray matter decrease with age . . . . . . . . . . . . . . . 138
4.2 Probabilistic template used by the SPM12 segmentation routine . . 139
4.3 Population-level modelling of brain scans . . . . . . . . . . . . . . . 141
4.4 Demonstrating computing two image quality metrics . . . . . . . . 156
4.5 Learned model for MR image translation . . . . . . . . . . . . . . . 156
4.6 Inferring MR images with differing FOV . . . . . . . . . . . . . . . . 157
4.7  Inferring non-acquired MR contrasts  . . . . . . . . . . . . . . . . . 158
4.8  MRI to CT translation  . . . . . . . . . . . . . . . . . . . . . . . . 160
4.9  MR images from two public datasets  . . . . . . . . . . . . . . . . 165
4.10 Architecture of the MRF CNN  . . . . . . . . . . . . . . . . . . . . 171
4.11 MRF CNN training data and results  . . . . . . . . . . . . . . . . 173
4.12 MRF CNN validation  . . . . . . . . . . . . . . . . . . . . . . . . . 174
4.13 Template learned from ABCD subjects  . . . . . . . . . . . . . . . 178
4.14 Intensity model learned from ABCD subjects  . . . . . . . . . . . . 179
4.15 Example ABCD image  . . . . . . . . . . . . . . . . . . . . . . . . 179
4.16 Template learned from haemorrhagic CT data  . . . . . . . . . . . . 180
4.17 Intensity model learned from haemorrhagic CT data  . . . . . . . . 181
4.18 Example normalised CT image  . . . . . . . . . . . . . . . . . . . . 182
4.19 Example IXI subject  . . . . . . . . . . . . . . . . . . . . . . . . . 183
4.20 Template learned from IXI subjects  . . . . . . . . . . . . . . . . 183
4.21 Example normalised IXI subjects  . . . . . . . . . . . . . . . . . . . 184
4.22 Average of normalised IXI subjects  . . . . . . . . . . . . . . . . . 185
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The Hounsfield scale</td>
<td>31</td>
</tr>
<tr>
<td>1.2</td>
<td>Pros and cons of publicly available neuroimaging datasets</td>
<td>52</td>
</tr>
<tr>
<td>2.1</td>
<td>Validation for estimating the noise precision parameter</td>
<td>82</td>
</tr>
<tr>
<td>2.2</td>
<td>Validation for single- vs multi-channel super-resolution</td>
<td>86</td>
</tr>
<tr>
<td>2.3</td>
<td>Validation for segmenting super-resolved images</td>
<td>88</td>
</tr>
<tr>
<td>2.4</td>
<td>Super-resolution results for predicting age and sex from segmentations</td>
<td>90</td>
</tr>
<tr>
<td>2.5</td>
<td>Denoising results for predicting age and sex from segmentations</td>
<td>94</td>
</tr>
<tr>
<td>3.1</td>
<td>NMTV results on simulated data</td>
<td>121</td>
</tr>
<tr>
<td>3.2</td>
<td>Average registration runtime</td>
<td>123</td>
</tr>
<tr>
<td>3.3</td>
<td>Registration results on clinical MRI data</td>
<td>127</td>
</tr>
<tr>
<td>3.4</td>
<td>Registration results on clinical MRI, CT and PET data</td>
<td>129</td>
</tr>
<tr>
<td>4.1</td>
<td>Symbols for population-level modelling of brain scans</td>
<td>142</td>
</tr>
<tr>
<td>4.2</td>
<td>Results for inferring MR images with differing FOV</td>
<td>158</td>
</tr>
<tr>
<td>4.3</td>
<td>Results for inferring MR image contrasts</td>
<td>159</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Contents

1.1 Motivations and Aims ........................................... 23
1.2 ‘Big Data’ Neuroimage Analysis .............................. 24
1.3 Research vs Clinical Neuroimaging Data ................. 27
1.4 Some Neuroimaging Modalities .............................. 29
  1.4.1 Computed Tomography .................................. 30
  1.4.2 Magnetic Resonance Imaging .......................... 33
  1.4.3 Positron-Emission Tomography ...................... 37
1.5 A Modelling Framework for Medical Image Computing . . 38
  1.5.1 Probabilistic Modelling ................................ 39
  1.5.2 Inference in Probabilistic Models .................... 41
  1.5.3 Discriminative vs Generative Models .............. 45
1.6 Publicly Available Neuroimaging Data ..................... 47
1.7 Thesis Overview .............................................. 51

1.1 Motivations and Aims

The motivation for the work conducted in this thesis was to facilitate the use of predictive models – implemented by some machine learning algorithm – on routine
In this thesis, the images that make up this data have mostly been acquired via magnetic resonance imaging (MRI) or computed tomography (CT). Currently, much of the information contained within hospital datasets of such images lies latent, its potential uses yet untapped. Putting confidentiality issues aside, the major reason for this lesser focus is the necessary preprocessing steps (often involving some form of registration and/or segmentation) being far more challenging to perform, than for data acquired in a research context. Such preprocessing is the focus of this thesis.

In the following sections, I will introduce neuroimaging analysis, with a focus on large-scale processing, and research- vs clinical-grade data. The type of medical data that will be processed in this thesis will also be described. Then, I will give an overview of the modelling framework that was adopted. Finally, I will give a summary of the publicly available neuroimaging data that I used in this work.

1.2 ‘Big Data’ Neuroimage Analysis

Neuroimaging data can be analysed in order to investigate the structure and function of the brain. Automated analysis of neuroimaging data enables studies that would be either very time consuming, or simply not possible, if performed manually by a researcher. These automated methods can be used to, e.g., examine changes over time due to ageing and neurodegenerative disorders (Reuter et al., 2012), make predictions that separate healthy subjects from those with some disease (Klöppel et al., 2008), delineate the location of pathology (Seghier et al., 2008), and much more.

One popular neuroimaging analysis technique investigates differences in tissue composition between groups of subjects and is known as voxel based morphometry (VBM; Ashburner and Friston (2000); Mechelli et al. (2005)). VBM is a fully automated technique that allows investigating focal differences in brain anatomy, using the approach of statistical parametric mapping (SPM; Friston et al. (1994)). The statistical analysis uses the general linear model (GLM) to identify regions of

---

1Please keep in mind that although the application of this thesis are brain scans, the proposed models should be general enough to find usage on other types of clinical medical imaging data.
1.2. ‘Big Data’ Neuroimage Analysis

Figure 1.1: The image preprocessing steps for a VBM study involves, for each subject: (1) brain tissue extraction by some segmentation routine; (2) normalisation to a common space via nonlinear image registration; and (3), smoothing with a Gaussian kernel. Both segmentation and normalisation are non-trivial operations. Their accuracy may vary greatly depending on both the instrumental and biological variability of the input data.

brain tissue concentration that are significantly related to the particular effects under study. VBM methods largely owe their success to the underlying preprocessing of the neuroimaging data, which makes it possible to test for statistical significance using a relatively simple mass-univariate technique. In short, the preprocessing steps involved in a VBM study are to segment the grey matter (GM), then spatially normalise all images to the same stereotactic space, and finally smooth. Smoothing has the effect of rendering the data more normally distributed, increasing the validity of statistical tests; due to the central limit theorem

\[^2^\] The size of the smoothing kernel should be comparable to the size of the expected regional differences between the groups of brains being compared. The smoothing step also helps to compensate for the inexact nature of the spatial normalisation. A diagram detailing the processing steps of VBM is shown in Figure 1.1. VBM methods have proved valuable at uncovering structural differences between populations, for example, of different age (Good et al., 2002) and professions (Maguire et al., 2000). It is however sensitive to intersubject misalignments (Bookstein, 2001). The risk of such misalignment can increase when the input data has poor quality, or if subjects have pathology (Crinion et al., 2007).

\[^2^\] The central limit theorem says that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution, even if the original variables themselves are not normally distributed.
Another neuroimaging analysis technique, which specifically targets pathological data, is lesion-deficit mapping. Lesion-deficit mapping can be used to correlate patterns of impairment in patients, with a knowledge of exactly what parts of the nervous system are damaged. This is achieved, in its most simple form, by taking the overlap of a set of lesions and contrasting the peak with that derived from another control set of patients (Robertson et al., 1988). Also more sophisticated methods have been derived, using both univariate and multivariate techniques (Bates et al., 2003; Husain and Nachev, 2007; Mah et al., 2014a). Lesion-deficit mapping naturally enables the investigation of how behaviour and cognition are influenced when a region of the brain is disrupted, as inactivating brain areas experimentally cannot be done, easily, in humans. Localising human brain function by studying the correlation between a behavioural disorder and the location of brain lesions has greatly advanced the understanding of brain function (Rorden and Karnath, 2004). Investigating a large number of subjects with lesion-deficit mapping could result in the decipherment of population patterns that closely map to underlying disease mechanisms and cognitive behaviour (Insel et al., 2010).

Lesion-deficit mapping is, however, complicated by the need for time-consuming and subjective manual segmentation of the lesions in the imaging data. This greatly limits the practicability of the approach. In order to perform experiments on larger numbers of subjects, automated methods for segmenting and spatially normalising the images are required. However, some of the most widely used software packages for performing these operations struggle to do so, accurately, when faced with lesioned data. This is because such scans are found mostly in clinical data, which is challenging to process. To better enable studies of neuropsychology, such as lesion-deficit mapping, the methodology of current software packages needs improving to better handle clinical-grade data.

A large number of neuroimaging software packages exist for analysing brain

---

3 A technique known as transcranial magnetic stimulation can ‘deactivate’ a brain region short-term (i.e., virtual lesions), by using a changing magnetic field. This causes electric current at a specific area of the brain (through electromagnetic induction) and a disruption to that area’s pattern of activation.
1.3 Research vs Clinical Neuroimaging Data

In this thesis, when clinical neuroimaging data is discussed, what is meant are brain scans acquired in a hospital setting, for patient diagnosis, treatment or follow-up; scans, using various methods. Three of the more popular software packages are SPM (Ashburner and Friston, 2005), FSL (Smith et al., 2004) and FreeSurfer (Fischl et al., 2004). Although there are differences in how each package models and processes data – and what type of analysis they can perform – they have all been shown to work well on images with small, almost isotropic voxels, having high contrast between brain tissues, e.g., GM and white matter (WM), (Kazemi and Noorizadeh, 2014; Heinen et al., 2016). As they require more time and higher field strengths, these types of features are mostly found in images acquired in a controlled, research context. Images with such properties are, however, just a tiny subset of the neuroimaging data acquired worldwide. The majority of neuroimaging data is of routine clinical type, collected in huge numbers in hospitals everyday.

Routine clinical medical imaging data exhibits a diverse range of pathologies, which small scale research studies can only dream of approaching. There is therefore an incredible potential associated with applying machine learning models to such datasets, e.g., for studies of neuropsychology in neuroimaging (Mah et al., 2014a). Currently, the use of ‘Big Data’ machine learning models (e.g., deep neural networks) is exploding (LeCun et al., 2015). This is due to a number of reasons, such as improved hardware, increased amounts of data and refined methodology. If these models could be deployed, successfully, on large datasets of clinical data, the aforementioned potential could be tapped into (Smith and Nichols, 2018). However, to do so often requires the type of image processing discussed earlier: segmentation and registration. But, for reasons covered in the next section, these operations are much more challenging to perform successfully on clinical data, than for its research counterpart. It is therefore difficult to access the information contained in these large datasets. This in turn impedes the development of complex inferential models that require large scale data to become computationally tractable.
not data acquired with the aim of being used in a research study. While, clinical data is optimised with a patient’s need in mind, research data is optimised in order to establish facts and reach new conclusions. This leads to very different imaging characteristics between these two types.

Research-grade data is most often acquired by investigators interested in applying some automated analysis method. They do therefore tend to acquire high-resolution volumetric images, which are optimal for detecting the signal of interest. This naturally makes large-scale studies, with thousands of subjects, difficult to perform due to the expense of scanning with such parameters. For example, UK BioBank took a decade of planning and acquisition (Ollier et al., 2005), at a cost of more than 60 million British pounds (Palmer, 2007). However, such studies can provide great opportunities for making new discoveries about the brain (Van Essen et al., 2013; Miller et al., 2016; Smith and Nichols, 2018). Note that research neuroimaging data is most often MRI since these type of images can be acquired without the need for radiating a patient, as would be the case with CT imaging.

Neuroimaging data found in hospitals do not have this problem, as huge amounts of population-representative scans exist (Roobottom et al., 2010), accumulated over years of clinical service. Furthermore, these scans are available for research at cost neutrality and have a much higher prevalence of pathology, which is of interest if disease is to be investigated rather than normality. Neuroimaging data acquired in a clinical setting have much greater variability than its research counterparts, as clinicians do not target reproducibility, but aim at maximising sensitivity to a suspected pathology. This variability can be disentangled into biological and instrumental:

**Biological variability:** The biological variability relates to the diverse morphological variation in patient brains due to different clinical conditions, as well as a wide age age distribution.

**Instrumental variability:** The instrumental variability stems from: (1) clinicians

\[\text{Not the general population – rather the population who are likely to have had strokes, tumours, etc.}\]
preference for speed over volume resolution and therefore typically acquiring images with high in-plane resolution, but fewer and thicker slices (i.e., so that it is possible to quickly scroll through a patient’s scans); (2), the scans being acquired across a diverse set of sequences and modalities, sensitive to different tissue properties and, therefore, different pathologies; and (3), artefacts, which are misrepresentations of tissue structures produced by the imaging technique. These artefacts may be caused, e.g., by data acquisition errors (such as patient motion), or a reconstruction algorithm’s inability to represent the anatomy.

All in all, this leads to a huge diversity in clinical neuroimaging data. As an example, a comparison between MR images acquired in a research and a clinical context is shown in Figure 1.2.

The instrumental and biological variability is why applying automated analysis pipelines, successfully, to images acquired in a routine clinical setting is so much more challenging than applying them to research scans. For example, the software packages discussed in the previous section: SPM, FSL and FreeSurfer, were recently evaluated at the task of brain GM and WM volume estimation from isotropic and anisotropic volumes acquired in the same subjects. Results were improved by several percentage points when isotropic volumes were used over thick-sliced ones (Adduru et al., 2017). There is therefore a need for: (1) tools that could generate closer-to-research quality images from clinical scans, because this would enable using many already developed tools on clinical data; and (2), tools that can directly and accurately process such data.

1.4 Some Neuroimaging Modalities

Various medical imaging technologies exist in order to create visual representations, images, of the interior of the brain. These images are used for clinical analysis and medical intervention. This section will focus on the three modalities used to acquire the imaging data encountered in this thesis. The majority of images will be CT and MRI, but also some positron-emission tomography (PET) data will be encountered.
Figure 1.2: Comparing research and clinical MRI. The top row shows PD-weighted (PDw), T1-weighted (T1w) and T2-weighted (T2w) MR images of a subject from the high-resolution IXI dataset (described in Section 1.6). The three scans have close to 1 mm isotropic voxels. The bottom row shows FLAIR, T1w and T2w MR images of a subject from the clinical dataset that was used for the validation in Chapter 2. These subject’s scans have thick-slices (6.5 mm), with different slice-select directions (in parenthesis), and partial brain coverage.

CT scanning is more widely used than MRI, as it is cheaper to acquire. MRI does however not radiate a patient, meaning it incurs no dose of X-ray radiation. MRI does, however, use strong magnetic fields and, as doctors may not know whether a patient has any metal implants, CT may be preferred (e.g., on emergency patients who are unconscious). CT images show superior bone contrast, while MR images have very good soft tissue contrast. MRI and CT are structural modalities. A structural modality images the structure of the brain (e.g. shows contrast between different tissues). PET, on the other hand, is a functional modality. A functional modality measures changes in metabolism, blood flow, regional chemical composition and absorption. Next, these modalities will be described in more detail.

1.4.1 Computed Tomography

CT operates by using an X-ray generator, contained within the doughnut shaped gantry of the scanner, which rotates around a patient to produce 3D data. This data is then processed using a form of tomographic reconstruction to produce a series of cross-sectional images (Arridge 1999). Voxels in an image obtained by CT scanning are displayed in terms of attenuation: the mean reduction of the intensity of the X-ray beam as it traverses substances. So that a voxel is displayed according to the attenuation of the tissue (or tissues) that it corresponds to on the Hounsfield
Table 1.1: The Hounsfield scale for medical-grade CT scans of the brain (Buzug and Mihailidis, 2009; Heymsfield et al., 2005; Prokop and Galanski, 2003). Shown are the approximate Hounsfield units for various substances in an image.

<table>
<thead>
<tr>
<th>Substance</th>
<th>Approximate Hounsfield unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>-1,000</td>
</tr>
<tr>
<td>Scalp fat</td>
<td>-100</td>
</tr>
<tr>
<td>Cerebrospinal fluid &amp; ventricle</td>
<td>+10</td>
</tr>
<tr>
<td>White matter</td>
<td>+25</td>
</tr>
<tr>
<td>Blood</td>
<td>+30</td>
</tr>
<tr>
<td>Gray matter</td>
<td>+40</td>
</tr>
<tr>
<td>Clotted blood</td>
<td>+65</td>
</tr>
<tr>
<td>Soft tissue</td>
<td>+200</td>
</tr>
<tr>
<td>Bone</td>
<td>+200 – 3,000</td>
</tr>
<tr>
<td>Foreign object</td>
<td>+500 – 30,000</td>
</tr>
</tbody>
</table>

The Hounsfield scale is a quantitative scale of measuring X-ray attenuation. Hounsfield units (HU) for substances commonly found in CT scans of the brain are shown in Table 1.1 (Buzug and Mihailidis, 2009; Heymsfield et al., 2005; Prokop and Galanski, 2003). Water has an attenuation of 0 HU, while air is -1,000 HU. Bone covers a wider range where craniofacial bone typically is +200, cancellous bone is +700 HU, and cortical bone can reach over +3,000 HU. Foreign objects can have any value on the scale. Because HU units overlap for different substances it can be difficult to differentiate between tissue types based on the intensity values alone.

Modern computer monitors can only display 256 levels of grey, but CT images have a much larger dynamic range. Linear windows are therefore commonly used to display a portion of the dynamic range on a computer monitor. This reduction is typically done via a process of windowing. For example, CT images of the brain are commonly viewed with a brain window extending from 0 HU to 100 HU, while a bone window would range from 200 HU to 3,000 HU (see Figure 1.3).

Some major sources of instrumental variability in CT scans include (but are not limited to):

**Beam hardening:** Beam hardening originates from an X-ray beam, comprised of polychromatic energies, passing through an object and becomes ‘harder’ as the lower energy photons will be absorbed leaving only the higher energy photons. The resulting artefact has two distinct appearances, streaking or
Figure 1.3: Examples of windowing a CT image. Due to the high dynamic range of CT images they are usually windowed to better differentiate, visually, between different tissues of the brain. Here, a standard window that covers a large range of the Hounsfield scale is shown (a), as well as a brain (b) and bone (c) window.

Figure 1.4: Examples of some of the instrumental variability found in CT images. Shown are images that have variability in the form of: image noise (a), motion corruption (b), beam hardening (c), and variable slice-thickness (d: corrected, e: uncorrected).
1.4. Some Neuroimaging Modalities

dark bands and cupping artefacts.

**Motion:** Motion artefacts are the result of patient movement during scanning, commonly due to breathing or agitation secondary to intracranial pathology.

**Noise:** Limiting the X-ray dose for patient safety means that fewer photons are observed, which in turn leads to noise in the acquired CT image.

**Partial volume effects:** Partial volume effects (PVE) occurs when tissues of different attenuation are encompassed on the same CT voxel producing an attenuation proportional to the average HU of these tissues. This means that the Hounsfield value in a voxel is a combination of different tissues.

**Variable slice-thickness:** As CT scans often are acquired with a gantry tilt and thinner slices near the brain stem, a variable slice-thickness correction may need to be performed if combining slices from DICOM files into a single 3D volume. If uncorrected, the resulting reconstructed image can appear distorted.

Examples of the instrumental variability found in CT images are shown in Figure 1.4.

1.4.2 Magnetic Resonance Imaging

To obtain an MR image, a patient is placed inside a strong magnetic field. This magnetic field forces the protons in the body to align with the field direction. When a radiofrequency (RF) pulse is then sent through the patient, the protons are stimulated and spin out of equilibrium, straining against the pull of the magnetic field. To distinguish between spins at different positions in space an additional spatially dependent field is introduced. These fields, called gradients, are generated by coils installed inside the magnet bore, which produce magnetic fields that vary linearly in space. Gradient fields are used in three different ways for spatial encoding: slice

As part of the work conducted in this PhD, variable slice-thickness reconstruction of CT scans was added into the SPM12 software (see the `spm_dicom_convert` function).
selection is used to selectively excite the spins in a well-defined plane, while frequency and phase encoding produce an image in two dimensions. Because the ideal slice-selection profile is never a perfect rectangle, slices are rarely contiguous, but have either gaps or overlap. The slice profile is also what determines the slice thickness. When the RF pulse is turned off, a receive coil in the MRI scanner is able to detect the oscillation of the proton spins about the direction of the main magnetic field. The contrast between different tissues is determined by the rate at which excited atoms return to the equilibrium state.

An MRI sequence is a particular setting of RF pulses and gradients, resulting in a particular image appearance (or MR contrast/channel). Common MR contrasts are: T1-weighted (T1w), T2-weighted (T2w), proton-density weighted (PDw), fluid attenuated inversion recovery (FLAIR), blood-oxygen-level dependent (BOLD), and diffusion-weighted imaging (DWI). Some examples of images of different contrasts are shown in Figure 1.5. The listed sequences, however, are but a few of many used in clinical practice. This makes it difficult to develop general machine learning techniques for analysing MR images. In comparison to CT images, whose intensity values are encoded in the Hounsfield scale, tissues in MR images can have different intensity values even when the same sequence and the same scanner is used. That is, regular MRI is a non-quantitative imaging method.

Figure 1.5: Example axial slices of three different MR contrasts, of the same healthy subject (from the IXI dataset).
1.4. Some Neuroimaging Modalities

Some major sources of instrumental variability in MRI images are:

**Intensity inhomogeneity:** Intensity inhomogeneities are commonly present in MR images due to imperfections of the magnetic fields and inhomogeneous properties of the scanned object. The majority of intensity inhomogeneities appear as a low-frequency intensity variation, known as a bias field\(^6\) across the whole image. This variation results in the same tissue types having different intensity values, which can obstruct analysis. An example of intensity inhomogeneity in an MR scan is shown in Figure [1.6](#).

**Noise:** The noise in MR images may be due to numerous reasons: thermal agitation (the main source), field strength, RF pulses, RF coil, voxel volume, or receiver bandwidth.

**Thick slices:** The slice thickness is a parameter that can be selected when acquiring an MR image. This parameter will change the thickness of the MRI slice in millimetres. By increasing the slice thickness, many more different types of tissues will be collected in one 2D slice. This causes PVE.

All the above mentioned variability impedes automated analysis of MR images.

**Quantitative MRI**

In conventional MRI, the image intensity values have arbitrary units and the value in a given voxel will depend on a large number of factors (*e.g.*, sequence type, hardware effects and physical tissue properties). The voxel intensity values in MRI are therefore ‘meaningless’, in the sense that they do not represent some physical unit that can be compared between subjects – just as a consumer-type camera. This is one of the reason that learning from MR images is challenging, as intensity features do not contribute meaningful information. Methods for intensity normalisation of the imaging data are one way of dealing with this issue, bringing the intensities to a common scale across subjects.

\(^6\)Although the term ‘bias’ is usually meant to indicate something additive, I will in this thesis model the bias field as multiplicative.
The majority of intensity inhomogeneities in an MR scan appears as a low-frequency intensity variation known as bias field. This spurious intensity variation may significantly hamper automated segmentation routines. It can be difficult to identify the bias field from visual inspection of a greyscale MR scan (a). By increasing the range of the image colour map (b), it can be more clearly seen that the white matter has a non-uniform intensity distribution.

Quantitative MRI (qMRI) represents a paradigm shift in MRI where the scanner is no longer seen as a camera, but as a scientific measuring instrument (Cercignani et al., 2018). The acquired images are now rather maps of meaningful physical or chemical variables. This property may improve the sensitivity of multi-site studies and the capability of data-driven methods to extract patterns from datasets of images.

One instantiation of quantitative parameter map reconstruction is the hMRI-toolbox (Tabelow et al., 2019), which is used in this thesis when working with qMRI. The toolbox creates quantitative maps from unprocessed multi-echo T1w, PDw and magnetic-transfer weighted (MTw) spoiled gradient echo acquisitions by fitting a mathematical model known as ESTATICS (Weiskopf et al., 2014). The map creation module additionally corrects the qMRI estimates for spatial receive and transmit field inhomogeneities. For a detailed description of the map creation, see the appendix of Tabelow et al. (2019). Figure 1.7 shows example quantitative maps created with the hMRI-toolbox.
1.4. Some Neuroimaging Modalities

Figure 1.7: Sagittal views of the quantitative maps created with the hMRI-toolbox. The toolbox generates MT, PD, R₁ and R₂ maps. The voxel values of these maps are interpretable, meaning they quantify some physical unit of interest.

1.4.3 Positron-Emission Tomography

PET is a nuclear medicine\(^7\) functional imaging modality that is used to observe metabolic processes in the body. PET differs from other nuclear medicine techniques in that it detects metabolism within body tissues, whereas other types of nuclear medicine examinations detect the amount of a radioactive substance collected in body tissue in a certain location to examine the tissue’s function. PET is an important tool for both diagnosis (brain tumours, strokes, and neuron-damaging diseases all cause great changes in brain metabolism) and research (e.g., to map brain function or support drug development). One of the disadvantages of acquiring PET imaging is the operating cost, which largely results from the expense of producing the radioactively labelled tracers that are injected into the patient.

It is common practice to superimpose PET with MR images to produce special views (see Figure 1.8). This is known as image fusion or co-registration. These views allow a doctor to correlate and interpret information from two different ex-

\(^7\)Nuclear medicine is a medical speciality involving the application of radioactive substances in the diagnosis and treatment of disease.
Figure 1.8: MRI and PET scans in co-register (of a patient with dementia). This view allows a doctor to correlate and interpret information from two different exams on one image, which can lead to more precise information and accurate diagnoses.

Before the use of functional magnetic resonance imaging (fMRI) became widespread, PET imaging was the preferred method of functional brain imaging.

## 1.5 A Modelling Framework for Medical Image Computing

Medical image computing aims at developing computational and mathematical methods for solving problems pertaining to medical images and their use in biomedical research and clinical care. One way of looking at the general aim of medical image computing is as trying to find a solution to the problem: given some observed data ($x$), predict an unknown ($y$). The observed data could be an MR or CT image. The unknown could, for example, be a deformation aligning one image to some mean template (known as normalisation$^8$), a set of tissue types, or a non-degraded image; these operations are forms of registration, segmentation and denoising, respectively (see Figure 1.9). Mapping between inputs and outputs in such a way can be done using machine learning techniques. These techniques can often be ex-

---

$^8$Normalisation refers to the process of co-registering neuroimaging data (most often using non-linear methods) to some mean template, to overcome the issue of brain shape variability between different subjects.
1.5. A Modelling Framework for Medical Image Computing

Figure 1.9: Examples of three common medical image computations that can be framed in a probabilistic modelling framework: (1) the log Jacobian of a deformation computed with a nonlinear registration routine; (2) an MRI (top) segmented into six tissue classes (bottom); (3) a noisy MRI (top) that has been denoised (bottom).

pressed within the mathematical language of probability theory, which provides a principled framework for building machine learning models.

1.5.1 Probabilistic Modelling

In probabilistic modelling, the observed data are assumed to be drawn from some probability distribution. Machine learning models derived from probability theory possess several advantages (Jaynes, 2003):

- they have an inherent qualitative correspondence with common sense;
- they represent a degree of plausibility by real numbers;

and therefore provide a principled way of deriving models (or objective functions).

Inverse problems – the process of calculating, from a set of observations, the causal factors that produced them – can often be cast in a probabilistic modelling context. An inverse problem requires a forward model, which maps from some unknown parameters to observed data \(e.g., x = Ay\). For MRI, the forward model is based on the signal equation of the particular imaging sequence and the intrinsic tissue parameters; and for CT, the mathematical process that generated the images from the X-ray projection data, acquired at many different angles around a patient.

A model \(\mathcal{M}\) is in probabilistic modelling defined by some joint probability distribution over all observed and unknown variables \(p(x, y \mid \mathcal{M})\). The probability
distribution is constructed so as to express all forms of uncertainty and noise associated with the observed data. Unknown quantities can then be inferred from the model, in order to learn from data and make predictions. Commonly, just a single model is of interest and one can therefore write simply \( p(x, y) \). If multiple models are to be compared, this can be done within the framework of Bayesian model comparison or by cross-validation techniques (Bishop, 2006).

Two key components of probabilistic modelling are the product rule:

\[
p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x), \tag{1.1}
\]

and the sum rule (also known as the marginalisation rule):

\[
p(y) = \sum_x p(x, y) \quad \text{(for discrete } x), \tag{1.2}
\]
\[
p(y) = \int_x p(x, y)dx \quad \text{(for continuous } x), \tag{1.3}
\]

where \( p(x) \) is the probability of \( x \), \( p(x, y) \) is the joint probability of \( x \) and \( y \) and \( p(x \mid y) \) is the probability of \( x \) conditional on \( y \). Another important concept is independence; that is, two random variables are independent if the realisation of one does not affect the probability distribution of the other:

\[
p(x, y) = p(x)p(y). \tag{1.4}
\]

Combining the sum and product rules gives Bayes’ rule:

\[
p(y \mid x) = \frac{p(x, y)}{\sum_y p(x, y)} = \frac{p(x \mid y)p(y)}{p(x)}, \tag{1.5}
\]

whose individual components are denoted:

\[
\text{posterior} = \frac{\text{likelihood \times prior}}{\text{evidence}}. \tag{1.6}
\]

Note that depending on how the joint probability is factorised in the numerator of (1.5), and marginalised in the denominator, posteriors over different variables of the model can be formulated.
Inference in Probabilistic Models

The most complete characterisation of an unknown is expressed in terms of its posterior probability \( p(y \mid x) \). Once a model of the posterior is available, it is possible to assign observed data to the realisation of the unknown having the largest probability. A mode estimate can be computed as the peak of the posterior distribution. A more rigorous approach would involve inferring the full posterior distribution. However, the marginalisation in the denominator of Bayes’ rule is typically intractable, which makes it necessary to resort to approximation schemes. There exist numerous such methods for approximating various parts of Bayes’ rule. A few examples are (Barber, 2012): Laplaces method, Bayesian information criterion, Markov chain Monte Carlo, the junction tree algorithm and loopy belief propagation. All of these methods have their individual strengths and weaknesses. This section will however focus on one additional such technique known as variational Bayes (VB), which will be used extensively in Chapter 4.

Computing Mode Estimates

Obtaining mode estimates of a distribution is in general a much less computationally costly task than inferring complete distributions. Given \( N \) independent observations \( X = \{x_n\}_{n=1}^N \) (e.g., voxels) a mode estimate of the unknown is computed by:

\[
\hat{y} = \arg\max_y p(y \mid X) = \arg\max_y \left\{ \frac{p(y) \prod_{n=1}^N p(x_n \mid y)}{p(X)} \right\} = \arg\max_y \left\{ \prod_{n=1}^N p(x_n \mid y) p(y) \right\},
\]

where the denominator can be discarded as it is constant with respect to the unknowns and so does not affect the position of the maximum. This technique is known as maximum a posteriori (MAP) estimation. Maximum likelihood (ML) estimation can be seen as a special case of MAP estimation with a (possibly improper)
uniform prior distribution over the parameters, so that:

$$\hat{y} = \arg\max_y p(X \mid y)$$

$$= \arg\max_y \left\{ \prod_{n=1}^N p(x_n \mid y) \right\}.$$  \hfill (1.8)

Two rationales for including prior information into the model is: (1) to protect against overfitting, in which the model fails to predict future observations reliably; and (2), as a regulariser for ill-posed objective functions. The choice of prior is often based on some knowledge over the distribution of the unknowns. For example, that neighbouring voxels in an image should look similar. A prior distribution can also encode information from training data. For example, that the probability that a voxel in an MR image belongs to a certain tissue type of the brain \cite{Ashburner2005}.

Commonly, what is minimised is the negative logarithm of the ML or MAP formulation (known as minimum energy estimation). By the monotonic nature of the logarithmic function it is equivalent to maximising the posterior probability:

$$\hat{y} = \arg\max_y p(y \mid X)$$

$$= \arg\min_y -\ln p(y \mid X)$$

$$= \arg\min_y \mathcal{E},$$ \hfill (1.9)

where \(\mathcal{E}\) is known as the energy. Writing out the energy of the MAP formulation:

$$\mathcal{E} = -\sum_{n=1}^N \ln p(x_n \mid y) - \ln p(y),$$ \hfill (1.10)

perhaps more clearly demonstrates the prior distribution’s role as a regularising term. For example, for a Gaussian likelihood \(N(x \mid Ay, \tau_1 I)\) and prior \(N(y \mid 0, \tau_2 I)\), regularised least-squares is obtained:

$$\mathcal{E} = \frac{1}{2} \| x - Ay \|_2^2 + \frac{\lambda}{2} \| y \|_2^2 + \text{const},$$ \hfill (1.11)

with \(\lambda = \tau_2 / \tau_1\).
The Expectation Maximisation Algorithm

Latent (or hidden) variables can be thought of as modelling the underlying process that generated the observed data. Introducing latent variables can simplify calculations. This is because a complicated distribution over the observed data can be broken down into simpler components. The original distribution over the observed data can be obtained by marginalising out the latent variables \( z \) from the joint distribution \( p(x) = \int p(x, z) dz \). However, the number of latent variables tend to scale with the amount of observed data, which can be prohibitive in certain cases. For example, in a Gaussian mixture model (GMM) the dimensionality of the hidden space is the number of classes times the dimensions of the observed data.

A general method that can be applied for estimating parameters/unknowns \( (Y) \) of a distribution when the observed data depends on latent variables \( (Z) \), is the expectation maximisation (EM) algorithm \( (\text{Dempster et al., 1977}) \). In the EM algorithm, the model log-evidence (\( \ln p(X, Y) \)) is lower bounded by an expectation:

\[
\mathbb{E}_{p(Z | X, Y)} [\ln p(X, Y, Z)] = \int p(Z | X, Y) \ln p(X, Y, Z) dZ.
\]  

(1.12)

By choosing initial values for the unknowns \( Y^{(0)} \), EM iteratively maximises the log-evidence by:

1. **E-step**: Evaluate \( p(Z | X, Y^{(i)}) \)

2. **M-step**: Evaluate \( Y^{(i+1)} \) by

\[
Y^{(i+1)} = \arg\max_Y \left\{ \int p(Z | X, Y^{(i)}) \ln p(X, Y, Z^{(i)}) dZ \right\}.
\]  

(1.13)

Convergence is usually assumed once the change in the monotonically increasing lower bound is below some defined threshold.

Variational Bayes

VB inference involves estimating factorised approximations to a probability distribution \( (\text{Beal, 2003; Bishop, 2006; Arridge et al., 2018}) \). Letting \( Z \) encapsulate both model parameters and latent variables (as VB assumes the model parameters to be
random variables), a variational distribution $q(Z)$ is introduced. This distribution is an approximation to the true posterior, so that:

$$q(Z) \approx p(Z \mid X). \quad (1.14)$$

The variational distribution is restricted to belong to a family of distributions, such that a locally-optimal, analytical solution can be obtained and is in practice usually assumed to factorise over some partition of the latent variables:

$$q(Z) = \prod_{i=1}^{I} q_i(z_i), \quad (1.15)$$

where $Z = \{z_1, \ldots, z_I\}$ is a set of disjoint groups of variables. This factorized form of variational inference is known as mean-field approximation.\(^9\)

A dissimilarity function is used to measure how different the variational posterior is from the true posterior. Most commonly, the Kullback-Leibler (KL) divergence is used:

$$\text{KL}(q \parallel p) = \int q(Z) \ln \left( \frac{q(Z)}{p(Z \mid X)} \right) dZ. \quad (1.16)$$

The KL divergence between the true and the variational posterior can be reformulated as the log-evidence:

$$\ln p(X) = \text{KL}(q \parallel p) - \int q(Z) \ln \left( \frac{q(Z)}{p(X, Z)} \right) dZ$$

$$= \text{KL}(q \parallel p) + \mathcal{L}(q). \quad (1.17)$$

The lower bound $\mathcal{L}(q)$ is known as the evidence lower bound (ELBO) or negative variational free energy:

$$\mathcal{L} = \int q(Z) \ln \left( \frac{p(X, Z)}{q(Z)} \right) dZ \quad (1.18)$$

As the log evidence in (1.17) does not depend on $q$, maximising the negative variational free energy with respect to $q$ minimises the KL divergence of $p(Z \mid X)$ from $q(Z)$. Maximising the similarity between a variational and a true posterior is the main idea behind VB techniques.

\(^9\)A mean-field method can be conceptualised as the study of a large and complex stochastic model by considering its simpler form.
By variational optimisation of the lower bound in (1.18), with respect to all of the distributions \( q_i(z_i) \) in (1.15), a general expression for the optimal solution can be derived (Bishop, 2006):

\[
\ln q_j^*(z_j \mid X) = \mathbb{E}_{i \neq j}[\ln p(X, Z)] + \text{const.}
\] (1.19)

The expression in (1.19) provides the basis for applications of VB, saying that the log of the optimal solution for factor \( q_j \) is obtained simply by considering the log of the joint distribution and then taking the expectation with respect to all of the other factors. This technique is known as variational expectation-maximisation (VEM).

Finally, note that standard EM is a special case of VB when \( q(Z) \) in the ELBO in (1.18) is replaced with \( p(Z \mid X) \):

\[
\mathcal{L} = - \int q(Z) \ln \left( \frac{q(Z)}{p(X, Z)} \right) dZ
= \int p(Z \mid X, Y) \ln \left( \frac{p(X, Z)}{p(Z \mid X, Y)} \right) dZ
= \int p(Z \mid X, Y) \ln p(X, Z) dZ + \text{const.}
\] (1.20)

### 1.5.3 Discriminative vs Generative Models

Bayes’ rule allows a model to be categorised as either discriminative or generative (Bishop, 2006). In a discriminative approach, a parametric model for the posterior probabilities is introduced. The values of the parameters are inferred from a collection of labelled training data \( Y = \{y_n\}_{n=1}^N \) and \( X = \{x_n\}_{n=1}^N \). For a new observation \( x^* \), the aim is then to infer the value of its corresponding \( y \) value by:

\[
\hat{y} = \arg\max_y p(y \mid x^*, X, Y).
\] (1.21)

As can be seen, discriminative techniques directly maximise the model posterior by learning a mapping from training data, and therefore do not include a model of how the data was generated – the right-hand side of Bayes’ rule is not explicitly defined. As discriminative models are given both input observations and desired outputs, learning tasks based on such methods are categorised as supervised. Supervised learning is the machine learning task of inferring a function from labelled...
training data. Two examples of supervised learning are: (1) classification, in which the desired outputs are discrete class labels and the goal is to classify new inputs correctly; and (2) regression, in which the desired outputs are continuous and the goal is to predict the output accurately for new inputs. Many deep learning methods based on neural networks are examples of discriminative models.

Generative approaches, on the other hand, model the joint distribution $p(x, y)$ of observables and unknowns. The joint distribution can be obtained by learning the class prior probabilities $p(y)$ and the likelihoods $p(x \mid y)$, separately. The required posterior probabilities are then obtained using Bayes rule as:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}.$$  \hspace{1cm} (1.22)

As generative models can draw inferences from datasets consisting of input data without labelled responses, they can naturally be used for both semi-supervised and unsupervised learning tasks. The goal of unsupervised learning is fundamentally that of estimating the density which is likely to have generated the observed data. This density can be used for reasoning, decision making, prediction, and more. Examples of unsupervised learning tasks are: clustering, dimensionality reduction and modelling data densities. Semi-supervised learning is halfway between supervised and unsupervised learning. In addition to unlabelled data, the algorithm is in this case also provided with some supervision information, but not necessarily for all observations (in which case it would be fully supervised).

Compared to discriminative approaches, generative models typically have the following advantages:

- They explicitly model the data acquisition process, that is, how the data was generated. This approach is desirable if one wishes to build information about the generation process into the model \cite{Fischl2004, Ashburner2005Friston2005}.

- They do not require labelled training data, but can be improved with some form of labelling. Generative models can therefore be very valuable in situations where producing labelled data is costly, \textit{e.g.}, in medical imaging.
They can handle missing data by marginalising out unknown variables (Ghahramani and Jordan, 1994; Iglesias et al., 2015).

Disadvantages of the generative approach are that calculating the denominator of Bayes’ rule quickly becomes intractable, and that the full joint probability may model many aspects of the data that may be irrelevant for determining the posterior probabilities (Ulusoy and Bishop, 2005).

By contrast, discriminative models generally offer the following advantages:

- No assumptions about how the data was generated are needed, the posterior distribution is instead directly learned from the data (Zhang et al., 2015).

- Inference is usually simpler in a discriminative model. For example, in many neural network based models, no derivatives need to be derived as automatic differentiation is available in most software packages implementing these models (Baydin et al., 2018).

- The flexibility of the model is used in regions of input space where the posterior probabilities differ the most. Hence, if some of the key benefits of using a generative model are not required (for example, the ability to handle missing data), it would be expected that discriminative models would have better predictive performance since they are trained to directly predict the posterior distribution, e.g., a class label rather than an entire joint distribution.

Disadvantages of the discriminative approach is that it requires a specific form of training data, and may need to see a great deal of this data in order to work effectively. Furthermore, it can be unclear what amount and distribution of data is sufficient for discriminative methods to generalise.

### 1.6 Publicly Available Neuroimaging Data

Publicly available medical imaging data allows researchers to quickly begin testing hypotheses and conduct experiments, while sidestepping concerns regarding pri-
Figure 1.10: Example images from the five publicly available neuroimaging datasets used in this thesis. From top to bottom: BrainWeb, IXI, MICCAI2012, MRBrainS18 and RIRE.
1.6. Publicly Available Neuroimaging Data

vacy and data access. It also allows for replicating results and to more fairly compare methods. Throughout this thesis publicly available neuroimaging data will be used for both model training and evaluation. The next sections detail these datasets and Table 1.2 compares their strengths and weaknesses. Figure 1.10 shows example images from each dataset.

BrainWeb

The BrainWeb database contains a set of realistic MRI data volumes produced by an MRI simulator (Cocosco et al., 1997). This data is meant to be used by researchers to evaluate the performance of various image analysis methods in a setting where the ground-truth is known, e.g., for image registration and denoising. BrainWeb contains simulated brain MRI data based on two anatomical models: normal and multiple sclerosis. For both of these, full 3D data volumes have been simulated using three sequences (T1w, T2w and PDw), and a variety of slice thicknesses, noise levels, and levels of intensity non-uniformity. Furthermore, researchers can use the BrainWeb database to simulate their own ‘degraded’ versions with available ground-truth.

Available from: [https://brainweb.bic.mni.mcgill.ca/brainweb/](https://brainweb.bic.mni.mcgill.ca/brainweb/)

IXI

The IXI dataset contains nearly 600 MR images from normal, healthy subjects. The MR image acquisition protocol for each subject includes: T1w, T2w and PDw images, magnetic resonance angiography (MRA) images and DWI. The data was collected at three different hospitals in London, on 1.5T and 3T systems. The structural scans all have close to 1 mm isotropic voxels. The IXI dataset does not provide manual structural, or tissue, labels suitable for semi- or fully-supervised learning. This limits the range of modelling applications that can be trained to unsupervised approaches only.

MICCAI2012

The MICCAI2012 dataset was provided for use in the MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labelling (Landman and Warfield, 2012). The dataset contains 30 manually labelled MRI brain scans with 25 unique subjects (five subjects were scanned twice). The MR images were taken from the OASIS dataset (http://www.oasis-brains.org/). The MRIs are all T1w with 1 mm isotropic voxels. The age range of the subjects is from 18 to 96 years (mean: 34, median: 25).

The scans were manually segmented into 136 anatomical brain regions by Neumorphometrics Inc. These regions can be combined to form, e.g., a white or gray matter class. Note that the extracerebral cerebrospinal fluid (CSF) was not manually segmented.

Available from: my.vanderbilt.edu/masi/workshops

MRBrainS18

The MRBrainS18 dataset was provided for use in the MICCAI 2018 Grand Challenge on MR Brain Segmentation. The purpose of the challenge was to compare methods for segmentation of GM, WM, CSF, and other structures on MRI scans of the brain, and to assess the effect of (large) pathologies on segmentation and volumetry. The image data used in the challenge was acquired from 30 subjects on a 3T scanner at the UMC Utrecht (Netherlands). For each subject, fully annotated multi-sequence scans were acquired: T1w, T1w inversion recovery (T1-IR) and T2-FLAIR. The 30 subjects (age > 50) included patients with diabetes, dementia and Alzheimer’s disease, as well as matched controls. All subjects’ images could contain varying degrees of atrophy and white matter lesions.

All subjects’ images were manually segmented by the same neuroanatomist into 11 classes: background, cortical GM, basal ganglia, WM, WM lesions, CSF in the extracerebral space, ventricles, cerebellum, brain stem, infarction and other. The MR images were provided both in ‘raw’ format (non-preprocessed NIfTIs) or preprocessed (by registration and bias field correction). The raw format MR volumes have voxel sizes: [1, 1, 1] mm for the T1w image and [0.96, 0.96, 3]
1.7 Thesis Overview

The next three chapters contain the main contributions of this thesis, which all focus on the development of probabilistic models for improved segmentation and registration of routine clinical data. The organisation of these chapters follows next, with related publications, as well as links to implementations.
Table 1.2: Pros and cons of publicly available neuroimaging datasets used in this thesis.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>BrainWeb</td>
<td>Simulated data enables ground-truth comparisons and controlled experiments</td>
<td>Simulated data is possibly a poor substitute for real data</td>
</tr>
<tr>
<td>IXI</td>
<td>Close to 600 subjects • Small voxel sizes • Multiple contrasts (T1w, T2w, PD2 and MRA) • Multiple scanners • Neck coverage</td>
<td>No manual labels makes supervised learning difficult</td>
</tr>
<tr>
<td>MICCAI2012</td>
<td>30 subjects • 136 manual labels • Small voxel sizes</td>
<td>Labels do not cover whole brain (no extracerebral CSF) • Single contrast (T1w)</td>
</tr>
<tr>
<td>MRBrainS18</td>
<td>11 manual labels • Labels cover whole brain • Multiple contrast (T1w, FLAIR and IR) • Neck coverage</td>
<td>7 subjects • Cerebellum is one class (not separated for GM and WM) • Extracerebral CSF labels contain sinuses</td>
</tr>
<tr>
<td>RIRE</td>
<td>Multiple modalities (MRI, CT and PET)</td>
<td>Purposefully misaligned without ground-truth transformation • Thick-sliced</td>
</tr>
</tbody>
</table>

Chapter 2: Resolution Recovery in Clinical Neuroimaging Data

This chapter presents a tool for resolution recovery in routine clinical neuroimaging data. As has been described, such images exhibit great variability, both biological and instrumental. This variability makes automated processing with neuroimaging analysis software very challenging. This leaves intelligence extractable only from large-scale analyses of clinical data untapped, and impedes the introduction of automated predictive systems in clinical care. The tool presented in this chapter enables such processing, via inference in a probabilistic forward model. All model parameters are estimated from the observed data, without the need for manual tuning. The model-driven nature of the approach means that no type of training is needed for applicability to the diversity present in clinical images. It is shown on simulated data that the proposed approach outperforms conventional model-based techniques; and, on large hospital datasets, that the tool can successfully super-resolve very thick-sliced MR images, and denoise CT scans. Furthermore, the method is also used to improve the signal-to-noise ratio (SNR) of multi-echo MR volumes, acquired in a qMRI context.

Related Publication(s)
Chapter 3: A Robust Cost Function for Multimodal Image Registration

Automated registration algorithms, bringing a pair (or group) of images into alignment, usually do so by optimising some cost function. Information theoretic cost functions (such as mutual information) have been shown successful at this task. However, as they usually rely on joint image intensity histograms, whose size grows exponentially with the number of images, they become prohibitive for simultaneously aligning more than two images. In addition, cost functions based on intensity histograms can fail in the presence of very strong bias fields or large offsets. This chapter introduces a novel, more robust alternative for group-wise intrasubject, intermodality alignment. Its use is justified theoretically in the context of image registration, where it has the advantage of enabling simultaneous alignment of multiple (> 2) images, while being insensitive to strong bias fields and large misalignments. The approach is validated on both simulated and real data and is shown to perform on pair with methods included in the SPM12 software.

Related Publication(s)

• Brudfors, M., Balbastre, Y., & Ashburner, J. Multi-channel total variation as a cost function for group-wise, intermodal image registration. [In preparation]
Implementation

https://github.com/WCHN/MTV-Registration

Chapter 4: Improved Segmentation of Clinical Neuroimaging Data

This chapter proposes extensions to a widely used segmentation model, to make it more applicable to clinical-grade neuroimaging data. The original method uses a generative model, with a GMM likelihood term and a non-stationary, atlas-based prior term. The extensions in this chapter modifies these terms in two novel ways. The first method proposes a solution to the problem of variable fields of view (FOV) across multimodal scans of the same patient, and can furthermore be used to simulate entirely missing modalities (e.g., MRI to CT). In short, this is achieved by allowing the GMM component to handle missing voxels. The second method marries the generative model image segmentation with deep learning, which improves segmentation accuracy on images acquired at different centres and scanners. The connection between the two methods is in the prior term, which encodes a convolutional neural network via a Markov random field (MRF).

Related Publication(s)


1.7. Thesis Overview


Implementation

https://github.com/WCHN/segmentation-model
Chapter 2

Resolution Recovery in Clinical Neuroimaging Data

Contents

2.1 Introduction ............................................... 58
2.2 Super-Resolving Multimodal Clinical MRI ............... 58
2.3 Related Work ............................................... 61
2.4 Methods .................................................. 64
  2.4.1 The Generative Model ................................. 65
  2.4.2 Model Optimisation .................................. 72
  2.4.3 Denoising ............................................. 78
  2.4.4 Implementation Details ............................. 78
2.5 Evaluation .................................................. 79
  2.5.1 Super-Resolution ..................................... 80
  2.5.2 Denoising ............................................. 90
2.6 Discussion .................................................. 97
Chapter 2. Resolution Recovery in Clinical Neuroimaging Data

2.1 Introduction
This chapter tackles the problem of recovering resolution in clinical neuroimaging data. The main objective is to facilitate the processing of thick-sliced MR scans with segmentation and registration methods, using a super-resolution technique; but it also explores denoising CT scans and improving the SNR in multi-parameter qMRI maps.

The proposed model enables both super-resolution and denoising via inference in a generative model. The model relies on a very limited set of assumptions, makes optimal use of orthogonal acquisitions, and does not necessitate any parameter tuning. This data-free approach limits the risk of over-fitting, allowing very large clinical datasets to be processed without the need for much quality control. Importantly, it is demonstrated on a large hospital dataset that exploratory machine-learning tasks, such as age and sex prediction, obtain improved results when applied to images that are processed with the tool, compared with standard baseline methods.

Furthermore, as the task of image denoising is a special case of the proposed generative model, experiments are performed to validate the model in this setting. Good results on denoising clinical CT images are demonstrated – and a proof of concept idea for improving the SNR in qMRI maps is also shown. Next, the most general formulation of the model is introduced, for super-resolving multimodal MR images.

2.2 Super-Resolving Multimodal Clinical MRI
Commonly, for analyses that require isotropic voxels, thick-sliced MRI volumes are upsampled using simple interpolation, even though it can introduce artefacts, such as aliasing (Aganj et al., 2012), and hence bias subsequent analyses. Similarly, interpolation techniques are often used to reslice multimodal images (e.g., T1w and T2w MRIs) to the same lattice, which is a necessary step prior to multi-channel analysis. Later in this chapter it will be shown that this method is non-optimal, as it
does not use the fact that information is shared between modalities. The objective of the model presented here is, instead, to take advantage of information that is distributed across all MR scans of a patient. This is in order to generate closer-to-research quality images from clinical scans, which could enable large-scale studies of hospital-grade MR data.

The approach is based on a principled generative model that can reconstruct multimodal high-resolution (HR) images with isotropic voxels from multimodal, low-resolution (LR) clinical scans. Carefully crafted generative models have already been shown to generalise well in the case of medical image segmentation (Ashburner and Friston, 2005; Zhang et al., 2001; Fischl et al., 2004), and this chapter develops such a model for image quality improvement of clinical brain scans. The assumption is that the LR MR images all have orthogonal thick-slice directions (see Figure 2.1); an assumption that almost always is the case for clinical MR data. The multi-sequence capability of the proposed model stems from a cross-channel functional that has been studied extensively in the computer vision community, and which has also been used in MRI, for parallel image reconstruction (Chen et al., 2013; Chatnuntawech et al., 2016).

The model-based approach proposed in this chapter is data agnostic and does not require any complex learning. Learning a representative distribution of the large number of sequences used in clinical MR imaging is extremely challenging and could make the model much less applicable to hospital datasets. Having to train a model anew, each time a new combination of MR contrasts is encountered, or a different scaling factor is needed, would make a general tool much more difficult to develop. Furthermore, the model can easily be used for single- and multimodal denoising by a simple change to some of its parameters.

It is shown on a large number of patient scans that the model is capable of super-resolving very thick-sliced clinical MRIs (≈6.5 mm), with variable FOV over the images. Furthermore, denoising results on patient CT data are shown, as

1Here, the different MR contrasts acquired in a patient are considered as being equivalent to colour channels in computer vision (e.g., RGB).
Figure 2.1: Example of clinical MR images of a patient. Patient MRI scans commonly have thick slices (6.5 mm), with different slice-select directions, and partial brain coverage. The model proposed in this chapter combines information in the different images in order to reconstruct isotropic HR versions.

well as a novel method for improving the SNR of qMRI maps. To ensure good generalisability, all model parameters were estimated in a principled way. The implementation is publicly available, and can be used by researchers interested in analysing routine clinical neuroimaging data. The software is available from [https://github.com/brudfors/spm_superres](https://github.com/brudfors/spm_superres) and will shortly be incorporated into the SPM12 software.
2.3. Related Work

In image processing, the task of improving the resolution of an image after it has been acquired is known as super-resolution\(^3\). There are two cardinal ways of super-resolving: the first is to combine information from a multiplicity of images of the same subject, the second is to introduce guidance from a population of other subjects. For both these methods it is common to use some kind of inductive bias given mathematically or conceptually, \textit{e.g.}, regularisation. An example of super-

\(^3\)www.fil.ion.ucl.ac.uk/spm/software/spm12

\(^3\)The term super-resolution can also be used to describe techniques attempting to transcend the diffraction limit of an optical system, \textit{e.g.}, super-resolution microscopy.

\textbf{Figure 2.2}: An example of super-resolution to combine multiple LR images into a HR version. Here, a forward model is assumed have generated three LR images from an unknown HR image. These three LR images are combined using super-resolution to obtain an estimate of the unknown HR version. Compared to simply linearly interpolating one of the LR images, super-resolution produces visually a higher quality HR image.
resolution, combining multiple LR images, is shown in Figure 2.2.

Super-resolution was first studied in the computer vision field in the early 80s and has since grown to become an active area of research using tools from, e.g., inverse problems and machine learning (Huang and Yang, 2010; Yue et al., 2016). In the early 2000s, super-resolution found applications in the medical imaging community, in particular for brain MRI (Van Reeth et al., 2012). This was partly due to multi-LR super-resolution’s dependency on exact image alignment, where for brain imaging it is possible to obtain good alignment simply by rigid registration. It was furthermore shown that super-resolution improved resolution and SNR favourably, compared with direct HR acquisition (Plenge et al., 2012).

The earliest works investigating super-resolution applied to brain MR images, defined an observation model that did not take into account multiple MR contrasts, but rather specified how a set of LR images, of the same contrast, were generated from a HR image (Peled and Yeshurun, 2001; Greenspan et al., 2001). The basic idea was that it should be more beneficial (with respect to the trade-off between acquisition time and SNR ratio) to acquire multiple LR images, and then combine them using super-resolution, than to acquire one HR image. The observation model was constructed by linear operations (shift, down-sampling, blurring) and additive Gaussian noise. A ML estimate of the unknown HR image could then be found by a gradient descent style algorithm. Limited in handling scans only of the same orientations, but with subpixel offsets, Shilling et al. (2009) extended the methods based on observation models to handle multiple LR images, of different slice-select directions. The work of Poot et al. (2010) subsequently generalised this approach to images not necessarily rotated around a common frequency encoding axis, allowing for any slice orientation. They additionally solved the optimisation problem efficiently using a conjugate gradient algorithm.

The ML methods were quickly extended to include some form of regularisation via MAP estimation, to reduce the ill-posedness of the solution and incorporating prior knowledge (e.g., neighbouring voxels should look similar). The work of Peeters et al. (2004), Kainz et al. (2015) and Rousseau et al. (2010b) did so using
edge-preserving functionals, such as the Huber function or anisotropic diffusion, to regularise their solutions, while Zhang et al. (2008) and Poot et al. (2010) used the Euclidean norm of the image gradients. In Bai et al. (2004) a MRF was used, and in Shi et al. (2015) two sorts of regularisation were combined: the edge-preserving total variation (TV) and a low-rank term that enabled utilisation of information throughout the image. Also Tourbier et al. (2015) used TV as a regulariser, with an adaptive regularisation scheme. Besides including some sort of regularisation in the objective function, replacing its mean-squared error norm has also been explored, with the aim of increased robustness to incorrect noise models (Gholipour et al. 2010).

The methods discussed up until this point are non-optimal for processing clinical MRIs, as they assume multiple LR images of the same contrast and hence do not use the fact that routine clinical scans often are of different contrasts. A method utilising information from a HR reference image of a different contrast was first proposed by Rousseau et al. (2010a). The super-resolved image was obtained by iteratively denoising an estimate of the super-resolved image with a patch-based technique, using the HR image as a reference, and then solving an optimisation problem, where the denoised image regularised the solution. The patch-based work of Manjón et al. (2010b) used the information from a HR reference in a similar way, but instead of solving for the super-resolved image in an optimisation setting, they used an iterative reconstruction-correction scheme. Another patch-based approach estimated weights from a HR image of a given contrast and then regressed a HR image from a LR image with a different contrast (Zheng et al. 2017). Numerous other works on super-resolving MRIs have used these patch-based approaches, utilising the pattern redundancy present in image patches (Manjón et al. 2010a; Coupé et al. 2013; Plenge et al. 2013; Jog et al. 2014). However, although these methods enable super-resolving across MR channels, they require access to HR data. Such data is seldom available in a clinical setting. To mitigate this problem, methods have been developed that use the property that clinical MR images are inherently anisotropic to learn a regression between LR and HR images (Jog et al. 2016; Zhao...
Data-driven (or learning-based) methods have also been thoroughly investigated for the task of super-resolving brain MRIs. These methods aim at learning a LR to HR mapping from training data. One of the first such methods was proposed by Rueda et al. (2013), using sparse dictionary learning. Regression-based techniques have also been explored, such as patch-based random forest regression (Alexander et al., 2014; Sindel et al., 2018). Convolutional neural networks (CNNs) form another class of supervised method, which has gained interest in MRI super-resolution (Tanno et al., 2017; Cengiz et al., 2017; Pham et al., 2017). Deep generative models have also been used to directly learn a LR to HR mapping for super-resolution recovery (Chen et al., 2018). To combat the fact that HR training data is difficult to obtain, Dalca et al. (2018) proposed a generative model for sparse image patches, where LR clinical scans were used as training data. They additionally dealt with missing data in an elegant way. However, there is to the best of the authors knowledge not yet a learning-based super-resolution tool useable without the need for training anew, when faced with unseen contrasts.

2.4 Methods

Compared with some of the super-resolution approaches mentioned in the previous section, the model proposed here is multimodal, and does not require any complex learning, nor any HR reference data. It additionally avoids any learning from training data with the hope of improved generalisability when applied to diverse clinical MR scans. The model relies on the definition of a joint distribution of a patient’s multimodal MR images. A generative model that integrates multimodal images requires a component that relates signal across the various modalities. Here, this component is a novel prior in the context of MRI super-resolution, which promotes combining image information distributed across a patient’s MR images. In this section, each MR contrast (e.g., T1w, T2w, PDw) is considered constituting one channel of a multi-channel volume. When multiple images of the same contrast are acquired, they are considered as multiple noisy observations of the same channel.
2.4. Methods

Figure 2.3: Graphical model of the joint probability distribution in (2.1). Random variables are in circles, observed variables are shaded, plates indicate replication. There are $C$ unknown HR images ($y_c$) and $I_c$ observed LR images ($x_{ci}$) for each $c$. Model parameters are dotted, they are: the Gaussian noise precisions $\tau_{ci}$, projection matrices $A_{ci}$, and regularisation parameters $\lambda_c$.

2.4.1 The Generative Model

The model assumes that each LR image is generated by selecting thick slices, arbitrarily rotated and/or translated, from a HR image, with the addition of random noise. This assumption can be written as a conditional probability distribution, the data likelihood. It is also assumed that each HR image is the result of a random process, characterised by another probability distribution, the prior. This generative model is formalised by the joint probability distribution:

$$p(X, Y) = p(X | Y) p(Y) = \prod_{c=1}^{C} \prod_{i=1}^{I_c} p(x_{ci} | y_c) p(Y),$$  \hspace{1cm} (2.1)

where $Y = \{y_c\}_{c=1}^{C}$ denotes the unknown HR images of $C$ different channels, and $X = \{X_c\}_{c=1}^{C}$ a set of LR images. The variable $I_c$ is the number of observed LR images of channel $c$, $N$ is the number of voxels in the HR images, and $N_{ci}$ is the number of voxels in the $i$th LR image of the $c$th channel. Note that the model allows for missing data in the observed LR images (i.e., Not a Number), which enables these voxels to be filled in during model fitting.

Prior to super-resolving a set of MR images a rigid registration of the observed data to a common reference is performed, using the spm_coreg routine of SPM12.

Estimating the HR images ($\hat{Y}$), given a set of observed LR images ($X$), is cast

---

4 A more elegant way of doing this would be to include registration in the generative model, such that fitting the model would optimise also some registration parameters.
as MAP estimation in the joint probability distribution defined by (2.1):

$$p(Y | X) = \frac{p(X | Y) p(Y)}{p(X)}$$

$$\Rightarrow \hat{Y} = \arg\min_Y \{-\ln p(Y | X)\}$$

$$\Rightarrow \hat{Y} = \arg\min_Y \{-\ln p(X | Y) p(Y)\}.$$  \hspace{1cm} (2.2)

For the model to generalise, its parameters are estimated from either the observed data (likelihood and prior hyper-parameters) or set in a general and consistent way (projection matrices).

A graphical representation of the generative model is shown in Figure 2.3. As in practice, there is rarely more than one observation of the same channel, the summations over \(i\) will from now on be dropped, so that only one observation of each channel is assumed. All derivations stay the same, except for additional summations over conditional distributions of LR images. The individual components of the generative model will be further explained in the next three sections.

Model Likelihood

The likelihood function should describe the data generating process of LR images \((x_c)\) from unknown HR images. Its main component is a deterministic projection matrix \((A_c)\) that encodes the slice-selection parameters (orientation, thickness, gap, profile) of an LR image. It is a linear operator that, when applied to the corresponding HR image, creates a noiseless LR version. The second component of the generative process encodes acquisition noise. As MR images are usually reconstructed as the magnitude of an image that was originally complex – and Gaussian noise on complex data gives a Rice distribution in the magnitude image – a Rician noise model would be suitable\(^5\) (Aja-Fernández and Tristán-Vega, 2013). However, it has been shown that the mathematically more tractable Gaussian distribution closely approximates the true Rician noise distribution in MRI (Gudbjartsson and Patz, 1995).

\(^5\)Note that multi-coil MR images generally do not have Rician noise because of the way the images are reconstructed. The Rician assumption is therefore good for older MR scanners, but becomes a bit of an approximation for more modern systems.
The following forward model is therefore assumed:
\[ x_c = A_c y_c + \epsilon, \quad \epsilon \sim \mathcal{N} \left( 0, \tau^{-1}_c \right), \] (2.3)
so that the conditional distribution of an observed LR image, given an unknown HR image, is a multivariate Gaussian distribution:
\[ p(x_c | y_c) = \mathcal{N} \left( x_c | A_c y_c, \tau^{-1}_c I \right) = \frac{\tau_c^{N_c/2}}{(2\pi)^{N_c/2}} \exp \left( -\frac{\tau_c}{2} \| x_c - A_c y_c \|_2^2 \right), \] (2.4)
where \( \tau_c \) is the precision of the noise of LR image \( c \) and \( A_c \in \mathbb{R}^{N_c \times N} \) is the linear operator mapping from HR to LR space. Dropping all terms that do not depend on \( y_c \), the negative log-likelihood can be written as:
\[ -\ln p(x_c | y_c) = \frac{\tau_c}{2} \| x_c - A_c y_c \|_2^2 + \text{const}. \] (2.5)
The multivariate Gaussian distribution in (2.4) is a likelihood function that is already well-established in the super-resolution literature (Greenspan et al., 2001; Shilling et al., 2009; Poot et al., 2010).

Model Prior
The prior probability should encode one’s belief about the unknown HR images \( (Y) \). Many types of priors have been devised for image reconstruction problems. The most popular alternative is perhaps the Tikhonov (or \( \ell_2 \)) prior, that penalises the squared norm of some image features:
\[ p(y_c) = \mathcal{N} \left( y_c | 0, (\lambda_c L)^{-1} \right), \] (2.6)
where \( \lambda_c \) is a channel specific regularisation parameter and the precision matrix \( (L) \) is designed as to induce correlations between image voxels. This type of prior probability favours images that are smooth when the precision matrix encodes some differential operator \( L = D^T D \), so that:
\[ -\ln p(y_c) = \frac{\lambda}{2} \| D y_c \|_2^2 + \text{const}. \] (2.7)
If the differential operator encodes a first-order derivative, then the negative log-likelihood of this prior is known as a first-order Tikhonov regulariser. The differential operator is in dimension \( D \in \mathbb{R}^{NG \times N} \), where \( G \) are the number of differential
Figure 2.4: Empirical investigation of the distribution of MR image gradients. By computing the $x$, $y$, and $z$-gradients of a MR image, and fitting a Gaussian (yellow) and a Laplace distribution (red) to the histogram of these gradients (blue), it can be seen that the Laplace distribution more accurately captures the empirical distribution of MR image gradients.

The voxel size of the reconstructed images is accounted for by dividing the gradient directions with the corresponding voxel size.

The underlying assumption of the Tikhonov prior is that the gradients in the HR image have a Gaussian distribution. However, by studying the empirical gradient distribution of a HR, noise-free MR image it can be seen that a Laplace distribution is a more suitable alternative (see Figure 2.4). Discarding terms that do not depend on $y_c$, this gives the following prior distribution:

$$p(y_c) \propto \prod_{n=1}^{N} \exp\left(-\lambda_c \|D_n y_c\|_2\right),$$  \hspace{1cm} (2.8) \\

$$-\ln p(y_c) = \lambda_c \sum_{n=1}^{N} \|D_n y_c\|_2 + \text{const.}$$  \hspace{1cm} (2.9)

where $\lambda_c$ is the inverse of the scale parameter of the Laplace distribution, and the operator $D_n \in \mathbb{R}^{G \times N}$ returns the gradients at the $n$th data point of $y_c$. The log of the Laplace distribution in (2.9) is known as (isotropic) TV and is another popular method for regularising image reconstruction problems (Rudin et al., 1992). Rather
than favouring smooth reconstructions, TV retains edges and therefore leads to less blurry results. Here, to avoid biasing the reconstruction, both the forward and backward first-order finite differences along each dimension are extracted, giving $G = 6$ in 3D.

If it is assumed that the unknown HR images follow the distribution in (2.8), then there is no dependency between channels. Clearly this assumption is false for MR images of the same subject, where most edges should be shared between scans. I therefore propose to use the MTV functional (Sapiro and Ringach, 1996; Bresson and Chan, 2008) as the prior probability:

$$p(Y) \propto \prod_{n=1}^{N} \exp \left( - \sum_{c=1}^{C} \| \lambda_{c} D_{n} y_{c} \|_{2}^{2} \right), \quad (2.10)$$

$$- \ln p(Y) = \sum_{n=1}^{N} \left( \sum_{c=1}^{C} \| \lambda_{c} D_{n} y_{c} \|_{2}^{2} \right) + \text{const.} \quad (2.11)$$

The summation over channels ($C$) inside of the square root ensures that the channels are ‘mixed’, making the assumption that the MR images have large smooth regions and a few sharp edges, in similar places (note that for $C = 1$, TV is a special case of MTV).

MTV is part of the vectorial total variation (VTV) class of regularisation functions, which extends TV to vector fields. Other forms of multi-channel TV methods include: color TV (Blomgren and Chan, 1998), where the authors pool individual TV contributions across channels; and total nuclear variation (Holt, 2014), where the author uses various matrix norms on the image Jacobian (e.g., the nuclear norm).

Here, I chose to use MTV because of: (1) its interpretation as a type of Laplace distribution, making it possible to estimate the $\lambda$ parameters in a principled way; and (2), the resulting optimisation problem being straightforward to solve.

Model Parameters

Defining the generative model gives a set of parameters: the projection matrices ($A$), noise precisions ($\tau$) and prior parameters ($\lambda$). These parameters are all image-specific and it is critical that these parameters are set in a principled way for the
model to generalise – it is not feasible to expect users to do manual tuning. This section gives more detail about how the model parameters were chosen, for MR data. The model can be used on other types of data, but the assumptions regarding the parameters may then have to be modified (e.g., see Section 2.5.2).

**Projection matrices:** The projection matrix ($A_c$) is a linear mapping from HR to LR image space. It should reproduce the slice-selection process of the MRI scanner. In this work, the projection matrix is constructed as three linear operators applied in succession as:

$$A_c = R_c S_c T_c. \tag{2.12}$$

The operator $T_c$ resamples from the HR image’s FOV to the LR image’s FOV, but keeping the voxel size of the HR image. If the slice orientation is at an angle with respect to the HR grid, the corresponding rotation is accounted for. Furthermore, to improve numerical properties of the projection operator, the FOV of the resampled image is slightly increased, which is then adjusted for in the $R_c$ operator.

The operator $S_c$ should simulate the slice profile of the MRI acquisition. The slice profile depends on the shape of the RF pulse applied during slice selection. This RF pulse can vary a lot from one sequence to the next. Therefore, there is not a single slice profile that suits all acquisitions (Liu et al., 2002).
Here, the slice profile is assumed to be Gaussian. Applying the $T_c$ operator before the Gaussian convolution ensures that the kernel is applied in the correct directions. The full width at half maximum (FWHM) of the Gaussian kernel is set to zero in the in-plane directions and to the slice-thickness in the thick-slice direction. The thick-slice direction FWHM is additionally modulated by subtracting a slice gap. I compute an estimate of this slice gap from 29,026 patients MRIs. The slice gap estimate is one third of the width of the slice thickness. This information could of course be read from the DICOM header of the image. But as the tool works on NIfTI data, this information may not be present, as it can be lost during DICOM-to-NIfTI conversion. The slice gap can however be provided, when available.

The operator $R_c$ performs the down-sampling operation from HR to LR space. The process of applying $A_c$ to a HR image is illustrated in Figure 2.5.

**Noise precision:** MR scans are usually reconstructed as the magnitude of an image that was originally complex. The Gaussian noise model in (2.3) is therefore just an approximation to a true Rician noise model. Hence, the amount of Rician noise in each observed LR image ($\tau_c$) needs to be estimated. This is done by fitting a mixture of two Rician distributions ($K = 2$) to the intensity histogram of each MR scan (Ashburner and Ridgway, 2013). The precision of the image noise ($\tau_c$) is then obtained from the class with the smallest non-centrality parameter, which should correspond to air. Two example fits are shown in Figure 2.6.

**Regularisation parameters:** Each unknown HR image has a corresponding regularisation parameter ($\lambda_c$); here, it is detailed how this parameter is chosen. First, note that that gradient values are, in general, correlated with intensity

---

6. The implementation of the projection matrix however, gives a user the option to easily change the slice profile assumption by simply changing the convolution kernel.

7. The slice gap can be computed from the DICOM representation of a MR image, whose header contain both the field Spacing Between Slices (0018, 0088) and Slice Thickness (0018, 0050), as the Spacing Between Slices minus the Slice Thickness.
values; the regularisation parameter should therefore be set with respect to
the mean intensity in a channel. A linear relationship is observed empirically,
between the standard deviation of gradient magnitudes $\sigma_c$ and the mean tissue
intensity $\mu_c$ of an image. The mean tissue intensities of 1,728 scans from the
IXI dataset (described in Section 1.6) are therefore computed by fitting the
Rician mixture model to each image and taking the non-centrality parameter
of its non-air (brain) component. The standard deviation of the first-order gra-
dients of each image is then regressed against this mean intensity. The value
$k_\lambda = \sigma / \mu = 4.67$ is obtained and the fit is shown in Figure 2.7. Since the
variance of a Laplace distribution relates to the scale parameter through the
relationship $\sigma^2 = 2 / \lambda^2$, the parameter $\lambda_c$ can then be set according to:

$$
\lambda_c = \frac{\sqrt{2}}{\sigma_c} = \frac{\sqrt{2}}{k_\lambda \mu_c},
$$

(2.13)

where, for a new image, the estimate of the mean brain intensity is obtained
from the Rician mixture fit (Figure 2.6).

### 2.4.2 Model Optimisation

With all the individual components of the model defined, the negative log posterior
probability can be written as:

$$
- \ln(p(X \mid Y) \cdot p(Y)) = \sum_{c=1}^{C} \frac{\tau_c}{2} \| x_c - A_c y_c \|^2_2 + \sum_{n=1}^{N} \sqrt{\sum_{c=1}^{C} \| \lambda_c D_n y_c \|^2_2} + \text{const},
$$

(2.14)

where the first term on the right-hand side is known as the data term and the second
term as the penalty term. The expression in (2.14) is what needs to be minimised to
obtain the $C$ super-resolved images:

$$
\hat{Y} = \arg\min_Y \left\{ - \ln p(X \mid Y) \cdot p(Y) \right\}.
$$

(2.15)

This optimisation problem is hard to solve because the MTV penalty term is non-
differentiable (nonsmooth). It is, however, convex with a global optimum.
Figure 2.6: Two examples of estimating the noise variance ($\tau_c$) and regularisation parameter ($\lambda_c$) from the MR image intensity histogram. A two-class Rician mixture model was fit to the intensity histogram of each MR image. The image noise was computed as the variance of the air class, and the mean brain intensity as the non-centrality parameter of the brain class. Note how the values on the x-axis show the non-quantitative nature of MR images – the two histograms cover different intensity ranges. Estimating the model parameters from each image’s histogram is what ‘normalises’ intensities across different scans.
Figure 2.7: Line fit for estimating the regularisation parameter ($\lambda_c$). A straight line was fitted between the mean tissue intensities and the standard deviations of image gradients, computed from the IXI dataset; its coefficient gives the value of the parameter $k_\lambda$.

A change of variables allows the unconstrained minimisation problem in (2.15) to be rewritten as a constrained minimisation:

$$\min_Y \sum_{c=1}^C \frac{\tau_c}{2} \|x_c - A_c y_c\|_2^2 + \sum_{n=1}^N \sqrt{\sum_{c=1}^C \|z_{nc}\|_2^2}$$

s.t. \(\lambda_c D_n y_c = z_{nc}\), for all \(n\) and \(c\),

(2.16)

where the smooth data term and the nonsmooth penalty term have been decoupled. The constrained form of (2.15) can now be solved efficiently by an algorithm known as alternating direction method of multipliers (ADMM).

Alternating Direction Method of Multipliers

ADMM is part of a class of optimisation methods called proximal algorithms [Boyd et al. 2011]. In short, an augmented Lagrangian is formulated from a constrained minimisation problem. This Lagrangian is then minimised in an alternating fashion until a convergence criterion is met. Many methods exist for solving TV problems. Here, an ADMM algorithm is chosen because such algorithms are straightforward
2.4. Methods

to implement and well suited to the type of optimisation problem in (2.15).

Deriving the general form of the ADMM algorithm starts with an unconstrained minimisation problem:

$$\hat{y} = \operatorname{argmin}_y \left\{ \text{data}(y) + \text{penalty}(y) \right\}, \quad (2.17)$$

which is equivalent to the constrained problem:

$$\min_{y,z} \text{data}(y) + \text{penalty}(z) \quad \text{s.t.} \quad F_d y + F_p z = d, \quad (2.18)$$

with constraints defined by $F_d$, $F_p$ and $d$. Note how the optimisation problem in (2.18) has the same form as the super-resolution minimisation in (2.16).

From (2.18), the augmented Lagrangian can be formulated as:

$$L_\rho(y, z, w) = \text{data}(y) + \text{penalty}(z) + w^T (F_d y + F_p z - d) + \frac{\rho}{2} \| F_d y + F_p z - d \|^2_2; \quad (2.19)$$

where $w \in \mathbb{R}^P$ holds the Lagrange multipliers and $\rho > 0$ is a descent step-size. From the augmented Lagrangian in (2.19), the ADMM updates are given by:

$$\hat{y} = \operatorname{argmin}_y L_\rho(y, z, w), \quad (2.20)$$

$$\hat{z} = \operatorname{argmin}_z L_\rho(y, z, w), \quad (2.21)$$

$$\hat{w} = w + \rho \cdot (F_d y + F_p z - d), \quad (2.22)$$

where (2.20) and (2.21) are known as the proximal operators at parameter $\rho$ for the data and penalty term, respectively. These ADMM updates are usually iterated until some convergence criteria is fulfilled.

**ADMM Updates**

In this section, the ADMM updates for the super-resolution model are derived. Two tensors will be used: $Z$ and $W$, which are both of dimensions $N \times C \times G$. Column vectors are then extracted from these tensors, where the subscript of a vector indicates its length: $z_n = \text{vec}(Z[n,:,:])$, $z_c = \text{vec}(Z[:,c,:])$, $z_{nc} = \text{vec}(Z[n,c,:])$, $w_n = \text{vec}(W[n,:,:])$ and $w_c = \text{vec}(W[:,c,:])$. 


With the notations introduced the augmented Lagrangian in (2.19) is formulated from (2.16) as:

\[
L_\rho(Y, Z, W) = \sum_{c=1}^{C} \frac{\tau_c}{2} \| x_c - A_c y_c \|_2^2 + \sum_{n=1}^{N} \| z_n \|_2^2 \\
+ \sum_{c=1}^{C} w_c^T (\lambda_c D y_c - z_c) + \frac{\rho}{2} \sum_{c=1}^{C} \| \lambda_c D y_c - z_c \|_2^2, \tag{2.23}
\]

where, for a fixed \( n \), the following relationship has been used:

\[
\sqrt{\sum_{c} \| z_{nc} \|_2^2} = \sqrt{\sum_{c} \sum_{g} (z_{ncg})^2} = \| z_n \|_2. \tag{2.24}
\]

From the augmented Lagrangian in (2.23) the ADMM updates can be derived as:

\[
\hat{y}_c = \arg\min_{y_c} \left\{ \frac{\tau_c}{2} \| x_c - A_c y_c \|_2^2 + \lambda_c w_c^T D y_c + \frac{\rho}{2} \| \lambda_c D y_c - z_c \|_2^2 \right\}, \tag{2.25}
\]

\[
\hat{z}_n = \arg\min_{z_n} \left\{ \| z_n \|_2^2 - w_n^T z_n + \frac{\rho}{2} \| \lambda_c D_n y_c - z_n \|_2^2 \right\}. \tag{2.26}
\]

The updates in (2.25) and (2.26) give \( C \) optimisation problems for the HR images \( Y \) and \( N \) optimisation problems for the variables \( Z \). From (2.22) the estimate of the Lagrange multiplier is:

\[
\hat{w}_c = w_c + \rho \cdot (\lambda_c D y_c - z_c), \quad \text{for all } c. \tag{2.27}
\]

**ADMM update for Y:** The optimisation problem in (2.25) is simply regularised least-squares with a closed-form solution given by:

\[
\hat{y}_c = \left( \tau_c A_c^T A_c + \rho \lambda_c^2 D^T D \right)^{-1} \left( \tau_c A_c^T x_c - \lambda_c D^T (w_c - \rho z_c) \right). \tag{2.28}
\]

This system is too large to be inverted directly, and a conjugate gradient method is often used as an alternative ([Hestenes and Stiefel] [1952]). Here, for increased speed and convergence, an approximate Newton’s method (or majorisation-minimisation scheme) is instead used. Writing the objective function as:

\[
L(y_c) = \frac{\tau_c}{2} \| x_c - A_c y_c \|_2^2 + \lambda_c w_c^T D y_c + \frac{\rho}{2} \| \lambda_c D y_c - z_c \|_2^2, \tag{2.29}
\]

its gradient can be computed as:

\[
\frac{\partial L(y_c)}{\partial y_c} = (\tau_c A_c^T A_c + \rho \lambda_c^2 D^T D) y_c + \lambda_c D^T (w_c - \rho z_c) - \tau_c A_c^T x_c, \tag{2.30}
\]
and Hessian as:
\[
\frac{\partial^2 \mathcal{L}(y_c)}{\partial y_c \partial y_c} = (\tau_c A_c^T A_c + \rho \lambda_c^2 D^T D).
\]  
(2.31)

Since the problem is quadratic, Newton’s method would solve it in one step, which is equivalent to computing the closed-form solution in (2.28). However, obtaining the full Hessian is computationally difficult. I therefore replace the true Hessian by:
\[
H_c = \text{diag}(\tau_c A_c^T A_c 1) + \rho \lambda_c^2 D^T D,
\]  
(2.32)

where \(1 \in \mathbb{R}^N\) is a vector of ones. Lemma S.3 in Chun and Fessler (2017) says that \(H^*_c \preceq H_c\), where \(H^*_c\) is the true Hessian. Since \(H^*_c \succ 0\), the approximation in (2.32) ensures convergence. Approximating the Hessian in such a way enables solving with Newton’s method in, linear time, using a multigrid technique (Ashburner, 2007). The update step for channel \(c\) is therefore:
\[
\hat{y}_c = y_c - H_c^{-1} \frac{\partial \mathcal{L}(y_c)}{\partial y_c}.
\]  
(2.33)

**ADMM update for \(Z\):** By some algebraic manipulations the optimisation problem in (2.26) can be rewritten in the equivalent form:
\[
\hat{z}_n = \arg\min_{z_n} \left\{ \frac{\rho}{2} \| z_n - \left( \frac{1}{\rho} w_n + \lambda_c D_n y_c \right) \|_2^2 + \| z_n \|_2 + \text{const} \right\},
\]  
(2.34)

where the constant contains terms that do not depend on \(z_n\). This gives an optimisation problem with an \(\ell_2\) data term and an \(\ell_1\) penalty term. A more general formulation of this problem is:
\[
\hat{s} = \arg\min_{s} \left\{ \frac{\beta}{2} \| s - t \|_2^2 + \alpha \| s \|_2 \right\},
\]  
(2.35)

which has a closed-form solution given by (proof in, e.g., Yang et al. (2009)):
\[
\hat{s}(t) = \max\left\{ \| t \|_2 - \frac{\alpha}{\beta}, 0 \right\} \odot \frac{t}{\| t \|_2},
\]  
(2.36)

where \(\odot\) denotes the Hadamard product. The solution to the optimisation problem in (2.34) is therefore given by:
\[
\hat{z}_n^{k+1} = \max\left\{ \| \frac{1}{\rho} w_n + \lambda_c D_n y_c \|_2 - \frac{1}{\rho}, 0 \right\} \odot \frac{\frac{1}{\rho} w_n + \lambda_c D_n y_c}{\| \frac{1}{\rho} w_n + \lambda_c D_n y_c \|_2}.
\]  
(2.37)
Algorithm 1 Multimodal super-resolution

1: Coregister input images ($X$).
2: Estimate model parameters ($\tau, \lambda$).
3: Initialise variables to zero ($Y$, $Z$, $W$).
4: **while** not converged **do**
5:     **for** $c = \{1, \ldots, C\}$ **do**
6:         # Loop over channels (distributed)
7:         Compute $\hat{y}_c$ by (2.33), when given $z_c$ and $w_c$.
8:     **end for**
9:     **for** $n = \{1, \ldots, N\}$ **do**
10:         # Loop over HR voxels
11:         Compute $\hat{z}_n$ by (2.37), when given $\hat{y}_n$ and $w_n$.
12:     **end for**
13:     **for** $c = \{1, \ldots, C\}$ **do**
14:         # Loop over channels
15:         Compute $\hat{w}_c$ by (2.22), when given $\hat{y}_c$ and $\hat{z}_c$.
16:     **end for**
17: **end while**

2.4.3 Denoising

By simplifying the forward model to:

$$x_c = y_c + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \tau_c^{-1}) \ , \quad (2.38)$$

that is, assuming the projection matrices as identity, the model can likewise be used for either single- or multi-channel denoising. The likelihood function becomes:

$$-\ln p(x_c \mid y_c) = \frac{\tau_c}{2} \|x_c - y_c\|^2 + \text{const.} \quad (2.39)$$

The optimisation procedure follows through unchanged.

2.4.4 Implementation Details

Algorithm 1 shows the steps for processing a set of thick-sliced patient scans with the proposed model. The process is computationally intensive and care is taken to provide an efficient implementation. The software is written in a mixture of MATLAB and C code. There are large memory requirements, which are likely to exceed the RAM of some workstations. To save memory, all the computations are performed using single precision floating point, which has negligible effect on the numerical stability. It is solved for $Y$ efficiently using a multigrid technique.
2.5. Evaluation

(Ashburner, 2007). The Newton update can be iterated over to get a closer fit to the solution. In this work, however, a single update is used to reduce computational time. Furthermore, the loop solving for each $Y$ is distributed, so that a separate process handles each iteration. Finding the optimal step-size $\rho$ is an open-problem (Dohmatob et al., 2014), and though under mild conditions ADMM converges for any value of $\rho$, the convergence rate depends on $\rho$. Here, the heuristic is used:

$$\rho = \sqrt{\frac{\text{mean}(\{\lambda_c\}_{c=1}^C)}{\text{mean}(\{\tau_c\}_{c=1}^C)}},$$

(2.40)

which is observed empirically to give good convergence properties. Furthermore, algorithm convergence is defined by the relative change in objective value (i.e., the gain):

$$\text{gain} = \frac{l_{k+1} - l_k}{l_k},$$

(2.41)

being less than $10^{-4}$. The objective value $l_k$, for iteration $k$, is computed from (2.14). Finally, the HR images’ FOV is defined as the bounding box that contains all LR images’ FOV.

The voxel size of the HR images are set to 1 mm isotropic. For the assumptions underlying the projection matrices to hold I therefore have to reslice LR images to have at least 1 mm in-plane resolution. In reality, the voxel size of the HR images should be defined as the smallest voxel size in the corresponding observed LR image. The 1 mm assumption is made here solely for computational reasons.

2.5 Evaluation

In this section the aim is to evaluate the proposed model on clinical data, first for MRI super-resolution and subsequently for CT image denoising. By doing so the hope is to answer a series of questions: (1) Does the MTV prior improve upon other classically used regularisation methods for super-resolution; (2) can the model process large clinical datasets; and (3), does the performance of predictive machine learning models improve, after application of the proposed model? Finally, proof of concept results are shown for applying the denoising model as part of estimating qMRI maps.
Figure 2.8: Example of simulating and super-resolving from the IXI dataset. Two LR images were simulated from two HR images (PDw and T2w). The HR images were then reconstructed using four methods: BS, FOT and TV are single-channel SR techniques, capable of combining LR images of only one channel; MTV on the other hand, a multi-channel method, combines information from both channels.

2.5.1 Super-Resolution

Here, the effectiveness of the MTV prior is assessed, compared with classically used regularisers, for super-resolving multi-contrast MR datasets. Furthermore, the proposed model’s ability to process a large clinical MRI dataset is investigated.

Simulated MRI Data

Because of the challenge in quantifying the results of super-resolution applied to MR images, the model was validated on thick-sliced data simulated from the IXI dataset (detailed in Section 1.6). The IXI subjects were used to: (1) validate the robustness of the noise variance estimation by adding known amounts of Rician noise and estimating it; (2) compare the heuristic for setting regularisation parameters, with optimal values obtained by grid-search; (3) compare the efficiency of two iterative methods (conjugate gradient and approximate Newton) for solving a quadratic optimisation problem; (4) compare four different methods for
super-resolving MRIs: 4th degree B-spline (BS) interpolation, first-order Tikhonov (FOT), TV and MTV.

LR images were generated by applying the forward projection operator ($A$) to HR images, such that LR images had thick slices in one direction. Thick-slice directions were picked randomly, but were always orthogonal across contrasts (e.g., axial for T1w, coronal for T2w, sagittal for PDw). Figure 2.8 shows an example of simulated LR images from HR images. Comparisons between methods are based on the root-mean-square error (RMSE):

$$\text{RMSE}(\mathbf{y}_{\text{recon}}, \mathbf{y}_{\text{ref}}) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_{\text{recon}}^{cn} - y_{\text{ref}}^{cn})^2},$$

and the peak signal-to-noise ratio (PSNR):

$$\text{PSNR} = 20\log_{10} \left( \frac{\max \left( \mathbf{y}_{\text{ref}}^{cn} \right)}{\text{RMSE}(\mathbf{y}_{\text{recon}}^{cn}, \mathbf{y}_{\text{ref}}^{cn})} \right),$$

two widely used metrics for image reconstruction. These metrics were computed for each contrast. Dice scores between GM and WM segmentations were also computed:

$$\text{Dice score} = \frac{2 \times TP}{2 \times TP + FP + FN},$$

where TP refers to the number of true positives, FP to false positives and FN to false negative. The segmentations were obtained by applying SPM12’s unified segmentation algorithm (with default parameters) to both the super-resolved and the reference images. The probabilistic segmentations were thresholded at 0.5 in order to compute the number of misclassifications.

FOT, defined in (2.7) and TV, defined in (2.9), were implemented within the same framework as MTV. Therefore, they used the same projection matrices and parameters. B-spline interpolation is not technically a super-resolution technique, but is often used in practice to reslice LR images prior to automated processing. For B-spline interpolation, if multiple LR images of the same contrast are available, they are simply averaged.

**Noise Precision:** All 1,728 IXI scans were used in this experiment. A known amount of Rician noise, as a percentage of the mean image intensity (1%,
Table 2.1: Validating the estimate of the noise precision parameter (for 1,728 subjects). Showing the simulated noise percentages, and the noise percentages estimated by fitting the two class Rician mixture model. Shown as mean±std.

<table>
<thead>
<tr>
<th>Ground-truth (%)</th>
<th>1</th>
<th>2.5</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated (%)</td>
<td>1.10±0.76</td>
<td>2.54±0.51</td>
<td>5.23±1.01</td>
<td>10.31±2.49</td>
</tr>
</tbody>
</table>

2.5%, 5%, 10%), was added to each image. The two-class Rician mixture model was then fit to the intensity histogram of the noisy images. The variance of the cluster with the lowest noncentrality parameter was used as an estimate of the image noise percentage. Table 2.1 shows the mean and standard deviation, across subjects, of the estimated noise variance percentage. These results show that the Rician mixture model can accurately estimate a wide range of noise levels. Note that there is already Rician noise present in the images. However, the amount of Rician noise that is added is an order of magnitude greater than the intrinsic noise.

**Regularisation Parameter:** Four images from different subjects and contrasts (T1w, T2w, PDw and diffusion-weighted (DW)), were used to validate whether the heuristic devised in Section 2.4.1 to estimate the prior parameter ($\lambda$) yields proper regularisation values. LR images with 7 mm slice-thickness were generated and a grid-search over the regularisation parameter was performed, (in the range $[10^{-4}, 1]$, with step size 0.2). For each value, PSNR between the resulting MTV super-resolved images and the known ground-truth was computed. The DW image was included to investigate whether the heuristic generalises to contrasts not part of the training set. The result of each grid-search, with the corresponding heuristic estimates marked by crosses, can be seen in Figure 2.9. This experiment shows that the method devised to estimate the regularisation parameter allows near-optimal reconstructions to be produced, even for unseen contrasts.

**Approximate Newton Solver:** The update-step for $Y$ entails solving a large linear system, which is often done iteratively using the conjugate gradient method (Hestenes and Stiefel, 1952). Here, it is shown that for this particular prob-
2.5. Evaluation

Figure 2.9: Validation of the heuristic used to determine the regularisation parameter. A grid-search over the regularisation parameter ($\lambda$) was performed for four different contrasts, and PSNR between the reconstruction and reference image was computed. The estimated regularisation parameter is marked by a black dot. Here, the DW data was held out, in that it was not used to learn the estimation of the parameter.

Problem, an approximate Newton method (based on a majoriser of the full Hessian and solved with a multigrid algorithm) converges faster than the conjugate gradient method. T1w, T2w and PDw image were simulated, with 6 mm slice thicknesses and different thick-slice directions. The MTV super-resolution algorithm was then run, where the update for $Y$ was solved either using the conjugate gradient or the approximate Newton method. Both solution methods were run for 50 iterations. Figure 2.12 shows the evolution of the negative log-likelihood over computation time. This total computation time takes into account both the number of iterations and the computation time per iteration. The approximate Newton solver converges faster than the conjugate gradient solver. Note that no pre-conditioning was used, which may have led to
improved conjugate gradient solver performance.

**Increasing Number of LR Images:** A single-channel experiment is performed to verify that image quality improves for an increasing number of LR observations (when for a $c \in (2.1), I_c > 1$). For 20 subjects, LR images were simulated with a slice thickness seven times greater than the in-plane resolution. Every triplet of these LR images were simulated having orthogonal thick slice directions. Furthermore, for more than three images, the simulations additionally included a 1 mm translational shift. 1 mm isotropic images are then reconstructed using BS averaging, FOT and TV. Figure 2.11 shows the results of the experiment. It can be seen that the reconstruction quality increases as more LR images become available. Furthermore, TV consistently reconstructs the highest quality images. As for the optimal number of LR images, it seems that four images provide the best trade-off between acquisition

![Super-resolving images of increasing slice-thicknesses](image-url)
2.5. Evaluation

Figure 2.11: Super-resolving with an increasing number of LR images (average of 20 subjects). Working on only a single-channel, an increasing number of LR images are simulated and subsequently super-resolved. PSNRs are then computed. It can be seen that TV consistently outperforms the two other methods.

Multi-channel Super-Resolution: All IXI subjects were used to simulate LR images. For each subject, the slice direction and thickness (between 2 and 8 mm) were chosen at random. The simulated LR images were super-resolved using BS, FOT, TV and MTV. Among these, only MTV makes joint use of information distributed across contrasts. Example reconstructions obtained with each method can be seen in Figure 2.8.

For each contrast, PSNR was computed between the reference and super-resolved images. Table 2.2 shows the average PSNR and MTV obtained the greatest mean and lowest standard deviation. Figure 2.10 additionally shows average PSNRs for different slice thicknesses, in which MTV once again performs favourably.
Figure 2.12: Comparing the conjugate gradients and approximate Newton methods for solving for $Y$. Time in seconds is plotted against negative log-likelihood. It can be seen that the approximate Newton method has faster convergence.

Table 2.2: Results comparing single- vs multi-channel super-resolution. BS, FOT and TV are all single-channel techniques, MTV is multi-channel. PSNRs are shown as mean±std.

<table>
<thead>
<tr>
<th>Channel</th>
<th>BS</th>
<th>FOT</th>
<th>TV</th>
<th>MTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1w</td>
<td>29.89 ± 9.35</td>
<td>30.51 ± 10.76</td>
<td>30.71 ± 10.84</td>
<td>31.19 ± 9.41</td>
</tr>
<tr>
<td>T2w</td>
<td>28.10 ± 8.02</td>
<td>28.71 ± 8.85</td>
<td>28.98 ± 8.96</td>
<td>29.58 ± 7.86</td>
</tr>
</tbody>
</table>

Reference and super-resolved images were also segmented into GM and WM and CSF using SPM12, with the reference segmentation considered as ground-truth, and the Dice coefficient was computed. The results can be seen in Table 2.3 where MTV obtained the highest Dice score.

Clinical MRI Data

The previous section showed that multi-channel super-resolution outperformed some established single-channel techniques, on simulated data. This result is
2.5. Evaluation

Reconstruct to 1 mm using baseline method

Reconstruct to 1 mm using super-resolution method

Figure 2.13: Example results for baseline vs super-resolution, for a randomly selected patient’s MR scans. The box on top shows the three input clinical scans (FLAIR, T1w and T2w). These three scans are reconstructed to 1 mm isotropic voxel size, using either the baseline (left) or the super-resolution method (right). It is clear that the multi-channel segmentation output (GM and WM), produced from the super-resolved images have better anatomical detail. For example, the WM has a clearer delineation close to the cortex.
Table 2.3: Results for segmenting super-resolved images (for 576 subjects). Dice scores for different reconstruction methods were computed for GM, WM and CSF tissue segmentations, using HR segmentations as references. Results shown as mean±sd.

<table>
<thead>
<tr>
<th>Tissue class</th>
<th>BS</th>
<th>FOT</th>
<th>TV</th>
<th>MTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>0.831±0.001</td>
<td>0.838±0.001</td>
<td>0.842±0.002</td>
<td>0.887±0.002</td>
</tr>
<tr>
<td>WM</td>
<td>0.866±0.001</td>
<td>0.874±0.002</td>
<td>0.878±0.001</td>
<td>0.891±0.001</td>
</tr>
<tr>
<td>CSF</td>
<td>0.842±0.002</td>
<td>0.852±0.001</td>
<td>0.856±0.001</td>
<td>0.885±0.002</td>
</tr>
</tbody>
</table>

Figure 2.14: Evaluation of the super-resolution method on clinical data. Feature vectors are obtained by concatenating smoothed GM, WM and Other segmentations. These segmentations are produced from multi-channel MRIs that has been reconstructed to 1 mm isotropic voxel size using either super-resolution or a baseline method. The prediction targets are either the age or sex of the patients. Predictive performance is then evaluated, for both, methods using Gaussian process models in the PRoNTo toolbox.

promising as the method next will be evaluated on real, clinical-grade MR data – a far more challenging scenario. Conversely to simulated data, there is no ground truth for clinical data. The model will therefore be evaluated implicitly by giving its super-resolved images as input to machine learning models to predict known, noise-free features: age and sex.

The dataset consists of 1,046 patient’s MRIs, with 615 males and 431 females. The dataset was acquired on a diversity of scanners and clinical indications at UCLH (University College London Hospitals, London, UK) in the context of routine clinical care. The dataset comprises three contrasts (T1w, T2w and FLAIR), each with a different thick-slice direction; the average voxel size over the whole dataset is [0.54, 0.54, 6.50] mm. Many of the images have only partial brain coverage in the thick-slice direction, with the outer most slices excluded. The age distribution of the dataset is shown in Figure 2.15a.
The gist of the validation is to train machine learning models to predict age and sex from tissue segmentations of patient MR images. It is now well known that these features can be accurately predicted from brain images (Monté-Rubio et al., 2018; Smith et al., 2019; He et al., 2018; Cole et al., 2017). Here, the procedure of Monté-Rubio et al. (2018) is followed. Since (accurate) normalised segmentations capture relevant and important anatomical features of the MRIs, predictive accuracy can be used as a measure of the quality of a super-resolution model.

For each subject in the dataset, the LR images were super-resolved to 1 mm isotropic with BS or MTV, and then segmented using SPM12’s unified segmentation routine (Ashburner and Friston, 2005), with default parameters. This routine outputs normalised, non-modulated (i.e., not scaled by Jacobian determinants) GM, WM and Other (1 - GM - WM) maps, that were smoothed with a Gaussian kernel of 12 mm FWHM. The concatenated smoothed maps were used as a feature vector for machine-learning. Two Gaussian process models were trained, using 10-fold cross-validation, to predict age and sex from this feature vector. The PRoNTo toolbox (Schrouff et al., 2013) was used to make these predictions. Using a deep neural network model could have been an option; however, a recent study by He et al. (2018) suggests that a kernel-based method such as the GP model used in this section, can perform comparably for predictive tasks on brain data. The authors discuss different potential reasons for this result, including limited amounts of training data and hyper-parameter settings of the neural network models.

The results of both the regression and classification task are shown in Table 2.4. Age regression results are reported in years using the RMSE, error standard deviation (SD), mean error (bias) and Pearson’s correlation coefficient; sex classification accuracy is reported in percentage, and the lower and upper bound over the 10 folds are included. Furthermore, for the MTV case, a scatter plot of the individual predictions for the regressions task is shown in Figure 2.15b and a receiver

---

8Kernel-based methods are non-parametric machine learning models that predicts, on new test data, based on a learned weighted average of all training data. The weighted average is computed using a predefined kernel. Deep neural network models, on the other hand, does not require a predefined kernel, but can be seen as instead learning this kernel from the training data.
Table 2.4: Results for predicting age and sex from normalised brain segmentations, from hospital MR scans (for 1,046 patients). The super-resolution model was compared to a baseline method, which reslices using 4th degree B-spline interpolation. For sex classification, the lower and upper bounds over the 10 folds are included.

<table>
<thead>
<tr>
<th>Method</th>
<th>Age (years)</th>
<th>Sex (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>SD</td>
</tr>
<tr>
<td>Baseline</td>
<td>6.91</td>
<td>9.89</td>
</tr>
<tr>
<td>Super-resolution</td>
<td>6.32</td>
<td>8.66</td>
</tr>
</tbody>
</table>

operator curve (ROC) for the classification task is shown in Figure 2.15c. Images and non-smoothed native space segmentations, for an example patient, are shown in Figure 2.13. The results make it quite clear that the segmentations produced from the super-resolved images have more predictive power. This is true for both the regression and classification task (c.f. Table 2.4). In particular for the classification task, where the improvement in accuracy is 3.3 percentage points. It also evident from the example segmentations in Figure 2.13 that anatomical features are more clearly defined in the super-resolved segmentations. Furthermore, comparing, e.g., the super-resolved close-up of the T1w image with its baseline counterpart, it can be seen that the mismatched FOVs have been filled in for the super-resolution case. Finally, the runtime of the algorithm to super-resolve three LR images – as was performed in this evaluation – is under 30 minutes on a modern workstation. The runtime scales linearly with the number of channels, but parallelisation, when used, will improve it.

2.5.2 Denoising

In this section the model’s denoising capability will be explored on two types of data: clinical CT scans and MR images used to estimate qMRI maps. For denoising, the simplified generative model in (2.39) is used, where the projection matrix is set to identity. The evaluation on CT data follows the same idea as for the super-resolution experiment on hospital MRI data. The experiment on qMRI data is simply a proof of concept.
Figure 2.15: Super-resolving clinical MRI scans. (a) Age distribution of the 1,046 patients in the clinical MRI dataset: mean value is 40.1 and standard deviation is 14.3 years. (b) Scatter plot of the individual predictions for the regressions task. (c) ROC curve for the classification task, where an area under the curve (AUC) of 0.97 was obtained.
Figure 2.16: Denoising clinical CT scans. (a) Example axial view of a CT image (with two different intensity windows); (b) average intensity distribution of 100 CT images, the mode around -1,000 is air; (c) Example GMM fit to truncated CT intensity distribution, where the orange dashed line represents the air class.
2.5. Evaluation

Denoising Clinical CT Scans

For denoising CT data (an example CT image is shown in Figure 2.16a) the methods for setting the noise precision ($\tau$) and regularisation ($\lambda$) parameters must be modified. The noise distribution in CT images is Gaussian, not Rician (Gravel et al., 2004). Furthermore, the Hounsfield scale in Table 1.1 defines the intensity of air as -1,000. By empirically studying the average intensity histogram of 100 CT scans (Figure 2.16b), it can be seen that it should be possible to get an estimate of the noise precision from the variance of the mode that is located at -1,000. The image noise can therefore be estimated by simply fitting a GMM to the truncated intensity histogram of a CT scan. As there can be foreign objects that have intensities close to that of air (such as a head immobiliser – seen in Figure 2.16a), a $K = 2$ component GMM is fitted to the histogram, which has been truncated in the range -1,020 and 980, and then selecting the mode with a mean closest to -1,000 (Figure 2.16c). This method is found empirically to work well for estimating the noise in CT images. For the regularisation parameter, as CT images are quantitative, the value can be set to a fixed constant. Performing a grid-search with visual inspection gives $\lambda = 0.05$, which produces satisfactorily denoised images.

The dataset that is used consists of 1,025 patient’s CT images, with 520 males and 505 females. The dataset was acquired on a diversity of scanners and clinical indications at UCLH in the context of routine clinical care. The average voxel size over the whole dataset is $[0.43, 0.43, 0.78]$ mm. The age distribution of the dataset is shown in Figure 2.15a.

The evaluation on denoising CT data follows the same principle as the evaluation on hospital MRI data for super-resolution, illustrated in Figure 2.14. The baseline method is the CT data without denoising applied. Hence, predictive performance for age and sex is compared on denoised and non-denoised images. The evaluation metrics are the same as for MRI super-resolution.

The results of both the regression and classification task are shown in Table 2.5. Age regression results are reported in years using the RMSE, SD, bias and Pearson’s correlation coefficient; sex classification accuracy is reported in percentage, and the
Table 2.5: Results for predicting age and sex from normalised brain segmentations, extracted from hospital CT images (for 1,025 patients). Predicting from segmentations obtained from non-denoised and denoised images is compared. For sex classification, the lower and upper bounds over the 10 folds are included.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>Age (years) SD</th>
<th>Bias</th>
<th>Correlation</th>
<th>Accuracy</th>
<th>Sex (%) Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-denoised</td>
<td>5.88</td>
<td>3.79</td>
<td>-1.55</td>
<td>0.50</td>
<td>85.5</td>
<td>83.2</td>
<td>87.5</td>
</tr>
<tr>
<td>Denoised</td>
<td>5.21</td>
<td>3.07</td>
<td>-1.10</td>
<td>0.55</td>
<td>87.5</td>
<td>85.3</td>
<td>89.4</td>
</tr>
</tbody>
</table>

lower and upper bound over the 10 folds are included. Furthermore, for the denoised case, a scatter plot of the individual predictions for the regressions task is shown in Figure 2.15b, and a ROC curve for the classification task is shown in Figure 2.15c. An example noisy and subsequently denoised CT image is shown in Figure 2.18.

The results implies that the denoising step improves predictive performance for both the sex regression and the age classification task.

**Improving SNR of qMRI Maps**

This section will cover a simple qualitative proof of concept of a method for increasing SNR of qMRI maps. This work was done in collaboration with Dr Christian Lambert at the Wellcome Centre for Human Neuroimaging. Dr Lambert provided the data used in the evaluation.

The qMRI maps were estimated with the hMRI-toolbox (Tabelow et al., 2019). This toolbox creates quantitative maps from multi-echo T1w, PDw and MTw spoiled gradient echo acquisitions (Figure 2.19). The echoes all have sub-millimetre resolution, as it is believed that qMRI data at such resolution will facilitate improved biophysical modelling (Weiskopf et al., 2015). However, as the SNR in MRI inherently drops with smaller voxel volume (Edelstein et al., 1986), it is difficult to increase the spatial resolution of qMRI towards sub-millimetre voxels. Denoising methods could therefore be used to increase the SNR of sub-millimetre MR images, post-acquisition. Furthermore, the ESTATICS model, implemented in the hMRI-toolbox, assumes noise-free images. Denoising the images prior to model-fitting would therefore adhere more closely to this assumption.

The idea for using the MTV model to improve the SNR of qMRI maps is quite simple. Given a set of MR images of different echoes: \( X = \{x_c\}_{c=1}^C \) =
Figure 2.17: Denoising clinical CT scans. (a) Age distribution of the 1,025 patients in the clinical CT dataset: mean value is 84.0 and standard deviation is 7.0 years. (b) Scatter plot of the individual predictions for the regressions task. (c) ROC curve for the classification task, where an AUC of 0.92 was obtained.
Figure 2.18: Example of a CT image before (a) and after denoising (b) with the MTV model. The image on the right has visually less noise, while still retaining finer details.

\[ \{ x_{T1}^{1}, \ldots, x_{T1}^{C}, x_{PD}^{1}, \ldots, x_{PD}^{C}, x_{MT}^{1}, \ldots, x_{MT}^{C} \} \], so that \( C = C_{T1} + C_{PD} + C_{MT} \). Treat all images as separate channels of the MTV model and minimise:

\[
- \ln (p(X | Y) p(Y)) = \sum_{c=1}^{C} \frac{\tau_c}{2} \|x_c - y_c\|_2^2 + \sum_{n=1}^{N} \sqrt{\sum_{c=1}^{C} \|\lambda_c D_n y_c\|_2^2} + \text{const.}
\]

This gives \( C \) denoised images where gradient information has been propagated between all of the input MRIs. The ESTATICS model is then fitted to the denoised images using the hMRI-toolbox. Note that prior to denoising, the echoes are registered using \texttt{spm_coreg} of SPM12. This registration takes into account that, for one contrast (e.g., T1), the echoes are already in alignment. Therefore, only two image registrations are performed: the two first echoes of each contrast series (e.g., MT and PD), to the contrast chosen as reference (e.g., T1). All other echoes of the contrasts being registered only has the affine matrix applied to their headers.

In order to qualitatively, investigate if applying the MTV model in (2.45) can improve the SNR in qMRI, in total, 22 image volumes were obtained. These volumes correspond to multiple echoes of three different imaging modalities: eight
2.6 Discussion

Figure 2.19: The hMRI-toolbox estimates qMRI maps by fitting the MPM model to multi-echo T1w, PDw and MTw spoiled gradient echo acquisitions. Individual windowing has been applied to the qMRI maps (windowing level shown at the top of each map).

echoes of T1w images, eight echoes of PDw images and six echoes from a dual excitation FLASH MTw sequence. All images were acquired from a healthy volunteer and they all have isotropic 0.8 mm voxel sizes. These images are then given as input to the hMRI toolbox to estimate qMRI maps, having applied either no denoising or MTV denoising on the input echoes. The results from both approaches are shown in Figure 2.20. Although just a simple proof of concept study, on a single subject, it seems as if applying the MTV model prior to qMRI map estimation does improve the SNR. Further work by Dr Lambert showed that test-retest reliability was greatly improved by the denoising procedure (not presented here).

2.6 Discussion

This chapter presented a resolution recovery model that can be applied to large datasets of clinical neuroimaging scans, both for super-resolving and denoising. Commonly in such datasets of MR images, each patient has a collection of scans acquired with different sequences and thick-slice directions. Currently, when performing some multi-channel analysis, these images are often simply interpolated to
Figure 2.20: Proof of concept results on improving SNR in qMRI maps. To investigate if
the MTV model can be used to improve the SNR of quantitative MRI, \(8 \times T1w\), \(8 \times PDw\) and \(6 \times MTw\) images were given as input to the hMRI toolbox. The estimated maps produced by the toolbox without applying denoising are shown in (a), and with MTV denoising in (b). Visually, it is apparent that the SNR has improved in the denoised qMRI maps.

the same size. The proposed super-resolution model is an alternative to interpolation that better leverages these multiple patient scans. Additionally, for single-channel analysis (e.g., using T1w MRIs) when more than one patient scan exists (e.g., a T1w, T2w and PDw), the model can be used to propagate information among scans to obtain more informative images.

The approach builds on a principled probabilistic generative model. The novelty of the model lies in the MTV prior, which allows a joint probability distribution across image channels to be modelled. All model parameters are set automatically, and no fine-tuning is necessary, allowing extensive datasets, with large instrumental variability, to be processed. It was shown that when the super-resolution model was used as a preprocessing step for MR data, or the denoising model for CT data, subsequent machine-learning tasks had improved results.
2.6. Discussion

Data-driven models currently show state-of-the-art performances in many image processing tasks and could be an alternative to the approach proposed in this chapter. However, generalisability is still an issue with many such methods, as they excel in scenarios where the unseen data is close to data the model was trained on, but does not extrapolate well to out-of-sample test data. This is because data-driven models construct an empirical prior from the data, which leads to a flexible, highly parametrised model – where minor prior assumptions are necessary – but where overfitting to a training population can be an issue. This is especially risky with pathological imaging, because pathology adds further diversity on already hugely diverse normal biology. Model-based methods, on the other hand, require prior assumptions to be made. To design priors as flexible as ones learned by data-driven techniques is extremely challenging, and simplified assumptions are therefore often made. If the prior assumption is close to the data generating process, and the forward model and statistical properties of the data is too, then model-based methods can perform well on out-of-sample data (and will not require any retraining). Interestingly, methods have been developed that combine model- and data-driven approaches (e.g., Adler and Öktem (2018); Dalca et al. (2019b); and the method that I propose in Section 4.5). This combination would be an interesting future direction for the super-resolution model proposed in this chapter.

The proposed model could be of value in translating methods that have shown good results on research data to clinical imaging. For example, many techniques based on machine (deep) learning show promising results on analysing neuroimaging data (Litjens et al., 2017). However, a model trained on HR data may struggle when given as input LR clinical data. Applying the super-resolution model as a preprocessing step, reconstructing HR versions from the LR input, could facilitate this transition. The evaluation also showed that the tissue segmentation performance of a widely used neuroimaging package improves when the multi-channel input images are super-resolved, compared with simply being interpolated.

However, metric scores such as those used in this study are specific to the data that was used to evaluate the model. Any claim that a method generalises to the
huge variability present in clinical MR scans should therefore be taken with a pinch of salt. This is not only because the imaging data varies greatly: in image contrast, intrasubject alignment and voxel size; but also because the scanner acquiring the images may not have computed the correct voxel sizes, slice-thicknesses, etc. In the evaluation, a clinical dataset with a large variability among patient scans was used. Still, this is no guarantee that the model will work on any clinical MR data. However, data specificity is likely to be greater the more flexible the model is, such that highly parametrised models may suffer more from this specificity.

The validation on CT data presented in this work only dealt with denoising. However, CT volumes often have anisotropic voxels too, with the additional difficulty that their thickness can vary across the image. With an appropriate definition of the projection matrix, super-resolving CT images is possible. The projection matrix will then need to take into account the variable slice-thickness that CT scans sometimes have (example in Section 1.4.1). This variable slice-thickness could be read of the image DICOM header and then used to construct the projection matrix.

The proof of concept results for improving the SNR in qMRI (Figure 2.20) seems to imply that denoising could be a valuable part of a qMRI estimation pipeline. However, instead of applying the denoising prior to ESTATICS map creation, as was done here, or afterwards, which is done elsewhere (Mohammadi et al., 2017); improved results could be obtained by treating denoising as regularisation in a model-based reconstruction of the parameter maps. Some parameters could then be estimated precisely, whereas others less so. With a regularisation approach, there would be more smoothing for the imprecisely estimated parameters, and those that can be accurately estimated would be less smoothed. This could be an interesting avenue for future work, as accurate quantitative parameter maps of an individual could enable two interesting ideas:

1. Currently, most supervised segmentation machine learning methods require ground truth training labels. However, even ex vivo ‘ground truth’ data has limitations; anatomical boundaries are defined semi-objectively based on a limited number of tissue properties across small sample sizes. Instead, mul-
2.6. Discussion

Timodal qMRI maps, including diffusion and connectivity measures, high resolution R1, R2*, PD, MT and susceptibility could be used to automatically derive anatomical labels. Jointly modelling this multimodal data, whilst accounting for image resolution differences, will better characterise the microstructural tissue properties, allowing objective mapping of anatomical regions.

2. Given more accurate quantitative parameter maps of an individual, it becomes possible to simulate MRI scans with a wide variety of other MRI contrasts (Drobnjak et al., 2010). In future, deep learning – trained on synthetic images – may become an important component when mining information from a diverse variety of hospital scans.

A well known fact is that TV introduces stair-casing effects on flat areas, which are abundant in brain MRI. However, the MTV regularisation proposed in this chapter seems not to suffer from such artefacts. This is probably because MTV uses gradients distributed over all contrasts, which often have orthogonal thick-slice directions. Hence, an image area containing flat gradients in one channel may very well have more informative gradients in another channel. Another factor that suppresses these stair-casing artefacts is the parametrisation of the differential operator, which computes the gradient using both the forward and the backward finite differences.

The model proposed in this chapter could be made, possibly, more robust in a few ways. One way relates to the assumptions regarding the slice profile and slice gap. These parameters are highly variable and assuming them as fixed, as is currently done, can lead to inexact super-resolved images. Of course, a user could change these parameters themselves if they know their values. However, they sometimes cannot be obtained. A principled solution to this problem would be to extend the parametrisation of the projection matrices to model both slice gap and profile. Although a non-trivial modelling problem, this would allow for estimating the most likely such parameters. Misalignment between scans could be another explanation for poor super-resolved images, as the edge-based prior distribution is highly
dependent on well registered LR images. Rather than performing an initial rigid registration of the LR images (as is currently done) improved alignment could be achieved by modifying the forward model in (2.3) to incorporate a rigid transformation. The optimal parameters could then be found by Gauss-Newton optimisation, similar to what is done in [Ashburner and Ridgway, 2013]. This will be discussed further in Section 5.2.2.

Extending the generative model to embrace the physics of the MR image formation would be an interesting future direction. This has already been done in the context of MR image segmentation [Fischl et al., 2004]. In that work, the observed voxels were modelled by the equation underlying the spoiled gradient echo acquisition [Ernst and Anderson, 1966], with additive Gaussian noise. This is in contrast to the model presented in this chapter, in which an observed voxel intensity is simply assumed to be the realisation of some unknown signal. Modelling the image intensities as in [Fischl et al., 2004] would allow for a super-resolution method more robust to sequence parameters, as these could in practice be pulled from the DICOM header.

Finally, the super-resolution could, in theory, be performed as to reconstruct images with voxels smaller than the smallest observed voxel size. This should help in decreasing partial volume effects. However, for such a method to perform well – in the context of a generative model – a prior is needed for this higher resolution data, possibly more informative than the MTV prior used in this work. Such prior information could be encoded, for example, by a high-resolution atlas and a model for voxel intensities. Such models will be presented in Chapter 4 of this thesis.
Chapter 3

A Novel Cost Function for Groupwise Image Registration

Contents

3.1 Introduction .................................................. 104
3.2 Intermodal Image Registration ............................... 104
3.3 Related Work .................................................. 108
3.4 Methods ....................................................... 110
  3.4.1 Total Variation Functionals ............................ 112
  3.4.2 Normalised Multi-Channel Total Variation .......... 112
  3.4.3 Lie Groups and Rigid-Body Transformations ....... 114
  3.4.4 Groupwise Constraint .................................. 115
  3.4.5 Implementation Details ................................. 116
3.5 Evaluation ..................................................... 116
  3.5.1 Simulated Data ........................................... 118
  3.5.2 Clinical Data ............................................ 123
3.6 Discussion ..................................................... 129
3.1 Introduction

In clinical practice, it is common to acquire, for a single patient, a wide range of different imaging modalities, such as various MRI sequences and CT imaging. This is for highlighting different structures or pathologies. As patient movement between scans (or scanning sessions) is unavoidable, correcting this misalignment by image registration is often an essential step for any form of multi-channel analysis. For example, the multimodal super-resolution approach discussed in the previous chapter required edges to be aligned; and the multi-channel segmentation method, that will be presented in subsequent chapter, will require voxel intensities to represent the same tissue class, across images. Good image alignment is what fulfil these requirements.

This chapter introduces a novel cost function for registration of such multimodal images. It is based on the MTV distribution, which was used in Chapter 2 for image reconstruction. The MTV distribution has the advantage of enabling simultaneous alignment of multiple (more than two) images. Here, its use is justified theoretically in the context of image registration. It is shown that this distribution is well suited as a groupwise cost function, and that it is robust to both strong bias field corruptions and large misalignments.

3.2 Intermodal Image Registration

Intermodal rigid-body registration is a fundamental task in medical imaging, which aims to align volumes representing the same object, acquired with different modalities (e.g., MR sequences, CT, PET) or at different times. Automated registration methods often solve this problem by optimising a transformation parameter (R) of some cost function (C), which should have its optimum when images are well aligned. Figure 3.1 shown a set of multimodal patient images, before and after such image registration. For intramodality alignment (e.g., two T1w MR scans), the differences between scans can be assumed to be independent with Gaussian noise, so

\footnote{Different time-points are here close enough so that no morphological change due to, e.g., ageing, has occurred. If this is not the case, then other forms of registration is needed (similarity/affine/nonlinear).}
3.2. Intermodal Image Registration

Figure 3.1: Example of multimodal patient scans: CT, PDw, T1w, T2w (world-space axial slices), before (left) and after (right) image registration. MR-to-CT registration is, in general, a more challenging problem than MR-to-MR.

that the cost function reduces to the sum of squared differences. The difficulty with intermodality alignment (e.g., a T1w MR scan and a CT image) comes from the fact that such scans are not repeated measures of the same signal and, therefore, a simple model of ‘measurement error’ (i.e., sum of squared difference) cannot be used.

An increasing number of image analysis tasks calls for the alignment of not just a pair of images, but multiple images: \( I = \{I_1, \ldots, I_C\} \). For example, the three MR sequences used to estimate qMRI maps (see Section 1.4.2), or the multimodal registration example in Figure 3.1. Two methods for registering multiple images are by pairwise and groupwise alignment (see Figure 3.2).

A pairwise approach takes one image in the group as a reference \( (I_r) \) and then registers all the other images to this reference in a pairwise manner, by solving multiple optimisation problems:

\[
\hat{R}_c = \arg\min_{R_c} C_p(I_r, I_c \circ R_c), \quad \text{for all } c \neq r, \quad c = 1, \ldots, C, \quad (3.1)
\]

where \( C_p(\cdot, \cdot) \) is some cost function and \( I \circ R \) resamples an image \( I \) by some rigid transformation \( R \). However, such a pairwise approach has two distinct shortcomings: (1) the choice of the reference image inherently biases the resulting transfor-
Figure 3.2: Comparison of pairwise (left) and groupwise (right) image registration. One approach to register multiple (more than two) images is to take one image in the group as a reference and register all other images to this reference in a pairwise manner. Another alternative is to perform a groupwise registration in which all transformations are optimized simultaneously.

A groupwise approach, on the other hand, defines a global cost function $C_g(\cdot, \ldots, \cdot)$ and performs a groupwise registration:

$$\hat{\mathcal{R}} = \arg\min_{\mathcal{R}} C_g(I_1 \circ R_1, \ldots, I_C \circ R_C),$$

(3.2)

where $\mathcal{R} = \{R_1, \ldots, R_c\}$ is the set of transformations that maps the coordinates from some common reference domain to the domain of each image. In such groupwise registration, all the transformations are optimized simultaneously. Here, transformations are expressed with respect to a common reference space, thereby removing the need for choosing a particular reference image, and the bias associated with that choice. Additionally, a global cost function simultaneously takes into account all information in the group of images, which makes for a more principled method and possibly improved alignment.

Information theoretic cost functions, such as mutual information, have historically been successful at the task of multimodal, intrasubject image registration, and have been used in both pairwise and groupwise settings. However, as they com-
3.2. Intermodal Image Registration

Figure 3.3: MR images that have very strong bias field corruption can be difficult to register, accurately, with information theoretic approaches. Shown is an example of registering a PDw image to a T1W image, in the presence of either a very strong (left), or no bias field corruption (right). The registered images are shown, as well as their joint intensity histograms, before and after registration. It can be seen that, for the bias free images, the initial joint histogram has less mutual information, whilst the final histogram has more so (it is more diagonal). For the images with strong corruption, the histograms does not change as much after registration.

Commonly rely on joint image intensity histograms – whose size grows exponentially with the number of images – they become prohibitive for aligning larger number of images, and sparsity additionally becomes an issue (Wachinger and Navab, 2012). Furthermore, cost functions based on intensity histograms can struggle in the presence of large rigid offsets, variable FOV over the images, or strong bias field corruptions (Saad et al., 2009; Greve and Fischl, 2009), an example of the latter is shown in Figure 3.3. As such properties can all be present in clinical neuroimaging scans; could an improved registration method be devised for such data?

In this chapter, a novel cost-function for groupwise multimodal image registration is proposed, which is based on the MTV distribution from Chapter 2. It is demonstrated, using simulated multicontrast MRI, with different noise levels, bias fields and slice thickness, that the proposed MTV-based cost function consistently outperforms well-known registration methods part of the SPM12 software. The proposed cost function is additionally evaluated on real clinical data, where it performs
well for multicontrast MRI registration.

### 3.3 Related Work

Automatic image registration has a long history of research in medical imaging, hence many such methods have been proposed over the years and many review articles exist, e.g., by Maurer and Fitzpatrick (1993); Hill et al. (2001); Oliveira and Tavares (2014). Normally, for registering images, a cost function is defined and minimised, as in (3.1) and (3.2). Such cost function is commonly composed of a data term and a penalty term, which impose a trade-off between image similarity and a constraint imposed on the parametrisation of the transformation model. The most commonly used cost functions are based on intensity cross-correlation, intensity differences and information theory.

The cross-correlation and its derived measures, such as the Pearson’s correlation coefficient or correlation ratio, is a common image registration cost function (Lewis, 1995; Cideciyan, 1995; Roche et al., 1998; Maintz et al., 1995; Van den Elsen et al., 1995; Maes et al., 1997). The cross-correlation is based on the assumption that there is a linear relationship between the intensities of the corresponding structures in both images. The measures based on the intensity difference, on the other hand, are usually based on the sum of squared differences or their normalisations (Gerlot-Chiron and Bizais, 1992; Ashburner and Friston, 1997; Ashburner et al., 1997; Hajnal et al., 1995; Woods et al., 1998). The assumption behind the sum of squared differences computed from the voxel intensity is that the corresponding structures in both images should have identical intensities. Another method based on intensity is the Woods criterion (Woods et al., 1993), which seeks to minimise the standard deviation between corresponding voxels in different modalities. In Myronenko and Song (2010), a method was proposed that relaxed the assumption of intensity independence and stationarity, which most intensity-based similarity measures rely on.

Multimodal intensity-based pairwise registration is often solved using mutual information (Collignon et al., 1995; Viola and Wells III, 1997; Wells III et al., 1998).
Mutual information assumes a stochastic relationship between the images to be registered based on the Shannon entropy that is computed from the joint probability distribution of the image voxel intensity. This joint probability is often constructed from multidimensional image intensity histograms. In addition to joint histograms, GMMs, similar to those used in Chapter 4, could also be used to encode the joint probability distribution [Orchard and Mann, 2009]. Normalised mutual information was proposed a few years later [Studholme et al., 1998], being less sensitive to the dimensions of overlapping image regions. In fact, mutual information based cost functions received enough attention as to produce a review article just a few years later, with over 200 works on the topic [Pluim et al., 2003].

In the seminal RIRE challenge (described in Section 1.6), mutual information was most accurate for CT-MRI registration, while the Woods criterion and mutual information did equally well for PET-MRI registration [West et al., 1996]. These results are quite old, but little progress has been made since, and cost functions based on mutual information are still used by default for multimodal, intrasubject registration in many widely used registration packages (e.g., SPM, ANTs, FSL).

To reduce the dependency on the image intensities, registration approaches based on aligning edges have been investigated, which include gradient magnitude correlation [Maintz et al., 1996], Canny filters [Orchard, 2007] and normalised gradients dot product [Haber and Modersitzki, 2006; Snape et al., 2016]. To match surfaces, or segmentations, has also been used as a method for registering images [Pelizzari et al., 1989; Hemler et al., 1995; Greve and Fischl, 2009; Xiaohua et al., 2005; Aganj and Fischl, 2017]. However, such methods are sensitive to the quality of the surface or segmentation.

A number of methods have been proposed to perform multimodal groupwise registration. In Orchard and Mann (2009) it was proposed to use a GMM instead of histograms to approximate the joint probability density functions and in Spiclin et al. (2012) the joint probability density function was approximated with a non-parametric approach based on a hierarchical intensity-space subdivision scheme. Wachinger et al. (2007) proposed a ML approach that accumulated all pairwise es-
estimates of mutual information for all possible pairs of images in a group. In [Joshi et al. (2004)] the mean squared differences was used as a pairwise metric to compare every image in a group to an average template image. The linear scaling of the computational complexity with respect to the number of images in the group and possibility to parallelize the algorithm, makes it feasible for large groups of images. [Bhatia et al. (2007b)] proposed to use the normalized mutual information as a pairwise similarity metric in a similar framework as in [Joshi et al. (2004)], for monomodal registration. The framework has since been used in the literature as a metric for multimodal groupwise registrations ([Huizinga et al., 2016; Hallack et al., 2014; Ceranka et al., 2018]). The work of [Polfliet et al. (2018)] built on that of [Joshi et al. (2004)] to do multimodal registration where the cost function was based on conditional entropy and principal component analysis was used to re-estimate the template image. The idea was that not using the average image as the template image should be more appropriate in multimodal data with intensities of varying scales, ranges and contrast.

The approach proposed in the next section has similarities with some of these methods in that it is based on edges, to remove the dependency on image intensities, and that it minimises a group-wise cost function; however, without the need to estimate a template image as all images of the group are resampled to an average space, in which the cost function is computed.

### 3.4 Methods

If a distribution over images has its optimum when the images are in alignment, it can be used as a cost function for image registration. Recall the MTV prior introduced in Chapter 2, a distribution over gradients of an arbitrary number of images (of possibly different contrasts):

\[
p(Y) \propto \prod_{n=1}^{N} \exp \left( -\sqrt{\sum_{c=1}^{C} \| \lambda_c D_n y_c \|_2^2} \right), \quad (3.3)
\]
which is here presented in its discrete form. In this section, the distribution in (3.3) is introduced as a cost function for groupwise, multimodal registration.

The motivation comes from the fact that the TV of a one-dimensional discrete signal is the log of a Laplace distribution over its finite differences, and the Laplace distribution has been shown to fit best the distribution of gradients in natural images (Huang and Mumford, 1999). This leads me to use MTV as a cost function for groupwise image registration, as it is the multi-channel equivalent of that best fit distribution. There are multiple definitions of such vector-valued TV; but I will show that, out of two popular definitions, only MTV is suitable in the context of image registration. MTV finds its optimum when the edges are co-localised across channels. This property should make MTV well-suited for co-registering images of different medical imaging modalities, where edge information is shared between channels, but intensity distributions can vary substantially.

As mentioned in Section 1.4.2, MR data commonly suffer from image non-uniformities caused by inhomogeneities in the B1 or B0 magnetic fields. The intensity inhomogeneity can severely challenge quantitative image analysis algorithms, e.g., image registration software. It is commonly assumed that intensity inhomogeneities can be ascribed to a smooth, spatially varying (bias) field, which is a multiplicative component of the measured image. The spatial dependency of the bias field originates, among other things, in the distance from the measured objects to the receive coils. The smooth nature of the bias field implies that a registration algorithm based on edges should be less affected by this corruption, as multiplication with a smoothly varying field will have negligible affect on the image gradient.

\(^2\)In the context of image registration, it is preferable to work in a continuous setting. I will therefore do that in this section. Furthermore, \(x\) and \(y\) commonly represent spatial locations in such a setting. I will therefore here switch from using these variables to represent data and unknown, to here use them for locations; and instead use \(f\) for observed data and \(\mu\) for the unknown.
3.4.1 Total Variation Functionals

The (isotropic) TV of a differentiable scalar-valued function \( f: \mathbb{R}^D \rightarrow \mathbb{R} \) is the integral of the \( \ell_2 \)-norm of its gradient:

\[
\text{TV}(f) = \lambda \int_{\Omega} \| \nabla f(x) \|_2 \, dx, \quad \Omega \subset \mathbb{R}^D,
\]

where \( \lambda \) is a scaling parameter that relates to the variance of the Laplace distribution (see Section 2.4.1) and \( D \) is the dimensionality of the domain of the function (for volumetric medical images, commonly \( D = 3 \)).

For vector-valued functions \( f: \mathbb{R}^D \rightarrow \mathbb{R}^C \), where \( f = \{ f_1, \ldots, f_C \} \) and \( f_c: \mathbb{R}^D \rightarrow \mathbb{R}, \ c = 1, \ldots, C \), two TV-based functionals that can be devised are:

\[
\text{CTV}(f) = \sum_{c=1}^{C} \lambda_c \text{TV}(f_c),
\]

\[
\text{MTV}(f) = \int_{\Omega} \left\| \left( \lambda_c \nabla f_c(x)\right)_{1 \leq c \leq C} \right\|_2 \, dx, \quad \Omega \subset \mathbb{R}^D.
\]

Here, CTV considers channels independently (Blomgren and Chan, 1998), whereas MTV applies the norm to the joint gradients over all channels (Sapiro and Ringach, 1996). By assuming that each channel is composed with a rigid transformation \( R_c: \Omega \rightarrow \Omega \), I will now show that the value of CTV does not change, therefore making it unsuitable as a cost function for rigid image registration.

Defining \( u = R_c x \), integration by substitution on an individual TV term (in (3.4)) gives:

\[
\int_{\Omega_c} f_c(x) \, dx = \int_{\Omega_u} f_c(u) |J| \, du,
\]

where \( |J| \) is the absolute value of the determinant of the Jacobian matrix of \( u \). As the determinant of a rigid transformation is one, I get \( |J| = 1 \). Hence, the value of an individual TV term does not change with the application of a rigid transformation, and therefore neither do the CTV term (as it is a sum of independent TVs). This is not the case for the MTV, due to the application of the norm across channels.

3.4.2 Normalised Multi-Channel Total Variation

Images are not continuous but discrete, and can be defined on non-overlapping domains. Let \( \mathcal{I} = \{ I_1, \ldots, I_C \} \) be a set of discrete images representing the same object
that may have different FOV and numbers of voxels. This set of images are made to represent a continuous vector-valued signal using cubic B-spline interpolation, so that \( I \Rightarrow f : \mathbb{R}^D \rightarrow \mathbb{R}^C \).

Each image is assumed a displaced version of a true signal \( \boldsymbol{\mu} : \mathbb{R}^D \rightarrow \mathbb{R}^C \), so that:

\[
 f_c(y) = \mu_c(\xi_c(x)), \quad c = 1, \ldots, C,
\]

where

\[
 \xi_{q_c}(x) = I_{3,4}M_c^{-1}R_cM_{\mu} \begin{bmatrix} x \\ 1 \end{bmatrix},
\]

\[
 I_{3,4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.
\]

Here, \( M_c \) is the \( c \)th image’s subject voxel-to-world mapping (that is read from the image’s NIfTI header) and \( M_{\mu} \) is the template voxel-to-world mapping that is computed initially from all of the input subject images orientation matrices (together with suitable dimensions), defining a common space domain \( \Omega_{\mu} \).

The MTV of the true (aligned) signal can then be written as:

\[
 MTV(f, R) = \int_{\Omega_{\mu}} \left\| \lambda_c \nabla f_c \left( \xi_c^{-1}(y) \right) \right\|_{1 \leq c \leq C}^2 dy, \quad \Omega_{\mu} \subset \mathbb{R}^D,
\]

where \( R = (R_1, \ldots, R_C) \) are the set of rigid-body transformations to be estimated. Furthermore, since for each channel, the image of the common space \( R_c^{-1}\Omega_{\mu} \) might differ from the observed space \( \Omega_c \), the gradients are nulled outside of the observed FOV, so that the MTV term only involves observed voxels.

As interpolation has a significant impact on the gradients shape, it is prevented from biasing the optimum by removing the individual TV terms from the cost function in (3.11), arriving at the proposed, groupwise normalised multi-channel total variation (NMTV):

\[
 NMTV(f, R) = MTV(f, R) - \sum_{c=1}^{C} TV(f_c, R_c),
\]

\[\text{Here, a parallel can be drawn to the negative mutual information, which can be computed as the difference of the joint and individual entropies: } \text{MI}(f_1, f_2) = - (H[f_1, f_2] - H[f_1] - H[f_2]).\]
Figure 3.4: Explaining why the NMTV cost function is modulated by $1/C$. By plotting the gradient magnitude (gm) for one channel $gm(f_1) = \|\nabla f_1(x)\|_2$, vs the NMTV computed from keeping the gradient magnitude of all other channels fixed (for different values), it can be seen that NMTV without modulation does not reach the correct minimum (left), while NMTV with modulation does (right).

where the individual terms have been modulated with the reciprocal of the number of channels:

$$MTV(f, R) = \int_{\Omega_{\mu}} \left\| \left( \frac{\lambda_c}{C} \nabla f_c \left( \xi^{-1}_c(y) \right) \right) \right\|_2 \, dy, \quad \Omega_{\mu} \subset \mathbb{R}^D, \quad (3.13)$$

$$TV\left( f_c, R_c \right) = \frac{\lambda_c}{C} \int_{\Omega_{\mu}} \left\| \nabla f_c(\xi^{-1}_c(y)) \right\|_2 \, dy, \quad \Omega_{\mu} \subset \mathbb{R}^D. \quad (3.14)$$

This modulation is necessary for the cost function to find its optimum when gradient magnitudes are in alignment. This is shown in Figure 3.4.

3.4.3 Lie Groups and Rigid-Body Transformations

In this thesis, rigid-body transforms are considered in terms of their membership of the special Euclidean group in three dimensions $SE(3)$ (Eade, 2013; Ashburner and Ridgway, 2013). This group is a Lie group and can be equivalently encoded by its Lie algebra $se(3)$. Working with the Lie group representation of $SE(3)$ gives
3.4. Methods

A lower-dimensional, linear representation for rigid body motion\(^4\) one which is increasingly being used in computer vision these days.

An orthonormal basis of the Lie algebra \(\mathfrak{se}(3)\) is:

\[
\mathcal{B} = \begin{cases} 
\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\
\begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{cases}
\tag{3.15}
\]

Therefore, a 3D rigid-body transform can be encoded by a vector \(q_c \in \mathbb{R}^6\) and recovered by matrix exponentiating the Lie algebra representation:

\[
R_{q_c} = \exp \left( \sum_{i=1}^{6} q_i B_i \right).
\tag{3.16}
\]

Conversely, by using a matrix logarithm, the encoding of a rigid-body matrix can be obtained by projecting it on the algebra:

\[
q_i = \text{Tr} \left( B_i \log(R_{q_c})^T \right).
\tag{3.17}
\]

### 3.4.4 Groupwise Constraint

Given an optimal set of transforms, one can generate infinitely many optimal sets by composing them with an additional transform. This phenomenon is sometimes referred to as a drift of the common space. To avoid this issue, the common space is required to be a mean of the individual observed spaces. This is enforced by constraining the cost function in (3.12) as:

\[
\text{NMTV}(f, R) = \text{MTV}(f, R) - \sum_{c=1}^{C} \text{TV}(f_c, R_{q_c}), \quad \text{s.t.} \quad \sum_{c=1}^{C} q_c = 0.
\tag{3.18}
\]

\(^4\)For example, a rotation matrix \(R\) in the \(\text{SE}(3)\) group has nine parameters (constrained by \(R^T R = I\) and \(\det(R) = +1\)), although in reality, there are only three degrees of freedom. Using Lie algebra enables working with only a three parameter representation.
3.4.5 Implementation Details

The FOV of the common space \((\Omega_\mu)\) is defined as the maximum bounding-box that contains all individual images (defined from their orientation matrices and dimensions). This FOV is automatically adjusted as the transformation matrix is estimated, so to not make the cost function go to zero if images are moved outside the FOV. To improve the speed of the algorithm, a three step coarse-to-fine scheme is used where the voxel size of the common space decreases from 6 to 4 to 2 mm, while at the same time smoothing the input images with a Gaussian kernel whose FWHM decreases from 8 to 4 to 0 (no smoothing) mm. A random jitter is introduced to the sampling grid of the input images, as a trick to reduce interpolation artefacts (Penny et al., 2011). The variable voxel size of the input images are accounted for in the computation of the gradients, by dividing each gradient direction with the width of the voxel. Finally, the algorithm is parallelised over channels, that is \(C\) in (3.18), which significantly decreases the runtime for larger numbers of channels.

The scaling parameters \(\lambda = \{\lambda_1, \ldots, \lambda_C\}\), which normalise the cost function across modalities, are estimated from each individual image’s intensity histogram (see Section 2.4.1). If an image contains only positive values (e.g., an MR image), a Rician mixture model is used and the scaling parameter is set as the mean of the brain class. If an image contains also negative values (e.g., a CT image), a GMM is used. The scaling parameter is then set as the absolute difference between the mean of the air class and the mean of the brain class. This is for the parameter estimate to be invariant to the units of the data (see Figure 3.5).

3.5 Evaluation

This section compares, for rigid, intrasubject, multimodal image registration, the NMTV cost function in (3.18) against the cost functions implemented in the co-registration routine of SPM12 (\texttt{spm.coreg}). This routine implements mutual information (MI; Wells III et al. (1996b); Maes et al. (1997)), normalised mutual information (NMI; Studholme et al. (1998)), entropy correlation coefficient (ECC; Maes et al. (1997)) and normalised cross correlation (NCC; Lewis (1995)).
cost functions are optimised using Powell’s method (\texttt{spm_powell}; \textcite{Press2007}). Powell’s method is an algorithm for finding a local minimum of a function by repeated line-searches, where the function need not be differentiable, and no derivatives are taken. The same implementation of Powell’s method is used to optimise the NMTV cost function (with the same convergence settings).

For NMTV, the alignment is optimised over all images at once (a group-wise registration), while for the SPM12 cost functions, one image is set as reference and all other images are aligned with this fixed reference (a pair-wise registration). Each transform is encoded by three translations (in mm) and three Euler angles (in
degrees), such that $\mathbf{R}^* = \mathbf{T}_x\mathbf{R}_y\mathbf{R}_z$, where:

$$
\mathbf{T}_x\mathbf{R}_y\mathbf{R}_z = \begin{pmatrix}
1 & 0 & 0 & t_x \\
0 & 1 & 0 & t_y \\
0 & 0 & 1 & t_z \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & \cos(r_x) & \sin(r_x) & 0 \\
0 & -\sin(r_x) & \cos(r_x) & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\cos(r_y) & 0 & \sin(r_y) & 0 \\
0 & 1 & 0 & 0 \\
-\sin(r_y) & 0 & \cos(r_y) & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\cos(r_z) & \sin(r_z) & 0 & 0 \\
-\sin(r_z) & \cos(r_z) & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
$$

(3.19)

In the next two sections, the cost functions will first be compared on simulated data and then on real clinical data.

### 3.5.1 Simulated Data

Because of the challenge in quantifying the results of image alignment on real, clinical data – where a ground-truth is rarely available – an in-depth validation study is here performed on simulated data from the BrainWeb dataset (described in Section 1.6).

**Cost Function Analysis**

When deriving new cost functions for image registration it is common practice to plot the value of the cost function vs varying one of the transformation parameters. This is done in order to determine if there is a global minimum. A cost function without such a minimum is less likely to find an optimal alignment during optimisation.

Therefore, to investigate if the NMTV cost function has a global minimum, the value of the cost function in (3.12) is plotted, as a function of relative offset between two 1 mm isotropic, noise-free T1w and T2w BrainWeb images. The resulting plot is shown in Figure 3.7. From this plot it is clear that there is a global and sharp minimum when the images are aligned at 0 mm. I can therefore conclude that NMTV should be a suitable cost function for image registration.
3.5. Evaluation

**Figure 3.6:** Visualisation of the simulation process on BrainWeb data for an example T1w image. The reference T1w MRI scan is first degraded as to be more like routine clinical data (1 - 4) and then rigidly realigned (5).

**Figure 3.7:** The value of the NMTV cost function in (3.12), as a function of translation, for two noise-free BrainWeb images. There is a global and sharp minimum when the images are aligned at 0 mm, which implies that NMTV is a suitable cost function for image registration.
Cost Function Comparison

To investigate the performance of NMTV vs other commonly used cost functions, non-degraded 1 mm isotropic T1w, T2w and PDw images (i.e., without noise and bias field corruption) from the BrainWeb simulator are used. A series of degradations are then applied to these reference BrainWeb images. This is in order to make them more similar to clinical-grade scans:

1. Intensity inhomogeneity is simulated by sampling smooth multiplicative bias fields from a multivariate Gaussian distribution, which favours fields with low bending energy.

2. Thick-sliced data is simulated by downsampling the images, in one direction, by a factor chosen between 1 and 6.

3. A percentage of Rician noise is added, chosen uniformly between 0% and 50% of the maximum image intensity.

4. Partial brain coverage is simulated by cropping 20 mm on both sides of a randomly chosen anatomical direction (axial, coronal or sagittal). A few voxels along each other direction are additionally removed – to make sure that the algorithm does not simply align the borders of the images FOV.

5. Two of the images (e.g., the T1w and T2w) are randomly selected and rigidly realigned by uniformly sampled x-, y-, z-translations (between -50 and 50 mm) and x-, y-, z-rotations (between -15 and 15 degrees). This transform is composed with the image’s original voxel-to-world matrix and therefore enables a ground-truth comparison.

Figure 3.6 shows an example of the simulation process outlined in the above steps.

I performed in total 2,000 simulations. For each simulation, the three simulated images were registered with all cost function: MI, NMI, ECC, NCC and

\[ \frac{1}{2} \int_{\Omega} \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \left( \frac{\partial^2 f(x)}{\partial x_i \partial x_k} \right)^2 \, dx, \quad \Omega \subset \mathbb{R}^3. \]  

The bending energy (also known as the biharmonic or thin plate model) is given by: The bending energy can be encoded in the precision matrix of a multivariate normal distribution. This allows for sampling this energy.
Table 3.1: Results on simulated data. For each method, the geometric mean of absolute translation and rotation errors (columns 2 and 3) are provided, as well as the coefficients of the linear fit \( \log(\text{abs}(\text{error})) \sim 1 + \text{bias} + \text{noise} + \text{downsampling} + \text{offset} \) applied to the translation errors (columns 4 to 7). Shown as mean±std.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tr. [mm]</th>
<th>Rot. [deg]</th>
<th>1</th>
<th>Bias</th>
<th>Noise</th>
<th>DS</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMTV</td>
<td>0.048±6.20</td>
<td>0.019±6.64</td>
<td>-5.68</td>
<td>0.059</td>
<td>2.91</td>
<td>0.50</td>
<td>0.0044</td>
</tr>
<tr>
<td>MI</td>
<td>0.12±14.4</td>
<td>0.042±16.0</td>
<td>-5.18</td>
<td>0.50</td>
<td>1.80</td>
<td>0.31</td>
<td>0.041</td>
</tr>
<tr>
<td>NMI</td>
<td>0.12±14.5</td>
<td>0.042±15.8</td>
<td>-5.22</td>
<td>0.53</td>
<td>1.76</td>
<td>0.33</td>
<td>0.040</td>
</tr>
<tr>
<td>ECC</td>
<td>0.12±14.6</td>
<td>0.042±15.5</td>
<td>-5.19</td>
<td>0.45</td>
<td>2.12</td>
<td>0.33</td>
<td>0.041</td>
</tr>
<tr>
<td>NCC</td>
<td>0.69±9.5</td>
<td>0.40±12.6</td>
<td>-1.47</td>
<td>0.21</td>
<td>0.13</td>
<td>0.14</td>
<td>0.015</td>
</tr>
</tbody>
</table>

NMTV. I then computed errors between the estimated transformation parameters and the known ground-truths. The geometric mean\(^6\) \( \mu \) and geometric standard deviation\(^7\) \( \sigma \) of absolute errors are computed for each method. This simulated dataset gives an unprecedented ability to evaluate the impact of different corruption parameters on registration error. To this end, a linear model is fitted to the log of the absolute translation errors generated by each method, with noise level, downsampling factor, bias magnitude and simulated offset as regressors. The corresponding maximum-likelihood slopes are written as \( \beta_n, \beta_d, \beta_b \) and \( \beta_o \).

Results

Figure 3.8 shows an example of three simulated MR images, before and after NMTV registration, and Figure 3.9 shows an example of the voxel-wise contribution to the NMTV cost function. The distribution of absolute errors, for all methods, is shown in Figure 3.10. NMTV does consistently better (\( \mu_t = 0.048 \) mm, \( \mu_r = 0.019^\circ \)), and NCC consistently worse (\( \mu_t = 0.69 \) mm, \( \mu_r = 0.40^\circ \)), than all other approaches (MI: \( \mu_t = 0.12 \) mm, \( \mu_r = 0.042^\circ \); NMI: \( \mu_t = 0.12 \) mm, \( \mu_r = 0.042^\circ \); ECC: \( \mu_t = 0.12 \) mm, \( \mu_r = 0.042^\circ \)), which are indistinguishable. Additionally, there are far fewer outliers with NMTV than with the other approaches: a cut-off at 1 mm gives 97% of success for NMTV vs 85% for MI, NMI and ECC, and just 60% for NCC.

\(^6\)The geometric mean is defined as \( \mu = \left( \prod_{n=1}^{N} x_n \right)^{\frac{1}{N}} = \sqrt[N]{x_1 x_2 \cdots x_N} \).

\(^7\)The geometric standard deviation is defined as \( \sigma = \exp \left( \sqrt{\frac{\sum_{n=1}^{N} \left( \ln \frac{x_n}{\mu} \right)^2}{N}} \right) \), where \( \mu \) is the geometric mean.
Figure 3.8: Example of simulated images from the BrainWeb dataset (T1w, T2w, PDw), before (top) and after (bottom) NMTV registration. The degradations, such as noise and bias field corruption, are clearly visible in the three MR images.

Figure 3.9: Example of the voxel-wise contribution to the NMTV cost function, before (left) and after (right) registration. Computed from simulated images from the BrainWeb dataset (T1w, T2w, PDw).
Table 3.2: Average runtime for the registrations in the three different experiments. The values are given in minutes. The experiment on simulated data used three modalities (Simulated; T1w, T2w, PDw). The two experiments on clinical data used three (Multicontrast; T1w, T2w, PDw) and five modalities, (Multimodal; T1w, T2w, PDw, CT, PET) respectively. Note that these runtimes were computed from executing the experiments on a computational cluster, without using parallelisation. Parallelisation would have speeded up the NMTV method. Results are shown as mean±std.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>NCC</th>
<th>NMTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated</td>
<td>16.2±4.5</td>
<td>14.6±3.7</td>
<td>16.1±6.3</td>
<td>23.9±4.9</td>
<td>252.1±58.4</td>
</tr>
<tr>
<td>Multicontrast</td>
<td>20.4±7.1</td>
<td>17.6±7.1</td>
<td>20.3±6.4</td>
<td>27.3±6.8</td>
<td>321.2±77.9</td>
</tr>
<tr>
<td>Multimodal</td>
<td>29.3±8.8</td>
<td>25.6±5.6</td>
<td>24.0±13.1</td>
<td>36.9±10.3</td>
<td>447.5±99.6</td>
</tr>
</tbody>
</table>

The slopes and intercepts of the log-linear fits are provided in Table 3.1. Figure 3.11 illustrates these results in more detail, for NMTV and MI. NMTV is the method most impacted by noise ($\beta_n = 2.91$, compared to MI’s $\beta_n = 1.80$) and downsampling ($\beta_d = 0.50$, compared to MI’s $\beta_d = 0.31$), but the most robust to $B^{-1}$ bias ($\beta_b = 0.059$, compared to MI’s $\beta_b = 0.50$) and to original misalignment ($\beta_o = 0.0044$, compared to MI’s $\beta_o = 0.041$).

The average runtime is shown in Table 3.2. For the pair-wise approaches, the runtime is the sum of the time taken to register the two source images to the reference scan. Note that these runtimes were computed from executing the simulations on a computational cluster, without using parallelisation. Parallelisation would have decreased the runtime of the NMTV method.

3.5.2 Clinical Data

To investigate the performance of NMTV at registering routine clinical data, images from the RIRE dataset are used (described in Section 1.6). As the RIRE dataset provides fiducial-derived, ground-truth transformations, I can compare the results of NMTV to the approaches implemented in SPM12. Two types of validations are performed to assess if: (1) the NMTV cost function compares favourably to the methods in SPM12 for multicontrast registration of clinical MR images; and (2), if the NMTV cost function can register multimodal combinations of images (i.e., MRI, PET and CT). I separate between multicontrast and multimodal to be able to investigate if the NMTV group-wise registration is less robust in the presence of
Figure 3.10: Absolute translation (mm) and rotation (deg) errors obtained from 2,000 simulations. Individual errors along the x, y, and z directions are plotted in different shades of grey, whereas the boxplot was computed from the pooled data. The vertical axis is in log-scale, so that higher points represent a greater error. The plot at the top shows absolute errors (errors below the red horizontal line are less than 1 mm or 1 deg). The plot at the bottom shows errors after normalisation by the geometric mean across methods.
3.5. Evaluation

Figure 3.11: Comparing the absolute error in translation (mm) and rotation (deg) between NMTV and MI. The errors are plotted vs bias field strength, noise level, down-sampling factor and true offset. Offsets are defined as percentages of the maximum simulated shift (50 mm/15 deg). The vertical axis is in log-scale. Errors along each dimension are considered to be independent, and a regression line is plotted in black. It can be seen that NMTV, in contrast to MI, is immune to bias field and offset.
non-MR images.

**Multicontrast Study**

This section uses the non-rectified T1w, T2w and PDw MR images of the RIRE dataset. For each of the 18 subjects, and each of the five methods, the three MR images were co-registered. For the pair-wise SPM12 approaches the reference scan was chosen at random. This study resulted in a cross-subject mean error based on manual markers, for all five methods.

The results of the study are shown in Table 3.3 Figure additionally 3.12 shows an example of MR images, before and after NMTV registration. All methods achieve sub-millimetre errors, which for the thick-sliced MR images of the RIRE dataset is below the smallest error that can be accurately measured (West et al., 1996). The SPM12 approaches perform well, which was expected as the initial misalignment was fairly small (≈10 mm) and the bias field corruption mild (cf. Figure 3.11).

The average runtime is shown in Table 3.2 For the pair-wise approaches, the runtime is, once again, the sum of the time taken to register the two source images to the reference scan. Note that, just as in the experiment on simulated data, the runtimes were computed from executing the simulations on a computational cluster; so that the runtime would have been less for the NMTV method if parallelisation would have been used.

**Multimodal Study**

This section uses the non-rectified T1w, T2w and PDw MR images, as well as the CT and PET scans, of the RIRE dataset. For each of the 18 subjects, and each of the five methods, at most six co-registrations were performed. For the pair-wise approaches, the CT image and the PET image was registered to the three MR images. For the group-wise NMTV method, all scans were registered at once. This study resulted in a cross-subject mean error based on manual markers, for all five methods.

The results of the study are shown in Table 3.4 Figure additionally 3.13 shows an example of CT, PET and MR images, before and after NMTV registration. In
3.5. Evaluation

Figure 3.12: Example of MR images from the RIRE dataset (patient_008), before (left) and after (right) NMTV registration. The complete FOV of all images, as well as more close up versions, are shown. The NMTV registered images appear well aligned.

Table 3.3: Multicontrast registration results on clinical data (using only MR images). For each method, the arithmetic mean of absolute translation and rotation errors are provided. PreReg shows errors before registration.

<table>
<thead>
<tr>
<th>Error</th>
<th>PreReg</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>NCC</th>
<th>NMTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr. [mm]</td>
<td>9.01</td>
<td>0.75</td>
<td>0.60</td>
<td>0.72</td>
<td>0.74</td>
<td>0.64</td>
</tr>
<tr>
<td>Rot. [deg]</td>
<td>1.13</td>
<td>0.91</td>
<td>0.93</td>
<td>0.93</td>
<td>0.80</td>
<td>0.85</td>
</tr>
</tbody>
</table>

In this study, NMTV does not perform as well as the other cost functions (other than NCC), while NMI gets the lowest registration error. It therefore seems as if the NMTV method may struggle when faced with non-MR images, i.e., for multimodal registration. Possible reasons for this will be discussed in Section 3.6. The average runtime is again shown in Table 3.2.
Figure 3.13: Example of CT, PET and MR images from the RIRE dataset (patient_002), before (top) and after (right) NMTV registration. The complete FOV of all images, as well as more close up versions are shown. It can be seen that, by looking at for example the close ups of the T2w and CT images (after NMTV registration), the alignment can be improved.
3.6 Discussion

This chapter presented the first use of the MTV distribution as a cost function for image registration. This distribution has previously been used for image reconstruction, in both computer vision and medical imaging. The key benefits of the proposed method, compared to many other widely used cost functions such as MI, NMI, and NCC, is that it provides a principled way of performing groupwise alignment and is robust to intensity variations due to bias fields and large misalignments. These are all important properties if one is interested in automated processing of large populations of routine clinical neuroimaging data, as these datasets often contains multiple images of each patient, which can have misalignments due to patient movement, or large bias fields. Manual inspection, of whether registration succeeded, is not an option in such large collections of images.

In the evaluation on registering multimodal clinical images, NMTV did not perform as well as MI, ECC or NMI. The NMTV cost function therefore seems to struggle at registering between MRI and non-MRI modalities. The assumption that edges are co-localised between modalities may not always be true for medical images. This could be the reason for the decrease in registration accuracy for multimodal data, as the distribution of edges, between modalities, are not as similar as within modalities. Recall, MTV should be a suitable distribution for modelling images of the same ‘group’, e.g., natural images (Huang and Mumford, 1999). Therefore, MTV works well at registering between MRI contrasts (e.g., T1w and T2w), because MR image can be considered as the same image group. But for multimodal registration, images from different groups are considered. That is, the image formation between a MR and a CT scan is completely different, just as they are in comparison to natural images captured by an optical instrument (such as a cam-

<table>
<thead>
<tr>
<th>Error</th>
<th>PreReg</th>
<th>MI</th>
<th>NMI</th>
<th>ECC</th>
<th>NCC</th>
<th>NMTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr. [mm]</td>
<td>26.31</td>
<td>7.13</td>
<td>3.01</td>
<td>3.29</td>
<td>12.26</td>
<td>8.43</td>
</tr>
<tr>
<td>Rot. [deg]</td>
<td>6.57</td>
<td>1.52</td>
<td>1.21</td>
<td>1.28</td>
<td>3.74</td>
<td>2.31</td>
</tr>
</tbody>
</table>

Table 3.4: Multimodal registration results on clinical data (MRI, CT, PET). For each method, the arithmetic mean of absolute translation and rotation errors are provided. PreReg shows errors before registration.
Another reason for the higher error for MR-to-CT registration could be that CT image intensities cover a much larger dynamic range than MR images, so that the intensity range corresponding to brain is just a fraction of the full range in the CT scan. Bone, soft tissue and background cover a majority of the intensity range. Therefore, the CT gradients that drive the registration are mostly outside of the brain. This could potentially make for poor alignment. However, some improved normalisation method could alleviate this issue. One idea, proposed in Rorden et al. (2012), transforms the CT image voxel intensities so that the brain covers a wider range. More specifically, CT intensities from -1,000HU to -100HU are transformed to 0 to 900, intensities from -99HU to 100HU are scaled linearly to 911-3100, and intensities greater than 100HU are added by 3,000. Another option is to normalise the gradient magnitude. In Haber and Modersitzki (2006) this was done by using the normalized gradient field, i.e., the local orientation, which is purely geometric information.

In this work, I used Powell’s method to optimise the NMTV cost function. Powell’s method has the advantage of not requiring deriving and computing derivatives, but is therefore an inefficient optimisation scheme as it does not take into account the descent direction of the cost function. Furthermore, it only works for cost functions with a small number of parameters, such as the six parameter rigid-body registration investigated here. For nonlinear deformation models, where the parameter space can be many orders of magnitude larger, such an optimisation scheme is not feasible. Future work will therefore investigate more efficient derivative-based optimisation techniques. This should, additionally, improve the runtime of the algorithm.

Another optimisation technique that does not require an analytical expression of the objective function, nor its derivatives, is Bayesian optimisation (Frazier, 2018). In Bayesian optimisation, a global optima, of such a potential ‘black-box’ function, is found using a sequential design strategy. This method has proven suc-
cessful in tuning the hyper-parameters of deep neural networks \cite{Snoek2012}, whose objective functions can be very expensive to evaluate. In such a setting, Bayesian optimisation can find the optimal parameters much faster than using grid-search techniques. In the context of image registration, when the imaging data are large 3D volumes, the cost function is expensive to evaluate. A Bayesian optimisation scheme could then find an optima faster than, e.g., Powell’s method, as well as being less susceptible to getting caught in local minima. Bayesian optimisation requires a Gaussian process prior over the cost function. For rigid registration, the work of \cite{Lang2014} suggests such a prior over translations and rotations, by representing these by dual-quaternions. By doing so, they can suitably be incorporated into a Gaussian processes using continuous data. Exploring Bayesian optimisation for rigid image registration would be an interesting avenue of future research.
## Improved Segmentation of Clinical Neuroimaging Data

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4.1 Introduction</strong></td>
<td>134</td>
</tr>
<tr>
<td><strong>4.2 Subject-Level Modelling of Images</strong></td>
<td>135</td>
</tr>
<tr>
<td>4.2.1 Intensity Distribution</td>
<td>137</td>
</tr>
<tr>
<td>4.2.2 Tissue Distribution</td>
<td>138</td>
</tr>
<tr>
<td>4.2.3 Model Optimisation</td>
<td>140</td>
</tr>
<tr>
<td><strong>4.3 Population-Level Modelling of Images</strong></td>
<td>141</td>
</tr>
<tr>
<td>4.3.1 Learning the Intensity Model</td>
<td>142</td>
</tr>
<tr>
<td>4.3.2 Learning the Template Model</td>
<td>147</td>
</tr>
<tr>
<td><strong>4.4 A Model for Medical Image Translation</strong></td>
<td>151</td>
</tr>
<tr>
<td>4.4.1 Methods</td>
<td>152</td>
</tr>
<tr>
<td>4.4.2 Evaluation</td>
<td>155</td>
</tr>
<tr>
<td>4.4.3 Discussion</td>
<td>161</td>
</tr>
<tr>
<td><strong>4.5 A Convolutional Neural Network MRF</strong></td>
<td>163</td>
</tr>
<tr>
<td>4.5.1 Methods</td>
<td>167</td>
</tr>
<tr>
<td>4.5.2 Evaluation</td>
<td>171</td>
</tr>
</tbody>
</table>
4.1 Introduction

The previous chapters have focused on preprocessing of routine clinical neuroimaging scans, by techniques for resolution recovery and improved multimodal alignment. In this chapter, the focus will be on extracting meaningful information from such data, by means of image segmentation and normalisation. Investigating a large number of subjects using such methods could, as was insinuated in Section 1.2, enable the discovery of population patterns that more closely map to underlying disease mechanisms and cognitive behaviour. They are additionally near imperative operations in many neuroimaging analyses, e.g., for looking at age and group studies, and shape and volume morphology.

However, automatic segmentation and normalisation of routine clinical data is very challenging. This is due to the two types of variability discussed in Section 1.3: biological and instrumental. Although instrumental variability can be alleviated by preprocessing methods, such as the one described in Chapter 2, the biological variability remains. Abnormalities, such as lesions, make this variability even greater. However, studying brains with such pathology is of considerable interest, e.g., for the introduction of automated predictive systems in clinical care.

Most normal (healthy) brains have both similar structure and tissue composition; diseased brains, on the other hand, can have an endless variability in morphology (shape and appearance). This variability means that it is difficult to assume and/or learn prior information, which is of imperative importance for automated image processing methods. For example, if a brain image is deformed to a common mean, this deformation needs to be constrained to avoid modelling implausible deformations. However, an implausible deformation in a normal brain may very well
be plausible in a non-normal one. Another example is when segmenting a brain image into different components; then, pathology may shift the spatial distribution of tissues so that a learned representation no longer models the tissue distribution accurately.

This chapter will therefore focus on two methods for improved segmentation of clinical neuroimaging data. The first method introduces a solution to the problem of variable FOV across multimodal scans of the same patient, and can furthermore be used to simulate entirely missing modalities (known as image translation). The second method marries a classical model for image segmentation with deep learning, for improved robustness on images acquired at different centres and scanners. Both methods build on top of a generative model for segmenting and normalising brain scans, which has been shown robust to a wide array of such data. This subject-level generative model, unified segmentation, will be described in detail in the next section, followed by a population-level extension. Equipped with the details of this model, novel contributions will then be presented.

4.2 Subject-Level Modelling of Images

Within the field of medical imaging, automatic segmentation of images into their constituent components is an important task. Accurate segmentation is of value in both diagnostic procedures, e.g., lesion detection (Gillebert et al., 2014), and therapeutic interventions, such as treatment (Mazonakis et al., 2001) or surgical planning (Kikinis et al., 1996; Agn et al., 2019). Manual annotation of data requires extensive training of raters, whereas automated segmentation tools can make the analyses less time consuming and more reproducible (de Boer et al., 2010). Numerous algorithms exist for this purpose and they can furthermore be subdivided into many different types. For reviews, see for example, Sharma and Aggarwal (2010) or Norouzi et al. (2014).

Some of the most robust segmentation methods (and up until recently, some of the most accurate) were based on probabilistic mixture models (Klauschen et al., 2009). These models define a probability distribution over an observed image (X),
conditioned on unknown class labels (\(Z\)) and parameters (\(\theta\)). Assuming a prior distribution over unknown variables, Bayes rule is used to form a posterior distribution:

\[
p(Z \mid X, \theta) \propto p(X \mid Z, \theta) p(Z \mid \theta),
\]

which can be evaluated or approximated. Wells III \textit{et al.} (Wells III \textit{et al.}, 1996a) introduced these types of models for brain segmentation from MR images. By assuming that the log-transformed image intensities followed a normal distribution in the likelihood term \(p(X \mid Z, \theta)\), they segmented the brain into three classes (GM, WM and CSF).

As generative models require the data-generating process to be defined (through (4.1)), they can be extended to more complex joint distributions than in [Wells III \textit{et al.} (1996a)], allowing for segmentation methods robust to, \textit{e.g.}, slice thickness, MR contrast, field strength and scanner variability. In fact, many of today’s most widely used neuroimaging analysis software, such as SPM (Ashburner and Friston, 2005), FSL (Smith \textit{et al.}, 2004) and FreeSurfer (Fischl \textit{et al.}, 2004), rely on these kinds of models, and have been shown to reliably segment a wide variety of MR data (Kazemi and Noorizadeh, 2014; Heinen \textit{et al.}, 2016). However, to accurately define the joint distribution of the data generating process requires expertise in the field of application, as well as skills in mathematical modelling and optimisation. Considerable time can therefore go into the development of such models. In contrast, deep neural network models, like variational autoencoders (Kingma and Welling, 2013) and generative adversarial network (Goodfellow \textit{et al.}, 2014), learn the generating distribution directly from the data, by automatic differentiation and back-propagation (Goodfellow \textit{et al.}, 2016). However, these methods still does not generalise well to the large variability present in medical images (Knoll \textit{et al.}, 2019).

The paper titled ‘unified segmentation’ introduced a generative model based on a probabilistic mixture model, which combined tissue classification, bias field correction and nonlinear registration of a probabilistic template (Ashburner and Friston, 2005). This model produces accurate segmentations for normal brains, and
has also been extended to cope with lesioned brains (Seghier et al., 2008; Schmidt et al., 2012; Mah et al., 2014b). Furthermore, the model underlying ‘unified segmentation’ could be used for methods that would benefit from having spatially encoded mutual information (Zhuang et al., 2011). Its implementation, available open-source as part of the SPM software, is used by thousands of neuroscience researchers worldwide. As my contributions in this chapter extend on the core building blocks of this model – its intensity distribution and distribution over latent variables – I will in the next sections explain the unified segmentation model in more detail.

### 4.2.1 Intensity Distribution

The likelihood term $p(X | Z, \theta)$ in (4.1) models the intensity distribution of the observed data $X$. Let $X \in \mathbb{R}^{N \times C}$ be a multimodal dataset from one subject, where $C$ is the number of modalities and $N$ is the number of voxels in the images. Each voxel is assumed to belong to one of $K$ classes, where the classification is encoded by the label matrix $Z \in [0, 1]^{N \times K}$, with $z_{nk} = 1$ iff. voxel $n$ belongs to class $k$. Each tissue class is associated with a multivariate Gaussian distribution of dimension $C$, which encodes the intensities’ mean ($\mu_k \in \mathbb{R}^C$) and precision ($\Lambda_k \in \mathbb{R}^{C \times C}$) over the modalities (the covariance matrix is related to the precision as $\Sigma_k = \Lambda_k^{-1}$). The GMM can then be written as a conditional probability that factorises across voxels:

$$p(X | Z, \{\mu_k, \Lambda_k\}_{k=1}^K) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(x_n | \mu_k, \Lambda_k^{-1})^{z_{nk}}.$$  

(4.2)

As MR images usually are corrupted by a smooth, spatially varying artefact that can severely hamper automated tissue classification (see Section 1.4.2), it is possible to introduce such a bias field into the model by substituting the means and precisions with:

$$\mu_{nk}^* = B_n^{-1} \mu_k,$$

$$\Lambda_{nk}^* = B_n \Lambda_k B_n.$$  

(4.3)

(4.4)

Here, the diagonal matrix $B_n$ models a (multiplicative) bias field. This bias field is parametrised by a linear combination of a small number of low frequency ba-
Figure 4.1: Showing the influence of different values of the tissue weights ($\omega$), in the distribution over unknown tissue classes \([15]\). The tissue weights are included to account for variable amounts of different tissue types. For example, because the total amount of grey matter decreases with age. This is illustrated in the above figure, by visualising the grey matter class of the default SPM template modulated by varying the value of $\omega$.

sis functions ($\beta$). The parameter vector has a multivariate Gaussian prior whose precision ($\Lambda_{\beta}$) encourages smooth fields.

4.2.2 Tissue Distribution

The prior term $p(Z \mid \theta)$ in \([4,1]\) models the distribution of the unknown tissue classes $Z$. These tissue classes are assumed drawn from a categorical distribution whose probabilities are encoded by a deformable template $a \in \mathbb{R}^{M_a \times K}$. The template is registered to the subject’s brain using a deformation field $\phi$. This assumption can
4.2. Subject-Level Modelling of Images

Figure 4.2: Probabilistic template (atlas) used by the SPM12 segmentation routine. The template contains six tissue classes: GM (a), WM (b), CSF (c), bone (d), soft tissue (e) and air/background (f). The template is nonlinearly deformed (warped) to align with a subject’s brain.
be written as the conditional likelihood:

\[ p(Z) = \prod_{n=1}^{N} \text{Cat}(z_n | \pi_n), \quad \pi_n \in \mathbb{R}^K, \]  

(4.5)

where

\[ \text{Cat}(z_n | \pi_n) = \prod_{k=1}^{K} \pi_{nk}^{z_{nk}}, \]  

(4.6)

\[ \pi_{nk} = \frac{\exp(\omega_k + \hat{a}_{nk})}{\sum_{l=1}^{K} \exp(\omega_l + \hat{a}_{nl})}, \]  

(4.7)

where \( \hat{a}_{nk} \) is an element of the template image after it has already been warped to subject space \( \hat{a} = a \circ \phi \). The vector \( \omega \in \mathbb{R}^K \) contains global class proportions\(^1\) which is optimised to account for variable amounts of different tissues (see Figure 4.1). The default template used in the unified segmentation model contains \( K = 6 \) tissues: GM, WM, CSF, bone, soft tissue and air (shown in Figure 4.2).

Originally, the registration used an initial affine registration of the template by mutual information followed by the optimisation of a small deformation model. This approach was later refined to alternate affine and diffeomorphic registration integrated within the generative model (Ashburner and Friston, 2011; Blaiotta et al., 2018), which is what is used in this work. Note that the registration is an integral part of model fitting, but I will in this thesis not go into details regarding this part of the optimisation. Hence, the voxel-wise proportion in (4.7) assumes a template already in some optimal alignment with the subject’s imaging data.

### 4.2.3 Model Optimisation

The parameters of the original unified segmentation model includes the GMM parameters, the tissue proportions and the parameters of the deformation and bias field. The objective function is:

\[ \ln p(X, Z | \Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[ \ln \pi_{nk} + \ln |B_n| + \ln \mathcal{N} \left( B_n x_n | \mu_k, \Lambda_k^{-1} \right) \right] + \ln p(\beta), \]  

(4.8)

\(^1\)Note that the unified segmentation model additionally allows for each tissue class to be modelled by a GMM (Malone et al., 2015). This improves robustness when the unimodal assumption of a tissue distribution does not hold true.
4.3 Population-Level Modelling of Images

Figure 4.3: The joint probability distribution for population-level modelling of brain scans. Random variables are in circles, observed variables are shaded, plates indicate replication. Model parameters are dotted. Diamond shapes indicate either a function (π) or parameters that are given (γ, φ). The template (a) and intensity model hyper-parameters (\{μₖ, V₀, ν₀\}) are learned when fitting the model to a population of S subjects.

where Θ is the set of all model parameters. Recall that in this chapter I assumed that the deformation, warping the template to a subject’s brain, is known. That is why I do not include the parameters of the deformation in (4.8).

Fitting all model parameters is difficult and there exists no closed form solution for finding their optimal values. Additionally, the optimal values for different parameters are dependent upon the values of others. Therefore, an iterated conditional modes approach was used, which iteratively maximised the probability of each variable conditioned on the rest. The GMM parameters was updated using an EM algorithm. The bias field and deformation parameters was estimated using a Levenberg-Marquardt strategy. A detailed description of how the optimisation problem was tackled, with derivations of derivatives and details of regularisation approaches used, can be found in (Ashburner and Friston, 2005).

4.3 Population-Level Modelling of Images

When a large dataset is available, the optimisation of the tissue template can be interleaved with the mixture model fit to each individual subject (Blaiotta et al., 2018). Furthermore, priors over the intensity parameters of the Gaussian mixture
Table 4.1: List of symbols for population-level modelling of brain scans.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>Number of subjects</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of voxels</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of tissue types</td>
</tr>
<tr>
<td>$x$</td>
<td>Image intensity</td>
</tr>
<tr>
<td>$z$</td>
<td>Latent tissue membership probabilities</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Bias field parameters</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean tissue intensities</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Precision of tissue intensities</td>
</tr>
<tr>
<td>$m_0$</td>
<td>Mean of Gaussian prior distribution over $\mu$</td>
</tr>
<tr>
<td>$b_0$</td>
<td>Scaling hyper-parameter of Gaussian prior distribution over $\mu$</td>
</tr>
<tr>
<td>$\nu_0$</td>
<td>Degrees of freedoms of Wishart prior distribution over $\Lambda$</td>
</tr>
<tr>
<td>$V_0$</td>
<td>Scale matrix of Wishart prior distribution over $\Lambda$</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Initial velocity of deformation field</td>
</tr>
<tr>
<td>$a$</td>
<td>Tissue template</td>
</tr>
<tr>
<td>$\Lambda_a$</td>
<td>Precision of Gaussian prior on $a$</td>
</tr>
<tr>
<td>$\Lambda_\beta$</td>
<td>Precision of Gaussian prior on $\beta$</td>
</tr>
<tr>
<td>$\omega$</td>
<td>Tissue proportions</td>
</tr>
</tbody>
</table>

– its means and precisions – can be defined and optimised as well. This type of learning, where subject-specific parameters are marginalised while population parameters are optimised, is known as parametric empirical Bayes (Carlin and Louis, 2000). Here, as exact marginalisation is intractable, a variational approximation is used to infer the mean and precision parameters. This extension, where the subject-specific parameters are assumed to be drawn from prior distributions that describe their variability at a population level, is described in the next sections. A graphical model representation of this joint probability distribution is shown in Figure 4.3 and all symbols are described in Table 4.1. Note that I will from now on discard modelling the bias field, so that $B_n = I$ for all $n$ (the derivation remains the same if $\text{diag}(B_n) \neq 1$). This is to give less cluttered equations. In practice, point estimates of both bias and deformation field parameters are additionally obtained as part of model fitting.

4.3.1 Learning the Intensity Model

In the context of a GMM, in order to learn from the image intensities, a prior needs to be placed on its parameters ($\{\mu_k, \Lambda_k\}_{k=1}^K$). Here, for mathematical convenience,
this prior is assumed to be the conjugate Gauss-Wishart distribution:

\[
p(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k, \Lambda_k | m_{0k}, b_{0k}, V_{0k}, v_{0k})
= \mathcal{N}(\mu_k | m_{0k}, (b_{0k}\Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | V_{0k}, v_{0k}).
\]  

(4.9)

The logs of the individual distributions in (4.9) are:

\[
\ln \mathcal{N}(\mu_k | m_{0k}, (b_{0k}\Lambda_k)^{-1}) = -\frac{C}{2} \ln 2\pi + \frac{1}{2} \ln |\Lambda_k| + \frac{C}{2} \ln b_{0k}
- \frac{b_{0k}}{2} (\mu_k - m_{0k})^T \Lambda_k (\mu_k - m_{0k}),
\]

\[
\ln \mathcal{W}(\Lambda_k | V_{0k}, v_{0k}) = \ln B(V_{0k}, v_{0k}) + \frac{v_{0k} - C - 1}{2} \ln |\Lambda_k|
- \frac{1}{2} \text{Tr}(V_{0k}^{-1}\Lambda_k),
\]  

(4.10)

with

\[
\ln B(V_{0k}, v_{0k}) = -\frac{v_{0k}}{2} \ln |V_{0k}| - \frac{v_{0k} \cdot C}{2} \ln 2
- \frac{C \cdot (C - 1)}{4} \ln \pi
- \sum_{c=1}^{C} \ln \Gamma\left(\frac{v_{0k} + 1 - c}{2}\right),
\]

(4.11)

where

\[
\Gamma(x) = \int_{0}^{\infty} u^{x-1} e^{-u} du,
\]

(4.12)

is the Gamma function. With a prior on the GMM parameters, the joint distribution becomes:

\[
p(X, Z, \{\mu_k, \Lambda_k\}^K_{k=1}) = p(X | Z, \{\mu_k, \Lambda_k\}^K_{k=1}) p(Z) \prod_{k=1}^{K} p(\mu_k, \Lambda_k).
\]  

(4.13)

VB can be used to infer the posterior over tissue classes and GMM parameters via the following approximation:

\[
q(Z, \{\mu_k, \Lambda_k\}^K_{k=1}) \approx p(Z, \{\mu_k, \Lambda_k\}^K_{k=1} | X).
\]  

(4.14)

---

2If the posterior is in the same probability distribution family as the prior, then the prior and the posterior are called conjugate distributions. The prior is called a conjugate prior for the likelihood function. A conjugate prior gives a closed-form expression for the posterior (i.e., mathematical convenience – no numerical integration is necessary). Further, conjugate priors can give intuition, by more transparently showing how a likelihood function updates a prior distribution.
A mean-field approximation is then made so that the distribution in (4.14) factorises as:

$$q(Z, \{\mu_k, \Lambda_k\}_{k=1}^K) = \prod_{n=1}^N q(z_n) \prod_{k=1}^K q(\mu_k, \Lambda_k),$$  \hspace{1cm} (4.15)

with $q(z_n) = \text{Cat}(z_n | \tilde{z}_n)$ and $q(\mu_k, \Lambda_k) = \mathcal{N}(\mu_k, \Lambda_k | \tilde{m}_k, \tilde{b}_k, \tilde{V}_k, \tilde{\nu}_k)$. The subject-level posterior parameters are denoted by a tilde: $\tilde{z}_n, \tilde{m}_k, \tilde{b}_k, \tilde{V}_k, \tilde{\nu}_k$.

A posterior parameter can be inferred by maximising the ELBO:

$$L = \mathbb{E}_q \left[ \ln p(X, Z, \{\mu_k, \Lambda_k\}_{k=1}^K) \right],$$  \hspace{1cm} (4.16)

where the expectation is taken with respect to the approximate distribution in (4.15) over all variables, except the posterior parameter to be inferred (see Section 1.5.2). Optimising all posterior parameters in such a way, in turn, results in a VEM algorithm.

The Optimal Posterior over Tissue Classes

The optimal log posterior over tissue classes is given by:

$$\ln q^*(Z) = \mathbb{E}_{q(\{\mu_k, \Lambda_k\}_{k=1}^K)} \left[ \ln p(X, Z, \{\mu_k, \Lambda_k\}_{k=1}^K) \right]$$

$$= \ln p(Z) + \mathbb{E}_{q(\{\mu_k, \Lambda_k\}_{k=1}^K)} \left[ \ln p(X \mid Z, \{\mu_k, \Lambda_k\}_{k=1}^K) \right] + \text{const}$$

$$= \sum_{n=1}^N \sum_{k=1}^K z_{nk} \ln \rho_{nk} + \text{const},$$  \hspace{1cm} (4.17)

where any term independent of $Z$ has been absorbed into an additive constant and:

$$\ln \rho_{nk} = \ln \pi_{nk} + \frac{1}{2} \mathbb{E}_{q(\mu_k, \Lambda_k)} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right]$$

$$- \frac{1}{2} \mathbb{E}_{q(\mu_k, \Lambda_k)} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right]$$

$$- \frac{C}{b_k} + \bar{v}_k (x_n - \bar{m}_k)^T \bar{V}_k (x_n - \bar{m}_k),$$  \hspace{1cm} (4.18)

where

$$\mathbb{E}_{q(\mu_k, \Lambda_k)} \left[ (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right] = \frac{C}{b_k} + \bar{v}_k (x_n - \bar{m}_k)^T \bar{V}_k (x_n - \bar{m}_k),$$  \hspace{1cm} (4.19)

$$\mathbb{E} \left[ \ln |\Lambda_k| \right] = C \cdot \ln 2 + \ln |\bar{V}_k| + \sum_{c=1}^C \psi \left( \frac{\bar{v}_k + 1 - c}{2} \right),$$  \hspace{1cm} (4.20)

and

$$\psi(x) = \frac{d}{da} \ln \Gamma(a),$$  \hspace{1cm} (4.21)
is the digamma function. Taking the exponential of both sides of (4.17), and requiring the resulting distribution to be normalised, gives:

$$q^*(Z) = \text{Cat}(z_n | \tilde{z}_n) = \prod_{n=1}^{N} \prod_{k=1}^{K} (\tilde{z}_{nk})^{z_{nk}}, \quad (4.22)$$

where

$$\tilde{z}_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}. \quad (4.23)$$

Hence, the optimal posterior for \(q(Z)\) is a categorical distribution with parameters \(\tilde{Z}\) and expectation:

$$E[z_{nk}] = \tilde{z}_{nk}. \quad (4.24)$$

The quantities in (4.24) are known as the responsibilies.

The Optimal Posterior over GMM Parameters

The optimal log posterior over the GMM parameters is given by:

$$\ln q^*(\{\mu_k, \Lambda_k\}_{k=1}^{K}) = \sum_{k=1}^{K} \sum_{n=1}^{N} E[z_{nk}] \left[ \ln \mathcal{N}\left( x_n \mid \mu_k, \Lambda_k^{-1} \right) + \ln \pi_{nk} \right] + \sum_{k=1}^{K} \ln p(\mu_k, \Lambda_k). \quad (4.25)$$

The solution is a Gauss-Wishart distribution:

$$q^*(\mu_k, \Lambda_k) = \mathcal{N}\left( \mu_k \mid \tilde{m}_k, (\tilde{b}_k \Lambda_k)^{-1} \right) \mathcal{W}\left( \Lambda_k \mid \tilde{V}_k, \tilde{\nu}_k \right), \quad (4.26)$$

with parameters:

$$\tilde{b}_k = b_{0k} + \sum_{n=1}^{N} \tilde{z}_{nk}, \quad (4.27)$$

$$\tilde{m}_k = \frac{b_{0k} m_{0k} + \sum_{n=1}^{N} \tilde{z}_{nk} x_n}{\tilde{b}_k}, \quad (4.28)$$

$$\tilde{V}_k = v_{0k} + \sum_{n=1}^{N} \tilde{z}_{nk}, \quad (4.29)$$

$$\tilde{V}_k^{-1} = V_{0k}^{-1} + \sum_{n=1}^{N} \tilde{z}_{nk} x_n x_n^T + b_{0k} m_{0k} m_{0k}^T - \tilde{b}_k \tilde{m}_k \tilde{m}_k^T. \quad (4.30)$$
Population-Level Extension

The joint distribution in (4.13) can easily be extended to a population of $S$ images:

$$p \left( \{X_s, Z_s, \{\mu_{sk}, A_{sk}\}_{k=1}^{K} \}_{s=1}^{S} \right) = \prod_{s=1}^{S} \left[ p \left( X_s \mid Z_s, \{\mu_{sk}, A_{sk}\}_{k=1}^{K} \right) \right] p (Z_s) \prod_{k=1}^{K} p (\mu_{sk}, \Lambda_{sk})$$.

(4.31)

The combined ELBO can then be written as:

$$L = \sum_{s=1}^{S} \mathbb{E}_{q_{\theta}} \left[ \ln p \left( X_s, Z_s, \{\mu_{sk}, A_{sk}\}_{k=1}^{K} \right) \right]$$.

(4.32)

which is simply a summation of all the individual ELBOs. In this case, empirical population priors over image intensities can be obtained by optimising the combined ELBO in (4.32) with respect to the Gauss-Wishart prior hyper-parameters:

$$\arg\max_{\text{m}_{0k}, \text{b}_{0k}, \text{V}_{0k}, \nu_{0k}} \left\{ \sum_{s=1}^{S} \mathbb{E}_{q_{\theta}} \left[ \ln p \left( X_s, Z_s, \mu_{sk}, A_{sk} \right) \right] \right\}, \quad k = 1, \ldots, K.$$

(4.33)

Population-Level Intensity Updates

Here, given a set of individual posterior parameters, the optimal updates of the Gaussian prior hyper-parameters are provided (derived from (4.33)). In contrast to Blaiotta et al. (2018), the hyper-parameters are now updated using closed-form solutions, except for the degrees of freedom of the Wishart distribution, which is updated using an iterative Gauss-Newton scheme (as in Blaiotta et al. (2018)). These update equations are:

$$\hat{m}_{0k} = \left( \sum_{s=1}^{S} \hat{V}_{sk} \hat{V}_{sk} \right)^{-1} \left( \sum_{s=1}^{S} \hat{V}_{sk} \hat{V}_{sk} \hat{m}_{sk} \right),$$

(4.34)

$$\hat{b}_{0k}^{-1} = \frac{1}{C \cdot S} \sum_{s=1}^{S} \hat{V}_{sk} (\hat{m}_{0k} - \hat{m}_{sk})^{T} \hat{V}_{sk} (\hat{m}_{0k} - \hat{m}_{sk}),$$

(4.35)

$$\hat{V}_{0k} = \frac{1}{S \cdot \nu_{0k}} \sum_{s=1}^{S} \hat{V}_{sk},$$

(4.36)

$$\frac{\partial L}{\partial \nu_{0k}} = -\frac{1}{2} \left( S \cdot \left( \ln |\hat{V}_{0k}| + \psi \left( \frac{\nu_{0k}}{2} \right) \right) - \sum_{s=1}^{S} \left( \ln |\hat{V}_{sk}| - \psi \left( \frac{\nu_{sk}}{2} \right) \right) \right),$$

(4.37)

$$\frac{\partial^2 L}{\partial \nu_{0k}^2} = -\frac{S}{4} \psi' \left( \frac{\nu_{0k}}{2} \right).$$

(4.38)
The intensity hyper-parameters are initialised as:

\[ m_{0k} = 0, \quad b_{0k} = 1, \quad V_{0k} = I, \quad \nu_{0k} = C, \]

while the intensity posterior parameters are initialised from their corresponding images’ mean and precision.

### 4.3.2 Learning the Template Model

The construction of a tissue template could be done by simply averaging a number of individual segmentations after they have been spatially normalised. A more elegant approach would involve modelling the fact that individual segmentations are realisations of a stochastic process governed by a prior anatomical model (as in (4.31)), which can be inferred from a large data set of individual observations (Bhatia et al., 2007a; Ashburner and Friston, 2009; Ribbens et al., 2010). Templates constructed in such a framework could have the ability to capture the peculiar anatomical features of populations poorly represented by standard anatomical atlases, such as elderly or diseased populations. This would not only lead to more accurate segmentation results, but as a direct consequence, also increase the reliability of subsequent data analyses, which build models of the segmented data to infer or predict clinically meaningful information.

In the context of the generative model, just as in the case of the intensity model, simply putting a prior on the tissue template enables learning it from a population of images:

\[ p(a) = \mathcal{N}(a \mid 0, \Lambda_a^{-1}). \quad (4.39) \]

In contrast to Blaiotta et al. (2018), where the template was encoded directly as a categorical distribution (with a conjugate Dirichlet prior), I in this work instead assume it is drawn from a multivariate normal distribution (Ashburner and Friston, 2009). The precision matrix \( \Lambda_a \) of this normal distribution encodes a block diagonal matrix (\( K \) blocks of size \( N_a \times N_a \)). The idea is that a template generated from a smaller number of subjects may be a poor representations of the anatomy of the general population, hence some form of spatial blurring may regularise this average.
The joint distribution becomes:

\[
p\left(\{X_s, Z_s, \{\mu_{sk}, \Lambda_{sk}\}_{k=1}^K\}_{s=1}^S, a\right) = \prod_{s=1}^S \left[ p\left(X_s \mid Z_s, \{\mu_{sk}, \Lambda_{sk}\}_{k=1}^K\right) \right. \\
\left. \quad p\left(Z_s \mid a\right) \prod_{k=1}^K p\left(\mu_{sk}, \Lambda_{sk}\right) \right] \\
p(a), \tag{4.40}
\]

whose graphical model representation is shown in Figure 4.3. The new ELBO is:

\[
\mathcal{L} = \ln p(a) + \sum_{s=1}^S \mathbb{E}_{q_s} \left[ \ln p\left(X_s, Z_s, \{\mu_{sk}, \Lambda_{sk}\}_{k=1}^K \mid a\right) \right] \tag{4.41}
\]

Discarding terms that do not involve \(a\) gives the following optimisation problem:

\[
\arg\max_a \left\{ \ln p(a) + \sum_{s=1}^S \mathbb{E}_{q_s} \left[ \ln p\left(Z_s \mid a\right) \right] \right\}. \tag{4.42}
\]

**Population-Level Template Update**

No closed-form solution exists for the optimisation problem in (4.42). Instead an iterative Gauss-Newton method is used:

\[
a \leftarrow a - H_a^{-1} \nabla_a \mathcal{L}(a), \tag{4.43}
\]

which is solved using a multigrid method. Therefore, the gradient and Hessian of:

\[
\mathcal{L} = \ln p(a) + \sum_{s=1}^S \mathbb{E}_{q_s} \left[ \ln p\left(Z_s \mid a\right) \right], \tag{4.44}
\]

with respect to \(a\) are needed. Expanding the above expression gives:

\[
\mathcal{L} = -\frac{1}{2} a^T \Lambda_a a + \sum_{s=1}^S \sum_{n=1}^{N_s} \sum_{k=1}^K [\tilde{z}_{snk} \ln \pi_{snk}] + \text{const}, \tag{4.45}
\]

where \(\tilde{z}_{snk}\) is the responsibility from (4.24), but warped into the space of the template, and \(\pi_{snk}\) is given by (4.7).

The multinomial part \((\mathcal{L}_m)\) of (4.45), for a single subject and a single voxel,
can be written as:

\[
\mathcal{L}_m = \sum_{k=1}^{K} \tilde{z}_k \ln \pi_k \\
= \sum_{k=1}^{K} \tilde{z}_k \ln \left( \frac{\exp(a_k + w_k)}{\sum_{l=1}^{K} \exp(a_l + w_l)} \right) \\
= \sum_{k=1}^{K} \tilde{z}_k (a_k + w_k) - \sum_{k=1}^{K} \tilde{z}_k \ln \left( \sum_{l=1}^{K} \exp(a_l + w_l) \right) \\
= \sum_{k=1}^{K} \tilde{z}_k (a_k + w_k) - \ln \left( \sum_{l=1}^{K} \exp(a_l + w_l) \right) \sum_{k=1}^{K} \tilde{z}_k,
\]

(4.46)

where I have made use of the log-sum-exp trick.\(^3\) The term \(\sum_{k=1}^{K} \tilde{z}_k\) would normally sum to one at each point in the brain. However, the algorithm uses a ‘pushforward’ of the responsibilities, which means that voxels of the warped responsibilities, used for re-computing the template, may not sum to one. The pushforward operation is the adjoint (multiplication by the transpose of a big sparse matrix) of warping the template to the individual. This is the reason why I keep the sum over the pushed responsibilities in the derivations. Note that such adjoint operations are also performed in the super-resolution model of Chapter 2, by the projection matrices and their transposes.

Differentiating (4.46) once gives:

\[
\frac{\partial \mathcal{L}_m}{\partial a_k} = \tilde{z}_k - \frac{\exp(a_k + w_k)}{\sum_{l=1}^{K} \exp(a_l + w_l)} \sum_{k=1}^{K} \tilde{z}_k \\
= \tilde{z}_k - \pi_k \sum_{k=1}^{K} \tilde{z}_k.
\]

(4.47)

Differentiating a second time with respect to \(k\) gives:

\[
\frac{\partial^2 \mathcal{L}_m}{\partial a_k^2} = - \left( \frac{\exp(a_k + w_k) \left( \sum_{l=1}^{K} \exp(a_l + w_l) \right) - \exp(a_k + w_k)^2}{\left( \sum_{l=1}^{K} \exp(a_l + w_l) \right)^2} \right) \sum_{k=1}^{K} \tilde{z}_k \\
= (\pi_k^2 - \pi_k) \sum_{k=1}^{K} \tilde{z}_k.
\]

(4.48)

\(^3\)The log-sum-exp trick can be used to avoid numerical underflows (or overflows) when the log of the sum of exponentials is to be computed: \(\text{lse}(x_1, \ldots, x_N) = \ln \left( \sum_{n=1}^{N} \exp(x_n) \right)\) by instead computing: \(\text{lse}(x_1, \ldots, x_N) = x^* + \ln \left( \sum_{n=1}^{N} \exp(x_n - x^*) \right)\), where \(x^* = \max \{x_1, \ldots, x_N\}\).
Differentiating with respect to $j \neq k$ gives:

$$
\frac{\partial^2 L_m}{\partial a_j \partial a_k} = \left( \frac{\exp(a_k + w_k) \exp(a_j + w_j)}{\left( \sum_{l=1}^{K} \exp(a_l + w_l) \right)^2} \right) \sum_{k=1}^{K} \tilde{z}_k
$$

$$
= \pi_k \pi_j \sum_{k=1}^{K} \tilde{z}_k.
$$

(4.49)

Hence, the second derivative is given by:

$$
\frac{\partial^2 L_m}{\partial a_j \partial a_k} = \pi_k (\pi_j - \delta(j - k)) \sum_{k=1}^{K} \tilde{z}_k.
$$

(4.50)

The complete gradient and Hessian of $L_m$ can therefore be obtained by computing (4.47) and (4.50) for each voxel of a subject, and then summing these over all subjects. Finally, the gradient and Hessian of the prior term ($L_p$) is given by:

$$
g_p = -\Lambda a,
$$

(4.51)

$$
H_p = -\Lambda a.
$$

(4.52)

The template is initialised so as to represent a uniform probability distribution over tissues.

The generative model defined by (4.40) describes a stochastic process on a population-level. This model can be used for learning population-wide parameters, such as intensity hyper-parameters and a tissue template; but also for segmenting and normalising images on a per subject basis, when the population-level parameters have been already learned. However, there is still room for improvement. For example, the model does not take into account the fact that clinical neuroimaging data often has differing FOVs (see Figure 2.1); but simply masks voxels where not all modalities are observed. Could variable FOVs be inferred within the model (Section 4.4)? Furthermore, besides the template, there are limited spatial dependencies in the model. Could the introduction of a MRF type of distribution improve the segmentation accuracy (Section 4.5)? These two questions will be addressed in the next sections, starting with the differing FOVs. As a matter of fact, the method proposed for this purpose even allows for predicting entirely missing modalities from one, or a few, MR contrasts. This is known as image translation.
4.4 A Model for Medical Image Translation

Applications of medical image translation are numerous, and include e.g. harmonising data across scanners; synthesising CT images from MR images for PET attenuation correction (Burgos et al., 2013), or decrease the need for radiating a patient; or generalising machine learning techniques by transferring out-of-distribution input data to the domain of the model’s training data (Roy et al., 2010). My particular interests, as to why I want to be able to perform image translation within the framework of the generative model presented in this chapter, are:

- It could simplify the problem of multimodal image registration, by synthesising a modality that is easier to register, from an acquired modality that is more challenging (Cao et al., 2012).

- It would allow for cross-validating the generative model by masking out voxels from an image, predicting the values of these missing voxels, and then computing some metric between the held-out reference voxels and the inferred ones.

- It would allow for segmenting multi-channel data with differing FOVs; instead of, as is currently often done, segment only the voxels in which each channel have observed data. Differing FOVs is a common property of routine clinical MR images (see Figure 2.1).

Image translation can be loosely categorised as either optimisation- or learning-based. Optimisation-based methods rely only on the data at hand to optimise a mapping between modalities, and do not use training data. Examples include using non-parametric joint histograms (Kroon and Slump, 2009), estimating an intensity transformation during image registration (Guimond et al., 2001), and biophysical models (Wein et al., 2008). Learning-based methods use training data to learn the mapping, and can be applied to translating an unseen image from one domain into another. Some examples in this category use clustering (Hsu et al., 2013), random forests (Huynh et al., 2015), patch-matching (Iglesias et al., 2013a) and dictionaries (Roy et al., 2011). Learning-based methods based on various CNN
architectures are currently the most popular approach for this. Trained end-to-end, on either paired or unpaired training data (Chartsias et al., 2017; Nie et al., 2017; Wolterink et al., 2017), they show promising results at this task, although they can run the risk of hallucinating unwanted features (Cohen et al., 2018).

This section presents a more interpretable generative modelling approach to image translation. It could be classed as an optimisation-based approach, although it does use training data to learn priors that inform the optimisation of mappings (see Section 4.3). More specifically, it will be shown how the generative model for group-wise normalisation and segmentation of neuroimaging data in (4.40) can be extended to handle missing data. What enables inferring missing voxels are: (1) that population-level parameters, the template and intensity priors, have been learned from a population of subjects; and (2), that the generative model’s GMM component is extended to handle missing data. GMMs have already been shown to handle missing data (Ghahramani and Jordan, 1994). I here extended this methodology to its variational GMM representation. Without a variational GMM, I would not be able to incorporate the learned intensity model.

Fitting the proposed model, to various populations of medical images, allows me to predict, from a few MR contrasts, entirely missing modalities (e.g., non-acquired MR contrasts or CT images). I will furthermore show that the model can be trained on a fairly small number of subjects. The proposed model is validated on three clinically relevant scenarios. Results appear promising and show that a principled, probabilistic model of the relationship between multi-channel signal intensities can be used to infer missing modalities – both MR contrasts and CT images.

4.4.1 Methods

The prediction of one modality from another is here cast as a joint intensity modelling problem. The workhorse of the proposed method is the model defined by (4.40). As exact marginalisation is intractable, a variational approximation is again used.
4.4. A Model for Medical Image Translation

Variational GMM with Missing Data

For a single subject, the fully observed GMM from Section 4.3 is:

\[
p(X, Z, \{ \mu_k, \Lambda_k \}_{k=1}^K) = p(X \mid Z, \{ \mu_k, \Lambda_k \}_{k=1}^K) p(Z) \prod_{k=1}^K p(\mu_k, \Lambda_k).
\]  

(4.53)

Now, it will instead be assumed that some modalities are missing in a voxel.\footnote{For example, a multi-channel MRI might have three contrasts: T1w, T2w and PDw. In one voxel, only the T1w intensity is observed. The T2w and PDw intensities are then assumed missing in that voxel. Note that different voxels can have different combinations of contrasts/modalities missing.}

The vector \( o \) indexes the observed modalities and the vector \( m \) the missing modalities. Therefore, the observed channels can be written as \( g = x_o \) and the missing channels as \( h = x_m \), where the voxel index \( n \) has been temporarily dropped for clarity. For a voxel in class \( k \), the marginal distribution of the observed channels can then be written as (Bishop, 2006):

\[
p(g \mid \mu_k, \Lambda_k, z_k = 1) = \mathcal{N}(g \mid \mu_{ko}, \Sigma_{koo}),
\]  

(4.54)

where

\[
\Sigma_{koo} = (\Lambda_{koo} - \Lambda_{kom}(\Lambda_{kmm})^{-1}\Lambda_{kmo})^{-1},
\]  

(4.55)

and the conditional distribution of the missing channels as:

\[
p(h \mid g, \mu_k, \Lambda_k, z_k = 1) = \mathcal{N}(h \mid \mu_{km} - (\Lambda_{kmm})^{-1}\Lambda_{kmo}(g - \mu_{ko}), (\Lambda_{kmm})^{-1}).
\]  

(4.56)

The set of all missing values in an image is written as \( H = \{ h_n \}_{n=1}^N \). The mean field approximation becomes:

\[
q(H, Z, \{ \mu_k, \Lambda_k \}_{k=1}^K) = \left[ \prod_{n=1}^N q(h_n \mid z_n) q(z_n) \right] \left[ \prod_{k=1}^K q(\mu_k, \Lambda_k) \right],
\]  

(4.57)

where \( q(h_n \mid z_n) = \prod_{k=1}^K \mathcal{N}(h_n \mid \tilde{h}_{nk}, \tilde{S}_{nk})^{\tilde{z}_{nk}} \). The marginal posterior over missing values is a GMM that can be obtained by marginalising the labels:

\[
q(h_n) = \sum_{k=1}^K \tilde{z}_{nk} \mathcal{N}(h_n \mid \tilde{h}_{nk}, \tilde{S}_{nk}).
\]  

(4.58)
Chapter 4. Improved Segmentation of Clinical Neuroimaging Data

Its expected value is:

$$\mathbb{E}[\mathbf{h}_n] = \sum_k \tilde{z}_{nk} \tilde{h}_{nk}.$$  \hfill (4.59)

This is the expression that is evaluated in order to predict missing voxels.

**ELBO**

The set of all observed values is written as $\mathcal{G} = \{\mathbf{g}_n\}_{n=1}^N$. The ELBO can then be written in two equivalent forms:

$$\mathcal{L} = \mathbb{E}[\ln p(\mathcal{G} | \mathbf{Z}, \boldsymbol{\mu}_{1 \ldots K}, \Lambda_{1 \ldots K})] - \sum_{n=1}^N \sum_{k=1}^K \mathbb{D}_{KL}(q_{nk} \parallel p_{nk}),$$  \hfill (4.60)

$$\mathcal{L} = \mathbb{E}[\ln p(\mathbf{X} | \mathbf{Z}, \boldsymbol{\mu}_{1 \ldots K}, \Lambda_{1 \ldots K})] - \sum_{n=1}^N \mathbb{E}_{z_n} \mathbb{D}_{KL}(q_{nk} \parallel p_{nk}).$$  \hfill (4.61)

The first form is used to optimise the posterior over tissue classes, while the second is used to optimise, in turn, the missing values and the GMM posterior parameters.

**Model Updates**

Optimising the ELBOs in (4.60) and (4.61) gives the subject-level posterior parameters as:

$$\tilde{z}_{nk} = \frac{\exp \left( \mathbb{E} \left[ \ln \mathcal{N}(\mathbf{g}_n | \boldsymbol{\mu}_k, \Lambda_k^{-1}) \right] + \ln \pi_{nk} \right)}{\sum_{j=1}^K \exp \left( \mathbb{E} \left[ \ln \mathcal{N}(\mathbf{g}_n | \boldsymbol{\mu}_j, \Lambda_j^{-1}) \right] + \ln \pi_{nj} \right)},$$  \hfill (4.62)

$$\tilde{b}_k = b_{0k} + \sum_{n=1}^N \tilde{z}_{nk},$$  \hfill (4.63)

$$\tilde{\mathbf{m}}_k = b_{0k} \mathbf{m}_{0k} + \sum_{n=1}^N \mathbb{E}[\tilde{z}_{nk} \mathbf{x}_n] \tilde{b}_k,$$  \hfill (4.64)

$$\tilde{v}_k = v_{0k} + \sum_{n=1}^N \tilde{z}_{nk},$$  \hfill (4.65)

$$\tilde{V}_k^{-1} = V_{0k}^{-1} + \sum_{n=1}^N \mathbb{E} [\tilde{z}_{nk} \mathbf{x}_n \mathbf{x}_n^T] + b_{0k} \mathbf{m}_{0k} \mathbf{m}_{0k}^T - \tilde{b}_k \tilde{\mathbf{m}}_k \tilde{\mathbf{m}}_k^T.$$  \hfill (4.66)

The update equations for the Gaussian parameters in the missing data case are similar to the fully observed case in (4.27) - (4.30), except that expectations are taken
about the data. These expectations are evaluated as:

\[
E[\tilde{z}_{nk|x_n}]_o = \tilde{z}_{nk}g_n,
\]

\[
E[\tilde{z}_{nk|x_n}]_m = \tilde{z}_{nk}h_{nk},
\]

\[
E[\tilde{z}_{nk|x_n}]_{oo} = \tilde{z}_{nk}g_n\tilde{g}_n,
\]

\[
E[\tilde{z}_{nk|x_n}]_{mm} = \tilde{z}_{nk}(h_{nk}h_{nk}^T + S_{nk}),
\]

\[
E[\tilde{z}_{nk|x_n}]_{om} = \tilde{z}_{nk}g_nh_{nk}^T,
\]

\[
E[\tilde{z}_{nk|x_n}]_{mo} = \tilde{z}_{nk}h_{nk}g_n^T,
\]

where

\[
h_{nk} = \bar{m}_{km} - \Lambda^{-1}_{mim}\Lambda_{mjo}(g_o - \mu_{ko}),
\]

\[
S_{nk} = \Lambda^{-1}_{mim},
\]

and \(\Lambda_k = \tilde{v}_k\tilde{V}_k\) is the posterior expected precision matrix of a given class. The mean \(h_{nk}\) and covariance \(S_{nk}\) gives access to the full Gaussian distribution of the missing voxels. This distribution could be used to sample missing data or for cross-validation procedures.

### 4.4.2 Evaluation

In this section the aim is to explore the translational (or inference) capability of the proposed model. This is achieved by conducting three experiments investigating: (1) inferring missing voxels of MRIs with differing FOV; (2) inferring entirely missing MRI contrasts; and (3), inferring CT scans from MRIs. The findings are quantified by computing the PSNR for an image channel \((c)\) given by (2.43). The PSNR is a metric that is commonly used in the medical image synthesis literature (Chartsias et al., 2017; Wolterink et al., 2017; Nie et al., 2017). To facilitate interpreting the results in the experiments of this section, I provide Figure 4.4 to give the reader an intuition for what different values of the PSNR implies in terms of image quality.

**MRI Contrast Translation**

This section evaluates translating between MR contrasts. The model is trained on 50 subjects from the IXI dataset (described in Section 1.6), which was acquired on
Figure 4.4: Demonstrating computing the PSNR and RMSE image quality metrics. A T1w MR image is corrupted with additive Gaussian noise, whose standard deviation (SD) is log-linearly spaced between 0 and $m_x$ (the maximum intensity in the image). For each value of SD the PSNR and RMSE is computed between the known reference and its corrupted version.

Figure 4.5: Learned model for MR image translation. Shown are template and expectations of the Gaussians drawn from the Gauss-Wishart priors, learned from 50 IXI subjects. Densities are plotted using their $3\sigma$ isocontours. This model is fit to a new subject, which allows for inferring missing voxels.
4.4. A Model for Medical Image Translation

Figure 4.6: Example of inferring MRIs with differing FOVs. An MR image with three channels (PDw, T1w and T2w) is observed. The PDw scan has full brain coverage, while the T1w and T2w scans have partial brain coverage (50% of voxels removed in each channel). From the observed data the values of the missing T1w and T2w voxels are inferred. The reference T1w and T2w scans are shown for comparison, as well as PSNR values.

three different MR scanners. Each IXI subject has three MR images: a T1w, T2w and PDw scan. Twelve ($K = 12$) mixture components are used, resulting in the model shown in Figure 4.5. Note that the template learned by the algorithm does not need to represent real tissues. Here, the model has been treated as a method of representing a probability density function, rather than as a way to do clustering. Any ‘meaningful’ clusters are incidental.

**Inferring MRIs with Variable FOV:** Doctors often acquire routine clinical MR scans of multiple contrasts. Commonly, these contrasts have differing FOV, meaning the brain coverage varies (cf. observed T1w and T2w images in Figure 4.6). This can be problematic for image segmentation routines as voxels with non-observed contrasts need to be discarded. The model should prevent this issue by inferring the values of these missing voxels. To test this, T1w, T2w and PDw scans of 50 unseen IXI subjects are used. All of the voxels are retained in the PDw image, while an increasing number of voxels are removed from the T1w and T2w images (25%, 50%, 75% and 100%). The missing voxels are then inferred with the trained model. An example can be seen in Figure 4.6. The mean PSNR values, computed between

---

5 This scenario is more realistic in a clinical context. The results would improve if data from only one scanners was used.

6 The model is trained on IXI subjects IXI[064–118], and tested on IXI[002–063].
Table 4.2: Results for inferring MR images with differing FOV (for 50 subjects). The PSNR is computed between known T2w and PDw references and inferred images, where an increasing percentage of the FOV has been removed. Results are shown as mean±std.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1w</td>
<td>42.1±1.6</td>
<td>36.3±1.3</td>
<td>31.1±1.3</td>
<td>28.9±1.2</td>
</tr>
<tr>
<td>T2w</td>
<td>40.7±2.1</td>
<td>34.4±2.0</td>
<td>30.4±1.8</td>
<td>27.6±1.6</td>
</tr>
</tbody>
</table>

Figure 4.7: Example of inferring non-acquired MR contrasts. An MR image with two channels (T1w and T2w) is observed. The PDw scan is missing, but inferred from the observed T1w and T2w scans. The reference PDw scan is shown for comparison, as well as the PSNR value.

For routine clinical MRI, it is rare that more than 50% of the FOV is missing. The results therefore suggest that the model does a good job at filling in images with variable FOV, which could be of value in segmenting hospital data.

**Inferring MR Contrasts**: Could the proposed model be used to infer an entirely missing MR contrast? An interesting application for this type of MRI translation could be for segmentation methods based on deep learning. A deep learning model that has been trained on MR images of a specific contrast can overfit to its training
Table 4.3: Results for inferring MR image contrasts (for 50 subjects). PSNR is computed for all different permutations of observed and missing contrasts. Results are shown as mean±std.

<table>
<thead>
<tr>
<th>Contrasts</th>
<th>Observed</th>
<th>Missing</th>
<th>T1w</th>
<th>PSNR T2w</th>
<th>PDw</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1w, T2w</td>
<td>PDw</td>
<td>T1w, T2w</td>
<td>28.9±1.5</td>
<td>28.5±1.1</td>
<td></td>
</tr>
<tr>
<td>T2w</td>
<td>T1w, PDw</td>
<td>T1w, T2w</td>
<td>28.0±1.2</td>
<td>27.6±1.6</td>
<td></td>
</tr>
<tr>
<td>PDw</td>
<td>T1w, T2w</td>
<td>PDw</td>
<td>29.2±1.4</td>
<td>28.1±1.5</td>
<td></td>
</tr>
</tbody>
</table>

data. If images could be simulated as to match the training data of the deep learning model, it might generalise better.

To test how well the model predict a missing contrast the same IXI subjects as in the previous experiment are used. For each subject, all combinations of contrasts are permuted over, set as either observed or missing. For example, only the T1w image can be observed while the T2w and PDw scans are inferred, or the T2w and PDw scans can be observed and T1w inferred (see Figure 4.7). The results from this experiment are shown in Table 4.3. These results imply that the T1w image is the most predictive, as the lowest PSNR value is obtained when this contrast is missing. The example inferred PDw image in Figure 4.7 looks realistic when compared to the known reference, although more noisy. Denoising the T1w and T2w images using the methods in Chapter 2 might have reduced this effect. The results in Table 4.3 are close to those previously reported in the literature, for the same task but a different dataset (Chartsias et al., 2017).

MRI to CT Translation

Accurately translating MRIs to CTs is interesting for numerous reasons, e.g., for removing the exposure to radiation that CT imaging involves, for attenuation correction in MR-PET imaging, or for focussed ultrasound therapy, which needs corrections for the acoustic properties of the skull (Marquet et al., 2006; Bretsztajn and Gedroyc, 2018). The proposed model should allow for this type of translation, by training it on subjects who have both MR and CT imaging. The intensity model is therefore trained anew – retaining the template learned from the IXI dataset –
on eight patients from the RIRE dataset (described in Section 1.6). Each patient in this dataset contains a number of imaging modalities. Here, only the patients with T1w and T2w MR scans (non-rectified), and CT images, are used. Note that the RIRE dataset is challenging to use due to the images having thick-slices, sometimes pathology, as well as requiring an initial co-registration (the dataset is part of a registration challenge and therefore purposefully misaligned). Each subject’s scans are registered using the co-registration routine of the SPM12 software, using the NMI cost function (recall, NMI was the best performing method in the multimodal study of Section 3.5.2).

To test the model’s ability to translate MRIs to CTs, eight unseen RIRE patients are used\(^7\). The trained model is fit to each subject’s T1w and T2w scans. The expected marginal posterior distribution over the missing CT image can then be computed. An example is shown in Figure 4.8. The mean±std PSNR between the inferred CT images and the known references is 25.5±1.2. Considered the intensity hyper-parameters were trained on only eight subjects, the results are satisf-

\(^7\)The model is trained on RIRE patients patient[102-109], and tested on patient[001-007,101].
4.4. A Model for Medical Image Translation

The examples images in Figure 4.8 suggest that the model does not capture a detailed enough distribution of bone. Additionally, the meninges does not appear in the inferred image, but is instead modelled as cerebrospinal fluid. Fitting not only the intensity hyper-parameters to the CT data, but also the template, could resolve these issues. More training data would also help.

4.4.3 Discussion

This section showed how a popular model for segmenting brain scans – a probabilistic forward model with a Gaussian mixture part – can be extended to infer missing data. For multimodal data, which is commonly found in a routine clinical context, this extension circumvents the need to model only voxels that are observed in all channels. It furthermore enables predicting one MR contrast from another, or even CTs from MRIs. The model gives reasonable results if trained on a small number of subjects, but further improvements are expected with access to more training data. Interestingly, image translation is just a ‘by-product’ of learning the parameters of a joint probability distribution that models missing voxels. The same model can also be used to segment, bias correct and spatially normalise brain scans.

The model requires setting the number of Gaussian mixture components \( K \) at the start of the training. If this number is set too low, then the simulated images will look unrealistic. Here, this issue was resolved by using a fairly large number of components, which was found empirically capturing a detailed enough model distribution. Uninformative mixture components can then be driven to zero, due to the Bayesian setting of the Gaussian mixture model, by making point estimates of the values of the global tissue proportions \( \omega \). This is known as automatic relevance determination \( (\text{Neal, 1995; Bishop, 2006}) \).

An interesting future work would be to compare the method proposed in this chapter to deep neural network models. Such models have shown excellent results on medical image translation, for example from MRI to CT \( (\text{Wolterink et al., 2017}) \), and are currently the most widely used for this task, as well as considered the state-of-the-art. Given that an implementation of a deep learning-based medical image translation model could be applied to the clinical-grade data in the RIRE dataset,
a comparison could be performed by simple hold-out training on this data. Another validation method could involve training the GMM based model on medical 2D data, such as the HAM1000 dataset\footnote{The HAM10000 dataset constitutes 10,015 dermatoscopic images of skin lesions, with lesion class labels. It is available from \url{https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW96T}} – because the proposed model is a general technique, applicable not only to brain scans. The large quantity of available training data would allow for a comparison on different amounts of images, to investigate if the generative model proposed here performs favourably on smaller amounts of data. This type of comparison was performed in Ashburner et al. (2019), and did then favour a more classical generative model, when less data was available.

Generative modelling approaches integrating multi-channel images, like the one presented here, should involve a component that relates signal across the various channels. The approach presented in this section involves a probabilistic model of the relationship between signal intensities over channels. An alternative approach would be to use the MTV prior, which ensures that ‘edges’ appear in similar locations across channels. The MTV prior was used in Chapter 2 to achieve super-resolution or denoising of medical images. An avenue of future work could therefore be to incorporate both of these components into a super-resolution method, to improve resolution of thick-sliced, hospital-grade MR scans (discussed further in Section 5.2.1). By combining, for example, axial thick-sliced T2-weighted images and sagittal thick-sliced T1w images of the same subjects. In this example, the T2w image could provide some of the missing T1w signal in the left-right direction, whereas the T1w image could fill in some of the missing T2w signal in the inferior-posterior direction. Of course, this strategy would need to be formulated properly, but this work aimed to show a proof of the concept that one of those components, a probabilistic model between channels, does a good job at filling in missing data in MR images.
4.5 A Convolutional Neural Network MRF

Up until this point, there are no spatial dependencies in the segmentation model of (4.40), as it factorises across voxels (it is independent and identically distributed). The template does provide some such context, via prior knowledge about the expected location of different tissues: $p(z_n) = \text{Cat}(z_n | \pi_n)$, still it assumes conditional independence between voxels. Misclassification rates could therefore be decreased by the introduction of some spatial distribution into the model.

MRFs are such distributions defined by an undirected graphical model (Li, 1994). MRFs can be introduced to model dependencies between pixels in (4.40), in a relatively tractable way, by assuming that interactions are restricted to a finite neighbourhood:

$$p(z_n) = \text{Cat}(z_n | \pi_n) p(z_n | \{z_j\}_{j \neq n})$$

(4.75)

$$p(z_n | \{z_j\}_{j \neq n}) = p(z_n | z_{N_n}),$$

(4.76)

where $N_n$ defines pixels whose cliques contain $z_n$, and where the subject index $s$ has been discarded for less cluttering (derivations remain the same only a summation over subjects needs to be introduced). Here, the aim is to introduce such an MRF into (4.40). The common assumption is made that the MRF neighbourhood is stationary, meaning that it is defined by relative positions with respect to $n$: $N_n = \{n + \delta | \delta \in \mathcal{N}\}$. It is furthermore assumed that this conditional likelihood factorises over the neighbours and that each factor is a categorical distribution:

$$p(z_n | z_{N_n}) = \prod_{\delta \in N_n} \prod_{k=1}^K \prod_{l=1}^K (W_{kl})^{z_{nk} z_{n+\delta,l}}.$$  

(4.77)

VB Learning of the MRF Parameters

To learn the parameters (or weights) $W \in \mathbb{R}^{K \times K}$ of the distribution in (4.77) the following expectation term of the ELBO can be optimised:

$$\sum_{n=1}^{N} \mathbb{E}_{q_{n}} [\ln p(z_n | z_{N_n})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{\delta \in N_n} \tilde{z}_{n+\delta,l} \ln W_{kl},$$

(4.78)
where $\tilde{z}_{nk}$ represents a responsibility (see (4.45)). The following constraint is introduced to enforce the columns of $W$ to sum to one:

$$\sum_{l=1}^{K} W_{kl} - 1 = 0, \quad k = 1, \ldots, K,$$

(4.79)

using Lagrange multipliers. This gives the objective function:

$$L_W = \sum_{n=1}^{N} \sum_{k=1}^{K} \tilde{z}_{nk} \sum_{l=1}^{K} \sum_{\delta \in N_n} \tilde{z}_{n+\delta,l} \ln W_{kl} - \sum_{k=1}^{K} \lambda_k \left( \sum_{l=1}^{K} W_{kl} - 1 \right).$$

(4.80)

Differentiating (4.80) with respect to an element in the weight matrix gives:

$$\frac{\partial L_W}{\partial W_{kl}} = \sum_{n=1}^{N} \tilde{z}_{nk} \sum_{\delta \in N_n} \frac{1}{W_{kl}} - \lambda_k.$$

(4.81)

Equating to zero gives:

$$\sum_{n=1}^{N} \tilde{z}_{nk} \sum_{\delta \in N_n} \frac{1}{W_{kl}} - \lambda_k = 0.$$

(4.82)

Solving for $W_{kl}$ gives:

$$W_{kl} = \frac{1}{\lambda_k} \sum_{n=1}^{N} \tilde{z}_{nk} \sum_{\delta \in N_n} \tilde{z}_{n+\delta,l}.$$

(4.83)

Substituting (4.83) into the constraint in (4.79) gives the value of a Lagrange multiplier as:

$$\lambda_k = \sum_{i=1}^{K} \sum_{n=1}^{N} \tilde{z}_{nk} \sum_{\delta \in N_n} \tilde{z}_{n+\delta,l}.$$

(4.84)

Hence, updating $W$ consists of computing (4.83) for each element of the matrix. Here, the MRF weights are fitted during the segmentation (on a subject-level). However, the responsibilities in (4.78) can be replaced with manual segmentations from a population of subjects (introduce the summation over $S$), to learn, possibly, more representative weights. The idea is that learning at the latent tissue level ($Z$) may generalise better than learning directly from the image intensities ($X$).

However, modelling and learning an MRF in such a way restricts it to small neighbourhoods and a (log) linear representation of the weight parameters. Next, a novel method for modelling an MRF using a deep learning approach will be proposed. This allows for a nonlinear parametrisation with a, possibly, larger neighbourhood, which can be efficiently learned using the backpropagation algorithm.
4.5. A Convolutional Neural Network MRF

Figure 4.9: T1w MR images from two different, publicly available, datasets: MICCAI2012 and MRBrainS18 (on which the method is evaluated). It is evident that learning from one of these populations, and subsequently testing on the other is very challenging. The intensities are different by an order of magnitude, the bias is stronger in the MRBrainS18 subject. Additionally, age related change and pathology can be clearly seen, such as differences in ventricle size and white matter hyper-intensities, which further complicates the learning problem.

Convolutional Neural Networks

Fairly recent advances in CNNs have provided a new method for very accurate (and fast) image segmentation ([Long et al., 2015](#)), circumventing the need to define and invert a potentially complex generative model. Discriminative CNNs learn a function that maps an input (e.g., an MRI) to an output (e.g., a segmentation) from training data, where the output is known. They typically contain many layers, which sequentially apply convolutions, pooling and nonlinear activation functions to the input data. Their parameters are optimised by propagating gradients backwards through the network (i.e., backpropagation). For medical imaging, the U-net architecture ([Ronneberger et al., 2015](#)) is the most popular and now forms the basis for most top performing entries in various medical imaging challenges aimed at segmenting, e.g., tumours, the whole brain or white matter hyper-intensities[9]. The more classical segmentation frameworks based on probabilistic models seem to have met their match.

However, challenges on medical image segmentation can be seen as lab exper-
iments and – as with new medical therapies – there is a large gap to get from bench to bedside. CNNs excel in the context of challenges, factorising the commonalities in an image population of training data, which generalise to new data from the same population. They can struggle, however, when faced with new data that contain unseen features (Dolz et al., 2017), e.g., a different contrast (see Figure 4.9). This scenario usually requires the model being trained anew, on that unseen image contrast. In fact, even without considering interindividual variability (age, brain shape, pathology, etc), a CNN-based segmentation software has yet to be presented that is agnostic to the great variability in MR data (Akkus et al., 2017). Lack of such software is largely due to the limited amount of labelled data available in medical imaging, which is a clear obstacle to their generalisability. Some methods have been developed to address this problem, e.g., based on intensity normalisation (Han and Fischl, 2007), transfer learning (Van Opbroek et al., 2015) or batch normalisation (Karani et al., 2018). Still, none of these methods are yet general enough to solve the task of segmenting across scanners and protocols. Recently, approaches based on realistic data augmentation have shown promising results (Jog and Fischl, 2018; Zhao et al., 2019).

In the next section, an approach is proposed to bridge between the classical, but robust, generative segmentation models (from Section 4.2 and 4.3) and more recent CNN based methods. The link is in the MRF term of (4.75), which is modelled using a CNN. This allows to parametrise the MRF by a more complex mathematical function than in the regular linear case, as well as cover a larger neighbourhood than a second-order one. The approach is validated on two publicly available datasets, acquired at different scanners and centres. Favourable results are shown when applying the model trained on one of these datasets to the other.

Related Work

Two fairly recent additions to the computer vision field (Zheng et al., 2015; Schwing and Urtasun, 2015), are closely related to the method presented in the subsequent section. The idea of both these papers is to cast the application and learning of a conditional random field (CRF) into a CNN framework. A CRF is a statistical mod-
4.5. A Convolutional Neural Network MRF

The method that directly defines the posterior distribution in (4.1) is the Convolutional Neural Network Multivariate Random Field (MRF). To compute the CRF, both papers apply a mean-field approximation, which they implement in the form of a CNN.

Learning of image priors have been performed in a variational framework (e.g., the optimisation problem in (2.17)). In Roth and Black (2009); Chen et al. (2014), this was achieved by modelling a MRF using the Product-of-Experts, a generic method for learning high dimensional probability distributions (Hinton, 1999), which allowed for learning the MRF filters from training data. Also CNNs have been used as higher-order priors in imaging inverse problems (e.g., Ulyanov et al. (2018)). Besides the prior term, additionally proximal operators, used for solving for the regularisation term in a variational setting, have been learned using deep neural networks (e.g., Meinhardt et al. (2017); Adler and Öktem (2018)). Both of these approaches, focusing on learned prior and proximal operator respectively, are powerful methods for solving inverse problems using data-driven models. For a recent, detailed review on this topic see Arridge et al. (2019).

In contrast to the papers discussed above, the interest in this work is to instead define the full generative model, and compared to Zheng et al. (2015); Schwing and Urtasun (2015), keeping the separation between likelihood and prior in (4.1). Modelling these two components separately allows for including expert knowledge and image-intensity independent prior information over the segmentation labels. It also integrates easily with existing mixture-model-based approaches. Furthermore, modelling the prior as an MRF, without data-dependency in the neighbourhood model, may help in generalising among different image populations. Finally, the model allows an arbitrarily complex MRF distribution to be defined, including, e.g., nonlinearities.

4.5.1 Methods

In this section the generative model defined by (4.1) is used to encode an MRF over the tissue labels (Z). It is shown that computing the MRF term is analogous to the mathematical operations performed by a CNN. Learning the MRF clique potentials can then be done using backpropagation. This representation allows for introducing
nonlinearities and an increasing complexity in the MRF neighbourhood.

Mean-Field Inference

For simplicity, it will from now on be assumed that there is no template term \( \text{Cat}(z_n | \pi_n) \), and that the likelihood \( p(X | Z) \) is given by some distribution \( p_k(x_n), \forall k, \forall n \). Only the posterior distribution over categorical labels \( (Z) \) will be inferred.

Recall, the approach is to search for an approximate posterior distribution that factorises across voxels:

\[
p(Z | X) \approx q(Z) = \prod_{n=1}^{N} q(z_n).
\]  
(4.85)

Then, by assuming that the current approximate posterior distribution for the neighbourhood voxels is:

\[
q(Z) = \prod_{j=1}^{J} \text{Cat}(z_j | z_j),
\]  
(4.86)

I can use VB to get the optimal updated distribution for factor \( n \) by taking the expected value of the joint model log-likelihood, with respect to all other variables:

\[
\ln q^*(z_n) = \sum_{k=1}^{K} z_{nk} \left( \ln p_k(x_n) + \sum_{\delta \in \mathcal{N}} \sum_{l=1}^{L} \tilde{z}_{n+\delta,l} \ln \pi_{k,l,\delta} \right) + \text{const}.
\]  
(4.87)

This distribution is again categorical with parameters:

\[
\tilde{r}_{nk} \propto \exp \left( \ln p_k(x_n) + \sum_{\delta \in \mathcal{N}} \sum_{l=1}^{L} \tilde{z}_{n+\delta,l} \ln \pi_{k,l,\delta} \right).
\]  
(4.88)

Implementation as a CNN

Under VB assumptions, posterior distributions should be updated one at a time, in turn. Taking advantage of the limited support of the neighbourhood, an efficient update scheme can be implemented by updating at once all pixels that do not share a neighbourhood\(^\text{10}\) Another scheme can be to update all pixels at once based on the previous state of the entire field. Drawing a parallel with linear systems, this is

\(^{10}\)When \( \mathcal{N} \) contains four second-order neighbours, this corresponds to a chequerboard update scheme.
4.5. A Convolutional Neural Network MRF

comparable to Jacobi’s method, while updating in turn is comparable to the Gauss-Siedel method.

In the Jacobi case, updating the labels’ expected values in (4.88) can be implemented as a convolution, an addition and a softmax operation – three basic layers of CNNs:

\[ \tilde{R} = f(\tilde{Z}) = \text{softmax}(C + W \ast \tilde{Z}). \]  

(4.89)

The matrix \( C \) contains the conditional log-likelihood terms (\( \ln p_k(x_i) \)). The convolution weights \( W \in \mathbb{R}^{K \times K \times |N|} \) are now equal to the log of the MRF weights (\( \ln \pi_{k,l,\delta} \)) and, very importantly, their centre is always zero. Such filters are here called MRF filters, and the combination of softmaxing and convolving an MRF layer. Their centres are zeroed to satisfy the VB solution in (4.87) — so that the expectations are computed with respect to the neighbouring voxels. The MRF weights are parameters of the approximate posterior distribution \( q^*(Z) \). Note that multiple mean-field updates can be implemented by making the MRF layer recurrent, where the output is also the input.

Now, it is assumed that there exist a set of true segmentations \( \{ \hat{Z}_s \}_{s=1}^S \), along with a set of approximate distributions with parameters \( \{ \tilde{Z}_s \}_{s=1}^S \). One may want to know the MRF parameters \( W \) that make the new posterior estimate \( q^*(Z) \) the most likely to have generated the true segmentations. This reduces to the optimisation problem:

\[ W^* = \arg\max_W \left\{ \sum_{s=1}^S \ln q^*(\hat{Z}_s | W) \right\} = \arg\max_W \left\{ \sum_{s=1}^S \sum_{n=1}^N \sum_{k=1}^K \hat{z}_{snk} \ln \tilde{z}_{snk} \right\}, \]  

(4.90)

which is a ML, or risk-minimisation, problem. Note that this objective function is the negative of what is commonly referred to as the categorical cross-entropy loss function in machine-learning. If the optimisation is performed by computing gradients from a subset of random samples, this is equivalent to optimising a CNN, with only one layer, by stochastic gradient descent.

Post-Processing MRFs

MRFs are sometimes used to post-process segmentations, rather than as an explicit prior in a generative model. In this case, the conditional data term is not known,
and the objective is slightly different: approximating a factorised label distribution $q(Z) = \prod_{n=1}^{N} q(z_n)$ that resembles the prior distribution $p(Z)$. This can be written as finding such distribution $q$ that minimises the Kullback-Leibler divergence with the prior $p$:

$$q^* = \arg\min_{q} KL(q \parallel p). \quad (4.91)$$

Again, assuming all other factors fixed with $q(z_j) = \text{Cat}(z_j | \tilde{z}_j)$, the optimal distribution for factor $i$ is obtained by taking the expected value of the prior log-likelihood:

$$\ln q^*(z_n) = \sum_{k=1}^{K} z_{nk} \left( \sum_{\delta \in \mathcal{N}} \sum_{l=1}^{K} \tilde{z}_{n+\delta, l} \ln \pi_{k,l} \right) + \text{const}, \quad (4.92)$$

which is equivalent to dropping the conditional term in the generative case. Equation (4.89) is then instead written as:

$$\tilde{R} = \text{softmax} \left( W \ast \tilde{Z} \right). \quad (4.93)$$

**Nonlinear MRF**

The conditional prior distribution $p(z_n | z_{\mathcal{N}_n})$ that defines the MRF can, in theory, be any strictly positive probability distribution. However, in practice, they are usually restricted to simple log-linear functions, which are easy to implement and efficient to compute. On the other hand, deep neural networks allow highly nonlinear functions to be implemented and computed efficiently. Therefore, a more complex layer is proposed, based on multiple MRF filters and nonlinear activation functions – a nonlinear MRF density. To ensure that a conditional probability is implemented, a constraint is introduced so that the input value of a voxel may not be used to compute its posterior density. Therefore, the first layer consists MRF filters that do not have a central weight, and subsequent layers are of size one to avoid reintroducing the centre value by deconvolution. The first layer is therefore proposed to be an MRF filter $W \in \mathbb{R}^{K \times F \times |\mathcal{N}|}$, where $F$ is the number of output features. Setting $F > K$ allows the information to be decoupled into more than the initial $K$ classes and may help to capture more complex interactions. This first convolutional layer is
4.5. A Convolutional Neural Network MRF

Figure 4.10: An illustration of the architecture of the MRF CNN. Outlined are the operations performed for learning to predict the centre of a segmented pixel. The nonlinearities are introduced by the ReLU activations. By setting the number of MRF layers to $K$ and keeping only the final softmax layer, the linear MRF model is obtained. The convolution kernel applied by the MRF filter is shown left of the segmentations, with its centre constrained to be zero.

followed by a ReLU activation function, 1D convolutions that keep the number of features untouched, and another ReLU activation function. This allows features to be combined together. A final 1D linear layer is used to recombine the information into $K$ classes, followed by a softmax. Figure 4.10 shows the proposed architecture.

Implementation and Training

In this work, the number of MRF layers is set to $F = 16$, three by three convolutions are used and leaky ReLU activation functions with $\alpha = 0.1$. The CNN is optimised using the Adam optimiser (Kingma and Ba, 2014). To reduce overfitting, the data is augmented in two ways: (1) by simple left-right reflection; and (2), by sampling warps from anatomically feasible affine transformations, followed by nearest neighbour interpolation. Realistic affine transformations are sampled by parametrising them by their 12 parameter Lie group (see the methods section of Chapter 3) and then learning their mean and covariance from a large number of subjects’ image headers.

4.5.2 Evaluation

This section aims to answer a series of questions: (1) does applying a linear MRF trained by backpropagation to the output segmentations of a generative model improve the segmentation accuracy? (2) does complexifying the MRF distribution using numerous filters and nonlinearities improve the segmentation accuracy compared to a linear MRF? (3) do the learned weights generalise to new data from an
Datasets and Preprocessing

The validation was performed on axial 2D slices extracted from the MICCAI2012 and MRBrainS18 datasets (described in Section 1.6). Within each dataset, all subjects were scanned on the same scanner and with the same sequences, whilst between datasets, the scanners and sequences differ (Figure 4.9). Both datasets have multiple labelled brain structures, such as cortical GM, cerebellum, ventricles, etc. These are combined as to obtain the same three labels for each subject: GM, WM and Other (1 − GM − WM), which were used as targets when training the model.

All T1w MR scans were segmented with the algorithm implemented in the SPM12 software, which is based on the generative model described in Ashburner and Friston (2005). In this model, the distribution over categorical labels is independent across voxels, non-stationary, and encoded by a probabilistic atlas deformed towards each subject. The algorithm generates soft segmentations, that is, parameters of the posterior categorical distribution over labels. The GM, WM and Other classes were pulled from these segmentations. Figure 4.11 shows the T1w image of one subject from each dataset, with its corresponding target labels and SPM12 segmentations.

Model Training and Evaluation

Two different models (or networks) were trained: a regular, second-order MRF (Lin); and a second-order nonlinear MRF (Net). Figure 4.10 explains the differences in architecture between the two. For each subject and each class, the ground-truth labels were used to compute Dice scores (defined by (2.44)) for the ML labels obtained using SPM12 and those obtained after application of the linear MRF and the nonlinear MRF. Statistical significance of the observed changes was tested using two-sided Welch’s t-tests between paired measures. Multiple comparisons were accounted for by applying the Bonferroni correction.

First, the learning ability of the networks was evaluated. To this end, a 10-fold

---

11 Besides disabling the final MRF clean-up, the default parameters of SPM12 were used.
cross validation was performed on the MICCAI2012 dataset, where groups of three images were tested using a model trained on the remaining 27 images. This yielded Dice scores for the entire MICCAI2012 dataset (shown in Figure 4.12a).

Next, the generalisability of the networks was evaluated. That is, what kind of accuracy can be obtained when the models are tested on images from an entirely new dataset. This was done by randomly selecting models trained on one of the MICCAI2012 folds and applying them to the images from the MRBrainS18 dataset. The results are shown in Figure 4.12b.

Results

The 10-fold cross validation results in Figure 4.12a show that the increase in Dice scores for both GM and WM is statistically significant after applying either of the two MRF CNN models. (Figure 4.11 shows the results for a randomly selected
Figure 4.12: Validation of the model on the MICCAI2012 (a) and MRBrainS18 (b) datasets. Dice scores were computed for known labels (GM and WM) and: SPM12 segmentations (SPM); and linear (Lin) and nonlinear (Net) 8-neighbour MRF applied to the SPM12 segmentations. Asterisks indicate statistical significance of paired t-tests after Bonferroni correction: 0.05 (*), 0.01 (**), 0.001 (***), 0.0001 (****).

MICCAI2012 subject). With a mean Dice of \{GM = 0.867, WM = 0.921\} for SPM12, and \{GM = 0.901, WM = 0.929\} and \{GM = 0.909, WM = 0.931\} after applying the linear and nonlinear MRF, respectively. The results imply that the classical generative approach of SPM12, which currently ranks in the top 50 on the MRBrainS13 challenge website\footnote{mrbrains13.isi.uu.nl/results.php}, could move up quite a few positions by application of the proposed model trained on the challenge data. As can be seen in Figure 4.12a, for one of the subjects, all models perform substantially worse. On closer inspection, this subject suffers from major white matter hyper-intensities. This abnormality is currently not handled well by the MRF CNN models, which obtain lower Dice scores than the initial SPM12 segmentations.

Figure 4.12b shows results when applying the models to data from a different centre, not part of the training data (Figure 4.11 shows the results for a randomly selected MRBrainS18 subject). Mean Dice scores are \{GM = 0.722, WM = 0.816\}
4.5. A Convolutional Neural Network MRF

for SPM12, and \{GM = 0.761, WM = 0.831\} and \{GM = 0.755, WM = 0.829\}
after application of the linear and nonlinear MRF, respectively. Application of the
MRF improves both GM and WM segmentations. The nonlinear MRF performs
slightly worse than the linear version. This result could be due to the nonlinear
model – which possesses many more parameters than the linear model – overfit-
ting to the training subjects of MICCAI2012. Additionally, the nonlinear MRF
may struggle with the MRBrainS18 subjects that have pathology (e.g. white matter
hyper-intensities). Still, the fact that Dice scores improve when applying the model
to new data shows that the proposed model can successfully improve segmenting
images from different MR imaging protocols.

4.5.3 Discussion

This section introduced an image segmentation method that combines the robust-
ness of a well-tuned generative model with some of the outstanding learning capa-
ibility of a CNN. The CNN encodes an MRF in the prior term over the unknown
labels. The method was evaluated on annotated MR images and it was shown that
a trained model can be deployed successfully on an unseen image population, with
very different characteristics from its training population. The hope is that the idea
presented in this chapter introduces, to the medical imaging community, a princi-
pled way of bringing together probabilistic modelling and deep learning.

In medical image analysis – where labelled training data is sparse and im-
ages can vary widely – generalisability across different image populations is one
of the most important properties of learning-based methods. However, achieving
this generalisability is made difficult by the limited amount of annotated data; the
datasets that were used in this section contained, in total, only 37 subjects. This
issue may be addressed by realistic, nonlinear data augmentation, which is able to
capture changes due to ageing and disease. Learning this variability in shape from a
large and diverse population could be a step in that direction [Balbastre et al., 2018;
Dalca et al., 2019a]. On the other hand – manual segmentations suffer from both
intra- and inter-operator variability, is it clinically meaningful to learn from these
very imperfect annotations? Could automatic segmentations prove more anatomi-
cally informative than manual ones (c.f., Figure 4.11)? Semi-supervised techniques, leveraging both labelled and unlabelled data, could be an option for making the method less dependent on annotations (see e.g., Roy et al. [2018]).

The architecture of the proposed MRF CNN was chosen with the idea of keeping the number of parameters low (to reduce overfitting), while still introducing a more complex neighbourhood than in more classical MRF models. However, the effect of different CNN architectures (e.g., activation functions, number of layers, filter size, etc.) was not evaluated here and remains to be studied. There is therefore a possibility of improved performance by design changes to the network. Such a change could be to hierarchically apply MRF filters of decreasing size, which could increase neighbourhood size without increased overfitting. Another potentially interesting idea would be to ‘plug in’ the MRF filters at the end of a segmentation network, such as a U-net, emulating MRF post-processing inside the network. Finally, future work intends to integrate the proposed model into a generative segmentation framework and then validate its performance by comparing it to other existing CNN-based segmentation models.

4.6 Examples of Learned Models

This section will, briefly, show three examples of fitting the population-level model, defined by the joint probability distribution in (4.40), to various neuroimaging data. A fully working, general and robust implementation, is still under development. This section should therefore be seen more as a proof of concept – that the model is able to normalise images and extract meaningful commonalities between subjects.

4.6.1 Predicting Fluid Intelligence

I participated, together with a team from UCL’s Centre for Medical Image Computing (CMIC), in the MICCAI 2019 ABCD Neurocognitive Prediction Challenge.[13] The challenge involved predicting fluid intelligence from T1w MRIs – about 8,500 subjects in total, aged 9-10 years. The data of 4,100 individuals was provided for training. The accuracy of each method was then assessed on its ability to predict

fluid intelligence scores from 4,400 individuals (whose actual scores was revealed after the challenge deadline). Out of 29 submissions, our team won the challenge (Mihalik et al., 2019).

In short, the method was based on fitting a three \((K = 3)\) class, unsupervised model to the 4,100 training subjects and then using the resulting normalised tissue segmentations (smoothed with a Gaussian kernel of FWHM 12 mm) to train a kernel ridge regression model. The model was then used to predict the fluid intelligence of the subjects in the testing population. The (softmaxed) template that was learned during model fitting is shown in Figure 4.13 and the prior intensity distribution in Figure 4.14. Figure 4.15 shown an example MRI from the training population.

4.6.2 Normalising Haemorrhagic Images

Normalising neuroimaging data with lesions, e.g., due to haemorrhagic stroke, is a very challenging task – it could even be considered an open problem (Sarmento et al., 2019). If accurate normalisation is achieved, a multitude of interesting analyses can be performed, which could lead to new insights into the diseased brain. Lesion-deficit mapping, discussed in Section 1.2, is one such analysis. I have therefore, together with Dr. Parashkev Nachev at UCL’s Institute of Neurology, been working on the task of spatially normalising haemorrhagic CT data, using an extension of the model in (4.40). This model can make use of labelled data in a semi-supervised setting and handle intensity normalisation across images in an elegant way. As it is currently work in progress, these extensions will not be discussed here.

However, an example fit of a nine \((K = 9)\) class supervised model will be shown: Figure 4.16 shows the tissue template, Figure 4.17 the intensity model and Figure 4.18 an example normalised CT scan. The model was fitted to data from three datasets: patient CT data, healthy IXI subjects and labelled images from MR-BrainS18. MR and CT images of in total 127 subjects were used. The labelled data included cortical GM, subcortical GM, WM, CSF and ventricles. The results are encouraging and will be part of future research from the group I am working in.
Figure 4.13: A $K = 3$ class tissue template learned from 4,100 T1w MR images part of the ABCD challenge. There is a GM and a WM class, and one class that absorbed everything else (background, CSF, etc.). The template voxels are 1.0 mm isotropic.
4.6. Examples of Learned Models

Figure 4.14: A six class intensity model learned from 4,100 T1w MR images part of the ABCD challenge. Here, there are six, not three classes, because each tissue is represented by two Gaussians.

Figure 4.15: Example ABCD T1w MR image. Shown is the native space MRI and its normalised counterpart.
Figure 4.16: A $K = 9$ class tissue template learned from, among other things, haemorrhagic CT data. There are two GM classes, one WM class, two CSF classes, one lesion class and three non-brain classes. The template voxels are 1.5 mm isotropic.
4.6. Examples of Learned Models

4.6.3 Modelling Brain and Spine Data

The potential impact of having general computational frameworks to deal with combined brain and spinal data is very promising. In fact, numerous studies have already shown that spine MRI may help in differential diagnosis and disease progression monitoring, as opposed to solely using brain MRI scans (Bot et al., 2004; Freund et al., 2016). The development of such a modelling framework was the aim of Blaiotta (2017), and I show here how the modifications suggested in this chapter allows for learning such a model.

Figure 4.20 shows a tissue template that was fitted to 200 subjects from the IXI dataset. Each subject from this dataset has a T1w, T2w and PDw image. The T1w image covers both the brain and spine (see Figure 4.19). Handling missing data, as described in Section 4.4, enables fitting a model that covers both the brain and the spinal cord. The model was learned in a fully unsupervised setting, with a
Figure 4.18: Example normalised CT image. Shown is the native space CT scan, its normalised counterpart, and the nine normalised tissue segmentations.
4.6. Examples of Learned Models

Figure 4.19: Example MR images of an IXI subject. Note how the T1w image covers the spine, while the PDw and T2w images do not.

Unsupervised tissue template built from 200 IXI subjects

Figure 4.20: A $K = 14$ class tissue template learned from 200 IXI subjects. Each subject had a T1w, T2w and PDw image. The template voxels are 1.5 mm isotropic.

A large number of classes ($K = 14$) and one Gaussian per tissue. Figure 4.21 shows 16 normalised T1w scans, and Figure 4.22 shows the arithmetic mean of all 200 subjects’ T1w scans.
Example of normalised IXI subjects

Figure 4.21: Example of 16 T1w IXI subjects, in spatially normalised space.
4.6. Examples of Learned Models

Figure 4.22: Arithmetic mean of 200 T1w IXI subjects, in spatially normalised space.
Chapter 5

Discussion

Contents

5.1 Contributions . . . . . . . . . . . . . . . . . . . . . . . . 187

5.2 Future Directions . . . . . . . . . . . . . . . . . . . . . . . 189

5.2.1 Unified Super-Resolution . . . . . . . . . . . . . . . . . 189

5.2.2 Improved MTV Registration . . . . . . . . . . . . . . . 191

5.2.3 Unified Generative Models . . . . . . . . . . . . . . . . 192

5.1 Contributions

Medical images contained in hospital databases exhibit a diverse range of patients and pathologies. Applying powerful machine learning methods to such large datasets could uncover population patterns that more closely map to underlying disease mechanisms and cognitive behaviour. In the age of ‘Big Data’, small scale research studies, performed in controlled environments, could only dream of approaching such results.

This incredible potential is unfortunately challenging to tap into, as most medical image analysis tools were developed for images with relatively high isotropic resolution. These tools do, in general, not work as well when applied to routine clinical data. This is due to hospital data having a huge variability, which makes learning representative distribution of such data very challenging. Developing methods
that facilitates applying medical image analysis tools to clinical data is therefore of great importance and could help in introducing automated predictive systems into clinical care.

Transforming routine clinical neuroimaging data into a more easily processed form – making it more amenable to further analysis – has been the aim of the work conducted in this thesis and the methods developed have contributed towards this aim. Although all methods were applied to brain scans, the proposed models should be general enough to find usage on other types of medical imaging data. This is a strength of the principled and interpretable techniques used throughout this work. An example of such a transition can be found in, e.g., Ridgway et al. (2018). In that paper, the utility of voxel-wise statistical analysis following nonlinear registration to a group-wise average atlas is investigated, for paediatric liver analysis. This is a technique that is very commonly used in neuroimaging studies.

Much of my work in this thesis has concerned joint modelling of images with different modalities. Chapter 2 proposed a generative model that involved a fairly conventional use of MTV as a prior for image super-resolution or denoising, allowing information in one modality to inform the image restoration in other modalities. This model was shown to successfully recover resolution in large quantities of hospital imaging data and could therefore be an important preprocessing step, prior to any form of analysis on routine clinical data.

Chapter 3 showed that the same model could be used to tackle intermodality registration problems, a niche that information theoretic approaches currently dominate. The proposed cost function is defined symmetrically, over all images at once, and is therefore more principled than approaches that consider one of the images to be either fixed or moving. It was shown to be robust to intensity variations due to bias fields, and did not fail in the presence of large misalignments. These are important properties if one is interested in processing large numbers of routine clinical data.

Chapter 4 concerned models of joint intensity that used a type of spatially encoded Gaussian mixture model – a well known model for segmenting and spatially
normalising brain scans. This model was extended to better enable processing clinical data. First, I presented a method that extends this model to handle images of differing FOV, which even allowed for predicting entirely missing modalities from one, or a few, MR contrasts. Second, I introduced a principled way of combining the strengths of such a classic generative model with the unprecedented discriminative capability of more recent deep learning techniques. Finally, I showed results from training the full generative model, presented in Chapter 3, on large datasets of publicly available data. A by-product of training this model were normalised segmentations, which can be of value as features in predictive models.

5.2 Future Directions

Many potential avenues for future research have already been covered briefly, in the discussion of each individual chapter (2.6, 3.6, 4.4.3 and 4.5.3). This section will therefore focus on more detailed descriptions of two particularly interesting possibilities: unifying super-resolution (Chapter 2) with segmentation (Chapter 4), and improving the MTV registration (Chapter 3).

5.2.1 Unified Super-Resolution

Unified models resolve circular dependencies between parameters by estimating them alternately, in an iterative fashion. This technique for parameter estimation usually provides better results than simple serial (or pipeline) applications of each component. Generative models enable, in an elegant way, such unified approaches (Ashburner and Friston, 2005; Pohl et al., 2006; Van Leemput et al., 2009; Iglesias et al., 2013b; Blaiotta et al., 2018).

Recall the generative super-resolution model presented in Chapter 2. The idea of this work was to improve image quality prior to further processing. Often, this processing involves segmenting a brain scan into different classes. The generative model of Chapter 4 is one such segmentation method. Keeping in mind the improved results often obtained using unified models, would it be possible to unify the super-resolution and segmentation models into one principled framework?
served, thick-sliced scans are realisations of unknown HR images, which in turn are samples from a categorical distribution over brain tissues. Fitting such a model to a subject’s images would circumvent the need for having scans of the same size, as the model inherently makes the assumption that they are not (via the projection matrices). Furthermore, the results of both super-resolution and segmentation could improve, as both processes would benefit from each other, in the form of having access to an estimate of the tissue distribution, decreased effects of partial voluming, etc.

However, the model in its current form does not easily lend itself to unifying the two approaches in such a way. This is due to the non-smoothness of the MTV prior. The resulting optimisation problem is currently solved using an ADMM algorithm, which deals with the non-smoothness by introducing a dual variable and then using a proximal operator. Using this method would make marginalising over the unknown HR images difficult, which is necessary for constructing a unified objective function using probabilistic modelling.

An alternative could be to instead use an approach based on reweighted least-squares (RLS; Daubechies et al. (2010); Bach et al. (2012); Grohs and Sprecher (2016)), making use of the bound:

$$\|z_n\|_2 = \min_{w_n > 0} \left\{ \frac{1}{2w_n} \|z_n\|_2^2 + \frac{w_n}{2} \right\},$$  

(5.1)

where $z_n = \text{vec}(Z[n,:])$, from Section 2.4.2. We could replace the MTV component of the log-joint distribution in (2.14), with the bound in (5.1). However, this would, in its current implementation, require an extension of the multigrid solver to handle non-stationary penalties.

Furthermore, instead of looking for a MAP solution (as is currently done), an approximate Gaussian posterior that factorises over voxels and channels could be obtained using VB. When used in combination with RLS, this posterior would be bounded by a Gaussian. Combining super-resolution with segmentation in this way would mean that commonalities across channels would be modelled by both edges and mutual information between intensities. This type of model could result in a very robust, joint super-resolution and segmentation framework.
When a population of images are available, commonalities across subjects could be learned as well by using a population-level model, such as the one discussed in Section 4.3. That model learned the intensity variation across subjects and their tissue distribution. An even more representative distribution of population shape and appearance could be included into the model by using principal components techniques (Balbastre et al., 2018; Ashburner et al., 2019). Such an extended, hierarchical Bayesian framework could make for a very powerful segmentation and normalisation framework, especially for smaller datasets (Ashburner et al., 2019), such as those commonly found in medical imaging applications. In fact, the approach could, as in Section 4.5, even be defined to include a deep learning model in the generative process.

5.2.2 Improved MTV Registration

The multimodal image registration cost function derived in Chapter 3 was optimised using Powell’s method. This method is a derivative-free optimisation, which makes it a good choice for certain low-dimensional problems, or for simple proof of concepts. It is however, in most settings, an inefficient optimisation scheme. A more efficient implementation would use a derivative-based algorithm. These algorithms use derivative information to find good search directions during the optimisation and are therefore more efficient at finding optima, for continuous-domain and smooth problems. It would therefore be of interest to derive the gradient and Hessian of the NMTV cost function in (3.12). This would allow me to use an efficient second-order optimisation method, such as Gauss-Newton, to reach the optimum of the cost function. An alternative to Gauss-Newton would be to formulate the registration within a variational Bayesian framework, using a Laplace approximation. Such approaches could be orders of magnitude faster than using Powell’s method.

Another interesting extension to NMTV would be to, instead of nulling the gradients for where there are missing values (as is currently done), marginalise them, in a probabilistic setting. This would involve making a Laplace approximation of the cost function, as the normalisation term of MTV is unknown, and then deriving the expectation $\mathbb{E}[\text{MTV}]$ with respect to the conditional distribution of missing
gradient magnitudes, given the observed gradient magnitudes. This should lead to a more robust method when images have non-overlapping FOVs, which very often is the case for routine clinical data.

Another interesting extension to the NMTV methodology is to define the registration within a generative model. Currently, the NMTV cost function does not rely on a generative model and therefore does not take into account the respective quality of the observed images (i.e., noise level). A more principled method would be to compute MAP estimates of both the true signal and the transformation parameters:

$$\arg\min_{\mu, R} \left\{ \sum_{c=1}^{C} \ln \mathcal{N}(f_c | A_c \mu_c, \sigma^2_c) + \text{MTV}(\mu) \right\},$$  \hspace{1cm} (5.2)

where $A_c$ is a contrast-specific projection matrix that depends on the transformation matrix $(R)$ and on the image’s voxel size, and $\mu_c$ is the reconstructed noise-free image, from which the observed image has been generated. The model in (5.2) is now – just as in the previous section – an extension of the super-resolution model in (2.14). In this setting, also the rigid alignment of the unknown HR image is being optimised over. This could allow for improved reconstructions as edges becomes more aligned. Furthermore, encoding the transformation matrix inside of the projection matrix allows me to quite easily derive both gradient and Hessian of (5.2), with respect to the rigid parameters.

However, what drives the registration is then solely the observed image ($f_c$) and the corresponding, current estimate of its true signal ($\mu_c$). I implemented this approach and found that convergence is far too slow for a feasible algorithm. Parametrising it instead as $f_c(y) = A_c \mu_c(\xi_{q_c}(x))$, as in Chapter 3, will invoke the MTV term in the optimisation. However, as has already been discussed, it requires performing a slow Powell based model fit at each evaluation of the cost function. This is one more reason to implement a derivative-based method for optimising NMTV, or to explore methods for Bayesian optimisation.

5.2.3 Unified Generative Models

Finally, having discussed unifying super-resolution and segmentation; and then the same idea, but for super-resolution and multimodal registration, it should now be
clear that all these steps could be performed within the same generative model. Although such an approach should, in principle, provide improved results for all three parts, it can be (1) difficult to derive and (2) slow to update the model parameters. That is, if a PhD student requires years just to understand the model it gives him/her no time to develop the algorithm further. And if optimising the model takes days, development is hampered even further.

However, just as in the case with deep learning booming largely due to efficient implementations of automatic differentiation on powerful graphics processing units (GPUs); similar implementations might be needed for developing probabilistic models in medical image analysis. There are software packages implementing such frameworks \cite{Bishop_2003, Kulkarni_2015, Salvatier_2016}. These software may eventually allow for automatic inference on generative models of volumetric medical data, requiring a user to only provide a model structure (i.e., a Bayesian network). Methods based on principled, probabilistic forward models could then, quite possibly, find a wide usage and even compete with deep learning.

Of course, the operations performed in the forward and backward pass of a neural network are very well-suited to be parallelised on the GPU. However, the same is true for many of the operations performed by the models in this thesis. For example, applying the forward and adjoint operator of the super-resolution model can be conceptualised as a (sparse) matrix multiplication, which can be efficiently parallelised. This is true also for all the voxel-wise operations of the segmentation model as well. Porting the models presented in this thesis to the GPU is therefore an interesting avenue of future work. It would furthermore improve the time of processing large databases of clinical images, possible by orders of magnitude.


Shu-Hui Hsu, Yue Cao, Ke Huang, Mary Feng, and James M Balter. Investigation of a method for generating synthetic CT models from MRI scans of the head and neck for radiation therapy. *Physics in Medicine & Biology*, 58(23):8419, 2013.


Ronald R Peeters, Pierre Kornprobst, Mila Nikolova, Stefan Sunaert, Thierry Vieville, Grégoire Malandain, Rachid Deriche, Olivier Faugeras, Michael Ng,


Karsten Tabelow, Evelyne Balteau, John Ashburner, Martina F Callaghan, Bogdan Draganski, Gunther Helms, Ferath Kherif, Tobias Leutritz, Antoine Lutti,


William M Wells III, W Eric L Grimson, Ron Kikinis, and Ferenc A Jolesz. Adapt-


Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional ran-