

This electronic thesis or dissertation has been downloaded from the King's Research Portal at <https://kclpure.kcl.ac.uk/portal/>



Longitudinal neuroimaging features for discriminating early neurodegeneration

Aksman, Leon

Awarding institution:
King's College London

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT



Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. <https://creativecommons.org/licenses/by-nc-nd/4.0/>

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Longitudinal neuroimaging features for discriminating early neurodegeneration

Leon M Aksman

A thesis submitted to King's College London in partial
fulfilment of the requirements for the degree of Doctor of
Philosophy (PhD)

Department of Neuroimaging

Centre for Neuroimaging Sciences

Institute of Psychiatry, Psychology and Neuroscience

King's College London

To Grace, Clara and Henry

Acknowledgements

I am very grateful to Janssen Pharmaceuticals for generously funding this PhD and to my supervisors, Dr. David Lythgoe and Dr. Andre Marquand, for guiding me through the whole of it. They have been very patient with a complete neuroimaging outsider who took a long time to figure things out. Andre has been a great teacher and a friend who has heavily influenced my research while allowing me to choose my own path. Dave has been unwaveringly helpful and supportive throughout.

The years of my PhD would have been much harder and more lonely had it not been for those around me. Those unfortunate enough to have shared an office with me deserve first mention (in alphabetical order): Anna, Anthony, Luis, Natalie and (just squeaked in) Parker. I also would like to thank several other people at the IoPPN, past and present, for their friendship and kindness over the years: Rich Joules, Derek, Nuria and Muriel. Of course, many others in the Neuroimaging department have made my time at King's very enjoyable. It is a credit to Steve that the department has a friendly and vibrant atmosphere.

I am grateful to my parents for investing so much into all of their children. Particularly to my mother for blindly believing that we can achieve anything we attempt and my father for trying to convey his love of mathematics and science to his often unreceptive son. I would not have been able to do the work I do without his influence.

Finally, my wife Grace deserves my greatest gratitude. She bore much of the burden of supporting our new family over these years and in particular during the write up of this thesis. I have dedicated this thesis to her and our two children.

Abstract

Longitudinal designs are widely used in medical studies as a means of observing within-subject changes over time in groups of subjects, thereby aiming to improve sensitivity for detecting disease effects. Paralleling an increased use of such studies in neuroimaging has been the adoption of pattern recognition algorithms for making individualized predictions of disease. However, at present few pattern recognition methods exist to make full use of neuroimaging data that have been collected longitudinally, with most methods relying instead on cross-sectional style analysis. In this thesis we develop a feature construction method that uses longitudinal high dimensional data to improve the predictive performance of pattern recognition algorithms when classifying early neurodegeneration. Our method can be applied to data from a wide range of longitudinal study designs and permits an arbitrary number of time-points per subject. We apply the method to two problems: discriminating subjects with mild cognitive impairment (MCI) from healthy controls and discriminating subjects at risk for Parkinson's disease from healthy controls. We show substantial improvements in predictive accuracy relative to cross-sectional classifiers for discriminating disease subjects from healthy controls on the basis of structural magnetic resonance (MR) images. In addition, our method allows for the transfer of longitudinal information from one set of subjects to make disease predictions in another set of subjects. The proposed methodology is simple and, as a feature construction technique, flexible with respect to the choice of classifier, imaging modality and image registration algorithm.

Contents

Acknowledgements.....	3
Abstract.....	4
Contents	5
Figures	9
Tables.....	10
Chapter 1 Introduction	11
1.1 Overview	11
1.2 Alzheimer’s Disease	13
1.3 Parkinson’s Disease	14
1.4 Thesis Outline	15
Chapter 2 Literature Review	18
2.1 Overview	18
2.2 Pattern recognition in neuroimaging: an introduction and brief history	18
2.3 Diagnostic criteria	21
2.3.1 Alzheimer’s Disease	21
2.3.2 Parkinson’s Disease	24
2.4 MCI and AD discrimination	25
2.4.1 Structural MRI	26
2.4.2 Diffusion MRI.....	28
2.4.3 Resting State fMRI.....	29
2.4.4 PET	30
2.4.5 Multimodal.....	33
2.4.6 Longitudinal.....	35
2.5 PD discrimination	37
2.5.1 Structural MRI	38
2.5.2 Diffusion MRI.....	40
2.5.3 123I-Ioflupane SPECT.....	41
2.5.4 Transcranial sonography (TCS)	42
2.5.5 Multimodal and longitudinal.....	42
2.6 Pitfalls and validation	43
2.7 Conclusions.....	45
Chapter 3 Datasets	47
3.1 Introduction.....	47
3.2 Study Designs	48

3.3 Study Data.....	51
3.4 Discussion.....	53
Chapter 4: Preliminary Methodology	56
4.1 Overview.....	56
4.2 The Data Analysis Pipeline.....	56
4.3 Image Preprocessing	59
4.3.1 Cross-sectional Jacobian determinant features	60
4.3.2 Longitudinal Jacobian determinant features	60
4.3.3 Cross-sectional versus longitudinal Jacobian determinants	61
4.3.4 Whole-brain masks	62
4.4 Classification.....	64
4.4.1 Linear Kernels.....	65
4.4.2 Support Vector Machines.....	65
4.4.3 Cross-validation	75
4.5 Performance Evaluation.....	76
4.6 Significance Testing.....	78
4.7 Model Interpretation	79
4.7.1 Weight maps	79
4.7.2 Forward maps.....	79
4.7.3 t-stat Maps.....	81
Chapter 5 Preliminary Analysis	82
5.1 Overview.....	82
5.2 Materials and Methods.....	83
5.2.1 Datasets	83
5.2.2 Group Demographics	85
5.2.3 Image Preprocessing	88
5.2.4 Classification Algorithms	89
5.2.5 Classifier Performance Evaluation and Significance Testing	92
5.3 Results.....	93
5.4 Discussion	96
Chapter 6 Novel Methodology.....	100
6.1 Overview	100
6.2 Basic longitudinal trajectory model	101
6.3 Coefficient matrices in the general case.....	104
6.4 Coefficient matrices in the special case: two samples per subject.....	107
6.5 Dimensionality Reduction	108

6.6 Eigenslopes	112
6.7 Forming the longitudinal subspace	113
6.8 Within-set prediction problem	114
6.9 Information transferring prediction problems	115
6.10 Feature Formation	116
6.11 Kernel Formation	119
6.12 Nested Cross-Validation	120
Chapter 7 Application to Early Dementia	121
7.1 Introduction	121
7.2 Material and Methods	121
7.2.1 Datasets	121
7.2.2 Data Acquisition and Quality Control	122
7.2.3 Image Preprocessing	122
7.2.4 Classification	124
7.2.5 Performance Evaluation and Significance Testing	125
7.3 Results	126
7.3.1 Balanced Within-Set Prediction (HNRS dataset)	127
7.3.2 Balanced Information Transferring Prediction (HNRS dataset)	129
7.3.3 Balanced Within-Set Prediction (OASIS dataset)	130
7.3.4 Unbalanced Within-Set Prediction (OASIS dataset)	131
7.3.5 Longitudinal Information Transfer across datasets	133
7.3.6 Visualizing and comparing eigenspaces across datasets	134
7.3.7 Statistical comparison of methods	136
7.3.8 Discriminative and Forward Brain Maps	138
7.4 Discussion	141
Chapter 8 Application to Early Parkinson's Disease	147
8.1 Introduction	147
8.2 Materials and Methods	148
8.2.1 Study Description	148
8.2.2 Image Acquisition	148
8.2.3 Image Preprocessing and Quality Control	148
8.2.4 Image Registration	152
8.2.5 Group Demographics	154
8.2.6 Analysis Methodology	156
8.3 Results	157
8.3.1 Whole-brain and skeletonized features	158

8.3.2 Statistical Comparison of Feature Construction Factors.....	158
8.3.3 Discriminative and Forward Brain Maps	162
8.4 Discussion	167
Chapter 9 Conclusions	169
9.1 Summary	169
9.2 Novel contribution of this thesis	173
9.3 Future work.....	174
Appendices.....	177
Appendix A. Computation of principal components	177
Appendix B. Effect of SVM C-parameter variation	178
Appendix C. Extended dataset of the PMPP study	179
Appendix D. Quality of FA maps registration and thresholding	180
Appendix E. OASIS dataset demographic and clinical information.....	185
Appendix F. Method implementation	185
References.....	186

Figures

Figure 2-1 Hypothetical biomarker model of Alzheimer's disease progression.....	24
Figure 3-1 Design of the PMPP study	50
Figure 4-1 The data analysis pipeline	58
Figure 4-2 The procedure performed at each cross-validation fold.	59
Figure 4-3 A linearly separable problem with two hyperplanes	68
Figure 4-4 Confusion matrix for a two-class classification problem	78
Figure 5-1 Location of scalp electrodes in EEG data	85
Figure 5-2 Confusion matrix for the general case.....	93
Figure 5-3 Confusion matrix for SMLR classifier of EEG data.	96
Figure 6-1 Fitting two different polynomial functions to data.....	102
Figure 6-2 The principal components of a set of fifty two-dimensional samples.....	109
Figure 6-3 <i>Within-Set Prediction</i> and <i>Information Transferring Prediction</i> problems	115
Figure 6-4 A simple example of a longitudinally projected 3D vector.....	118
Figure 6-5 A CV fold implementing LM-PCA, <i>Balanced Within-Set Prediction</i> problem.	118
Figure 6-6 A CV fold implementing LM-PCA, <i>Balanced Information Transferring Prediction</i> problem	119
Figure 7-1 Histogram of number of cross-validation folds with given explained variance percentage	128
Figure 7-2 Visualization of eigenslopes.....	136
Figure 7-3 Pairwise correlations and angles between eigenslopes	136
Figure 7-5 Comparison of maps for <i>Balanced Within-Set Prediction</i> problem	141
Figure 8-1 Gradient direction specific artefacts.....	149
Figure 8-2 A uniformly noisy image.....	149
Figure 8-3 Examples of problematic FA maps.	151
Figure 8-4 Maximum subject motion across the three groups.	151
Figure 8-5 A comparison of the masks	154
Figure 8-6 Breakdown of number of subjects per group.	155
Figure 8-7 Whole-brain maps, PD vs controls.....	163
Figure 8-8 Skeletonized maps, PD vs controls	164
Figure 8-9 Whole-brain maps, HR _{PD} vs controls	165
Figure 8-10 Skeletonized maps, HR _{PD} subjects vs controls.....	166
Figure A-1	179

Tables

Table 5-1 Demographics of structural MRI based discrimination, HNRS dataset	87
Table 5-2 Demographics of structural MRI based discrimination, PMPP dataset.....	87
Table 5-3 Demographics of EEG based discrimination, PMPP dataset	87
Table 5-4 HNRS Dataset, structural MRI based discrimination.....	95
Table 5-5 PMPP Dataset, structural MRI based discriminations.....	95
Table 5-6 PMPP Dataset, EEG based discriminations.....	96
Table 7-1 HNRS Dataset, <i>Balanced Within-Set Prediction</i> discriminations	128
Table 7-2 HNRS Dataset, <i>Balanced Information Transferring Prediction</i> discriminations	130
Table 7-3 OASIS Dataset, <i>Balanced Within-Set Prediction</i> discriminations.	131
Table 7-4 OASIS Dataset, <i>Unbalanced Within-Set Prediction</i> discriminations.....	133
Table 7-5 Longitudinal information transfer across datasets.	134
Table 8-1 Demographics of analysis subjects at final (two year) follow-up time-point.	156
Table 8-2 Motion corrected whole-brain features, results	160
Table 8-3 Motion uncorrected whole-brain features, results	160
Table 8-4 Motion corrected TBSS features, results.....	161
Table 8-5 Motion uncorrected TBSS features, results.....	161
Table A-1 Acquisition details for available MRI sequences, PMPP dataset	180
Table A-2 Demographic and clinical information, OASIS dataset.....	185

Chapter 1 Introduction

1.1 Overview

The aim of this thesis is to develop new analysis methods to improve the automated diagnosis of early neurodegeneration using neuroimaging biomarkers. Early and accurate diagnosis, aided by methods that better detect subtle within-subject changes using longitudinal information, would allow for interventions that prolong the presymptomatic and prodromal phases of neurodegeneration, delaying the onset of clinical symptoms. In Parkinson's disease, for example, by the time a clinical diagnosis is made based on motor symptoms over half of the dopaminergic neurons in the substantia nigra portion of the brainstem have typically been lost (Fearnley and Lees, 1991; Ross et al., 2004). Identifying preclinical PD would help in recruiting patients for clinical trials of disease-modifying therapies and give such therapies the best chance of having a lasting positive effect (Olanow and Obeso, 2012).

Neuroimaging has been identified as a promising source of quantitative biomarkers¹ of neurodegeneration (Frisoni et al., 2010). At present, however, these and other biomarkers do not play a primary role in the diagnosis of the most prevalent neurodegenerative diseases (Albert et al., 2011; McKhann et al., 2011; Miller and O'Callaghan, 2015). There are several key reasons for this: (i) a simple set of criteria that can be widely administered within clinics is preferred; (ii) related to this point, biomarkers are expensive and are often add marginal diagnostic value and (iii) biomarkers have not been fully validated for clinical use due to the heterogeneity of diseases, particularly in their early stage. As a result, Jack et al., (2011), as part of a workgroup on Alzheimer's disease diagnosis, states that further work is needed to standardise, interpret and uniformly assess the biomarkers of neurodegeneration.

Automated pattern recognition based methods have the potential to address some of these issues by making systematic and objective diagnoses that do not rely on qualitative assessments. Such methods have been increasingly applied to

¹ A biomarker is an accurate and reproducible measure of a medical or biological state (Strimbu and Tavel, 2010).

neuroimaging data, particularly structural MRI, to build models that are able to make individualized disease predictions (Klöppel et al., 2012a). Furthermore, multivariate pattern recognition methods are able to aggregate information across multiple variables to make disease predictions with higher sensitivity and specificity than univariate methods (Sabuncu and Konukoglu, 2014). Importantly, though, automated methods still rely on ‘gold standard’ information (i.e. class labels or severity measures) that, at present, is provided by expert humans. This information may be error prone, especially when accurate diagnosis is challenging, such as in the most early and uncertain stages of disease. While there are algorithmic approaches to dealing with erroneous diagnostic labels such as probabilistic models that provide a measure of confidence with their prediction (e.g. Marquand et al., (2010)) or methods that directly account for so-called label noise (e.g. Natarajan et al., (2013)), automated methods themselves can improve the gold standard over time by identifying which spatial and/or temporal patterns within data do or do not predict disease, thereby refining the diagnostic criteria.

This thesis focusses on using longitudinal information in pattern recognition as a means of achieving our aim. The cardinal feature of longitudinal study designs is that they involve taking measurements from the same set of subjects over multiple points in time. The primary importance of such information is that it measures within-subject changes without the confounding effects of between-subject differences. Longitudinal study designs, i.e. those that aim to collect longitudinal information, are therefore a natural approach to studying diseases marked by change over time. Neurodegeneration is one such class of diseases, defined as the progressive loss of neuronal structure or function over time. There are many neurodegenerative diseases, including Alzheimer’s disease, Parkinson’s disease, Huntington’s disease and amyotrophic lateral sclerosis (ALS). We will focus on the two most prevalent neurodegenerative diseases, namely Alzheimer’s disease and Parkinson’s disease.

In the following sections we briefly describe both diseases and highlight the need for biomarkers that can aid in the early diagnosis of each. Despite the clearly recognised value of longitudinal data in studying such progressive diseases there is a paucity of methods that have been proposed to make use of this information to form biomarkers, and even fewer that are based on pattern recognition.

1.2 Alzheimer's Disease

Alzheimer's Disease (AD) is the most prevalent neurodegenerative disease worldwide. AD is the most common cause of dementia, accounting for 60-70% of all cases. Dementia is a syndrome defined as cognitive dysfunction that is sufficient to cause some impairment in activities of daily living, and can include a wide variety of symptoms such as memory loss, cognitive and language problems, behavioural changes and an impaired ability to perform daily activities. Dementia is a major cause of disability among older people and has a huge physical, social and economic impact on the individuals affected as well as their family and caregivers. According to the World Health Organization (WHO), dementia affects 47.5 million people globally, with 7.7 million new cases per year². The prevalences of AD and dementia in those aged 65 or over have been estimated to be 4.4% and 6.4% respectively (Lobo et al., 2000). With so many older people affected, dementia has a large impact on society as a whole, with one study estimating the direct and indirect costs of patient care to be between \$80 and \$100 billion per year in the US (Meek et al., 1998). The combination of high prevalence in older people and the ageing of populations worldwide have led to the prediction that the number of people living with AD will increase to 100 million in 2050 (Brookmeyer et al., 2007).

Much is known about the effects of AD on the brain, which include the deposition of amyloid plaques and neurofibrillary tangles (NFTs), neuronal and synaptic losses and inflammation (Serrano-Pozo et al., 2011). The causes of AD, in contrast, are not fully understood. They are thought to be a combination of genetic, lifestyle and environmental factors. Furthermore, despite decades of basic research and drug development, there is no treatment that can halt or reverse the pathology of AD. Early diagnosis is recommended for cost-effective care with treatments that may mitigate symptoms of the disease for some period of time (Winblad et al., 2016). It has been estimated that a modest one year delay in the onset and progression of AD would result in nearly 9.2 million fewer cases of the disease in 2050 (Brookmeyer et al., 2007). Towards this end, a great deal of research has focussed on characterising the early stages of AD and dementia. Researchers have identified and tried to refine

² <http://www.who.int/mediacentre/factsheets/fs362/en/> (accessed February 1st, 2017)

a diagnostic entity known as mild cognitive impairment (MCI), the symptomatic prodromal stage between healthy cognition and AD.

The diagnosis of MCI and AD is based on clinical criteria, using information that can be collected in a clinical setting. The criteria include both subjective evidence of a change in cognition within the patient and an objective assessment of cognitive impairment. The diagnostic criteria for MCI acknowledge the growing importance of biomarkers, including those based on neuroimaging, however, as we have noted, their role is presently secondary in making a diagnosis (Albert et al., 2011).

1.3 Parkinson's Disease

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease, affecting an estimated 1% of individuals over the age of 60 (de Lau and Breteler, 2006). In ninety percent of cases PD is idiopathic (i.e. of unknown cause). In the remaining ten percent of cases its cause is known: genetic causes explain roughly 4-5% of cases while secondary Parkinsonism (e.g. drug induced or caused by other diseases or illnesses) accounts for the remainder of cases. PD is characterized by a loss of dopaminergic neurons in the pars compacta of the substantia nigra (SN), a small structure located in the midbrain portion of the brainstem. The typical symptoms of PD, related to dopamine deficiency, include resting tremors, rigidity and postural instabilities. For these reasons, PD has traditionally been viewed as a motor disorder, although it is now considered to be a systemic disease, with non-motor symptoms that can manifest years before overt motor signs (Miller and O'Callaghan, 2015). These pre-motor symptoms include hyposmia (i.e. loss of smell), rapid eye movement (REM) sleep behaviour disorder, constipation and depression. Mid to late stage non-motor symptoms include cognitive decline, compulsive behaviour and hallucinations. There is currently no cure for PD, though treatments such as levodopa can alleviate symptoms, allowing good quality of life while the disease progresses (Chaudhuri et al., 2016).

Current diagnostic procedures typically identify PD at a relatively late stage in the disease where substantial numbers of SN neurons have already degenerated

and led to detectable motor symptoms. It is widely recognized in the research community that the severe loss of SN neurons that has occurred at the time of diagnosis likely limits treatment options and that PD needs to be diagnosed sooner, in its prodromal, premotor symptom phase. However, progress towards this goal has been limited by a lack of validated biomarkers. Miller and O’Callaghan, (2015) unequivocally state that there are no biomarkers at present that can predict the onset of PD nor any that can make a definitive diagnosis of PD. They identify the need for biomarkers that can identify early PD: namely the prodromal, preclinical or premotor stages. The need for early stage biomarkers is clear, as they would serve to identify PD before significant degeneration has occurred, presumably when future therapies would be most effective at halting the progress of the disease (Lang, 2011).

1.4 Thesis Outline

Part of the challenge in finding such biomarkers is that early stages of these diseases are marked by subtle pathological changes in the brain’s structure and function. The clinical diagnostic criteria for MCI we have outlined already take into account subjective changes in cognition within the patient. Such measures are important as there may be a large amount of variability in clinical measures across individuals or even within a single individual depending on the particular day. It seems natural that biomarkers should similarly use longitudinal information to capitalise on within-subject changes over longer periods of time, thereby increasing the sensitivity and/or specificity of diagnosis. It is commonly said in neurology that ‘time is the best diagnostician’ (Weinberger and Goldberg, 2014).

This thesis has a straightforward structure: it is built to present and apply a pattern recognition based method that makes use of longitudinal information and apply it to discriminate the early stages of AD and PD. Chapter 2 reviews the pattern recognition based disease discrimination studies in AD and PD that have used neuroimaging data. It puts emphasis on both early disease discrimination as well as two particular forms of problems structure, namely longitudinal and multi-modal information³. We point out that to date few studies have successfully capitalized on

³ We do not focus further on multi-modal information in this thesis, although it is a promising avenue of research given the biomarker progression models we discuss in Chapter 2.

longitudinal information, owing to a lack of pattern recognition based methods that directly make use of it, setting up the methodology we develop in Chapter 6.

Chapters 3 and 4 describe the data and preliminary methods used in subsequent chapters. Chapter 3 describes data from two studies, one focussing on MCI, the other on early PD. Both contain longitudinal neuroimaging data for disease subjects and healthy matched controls, allowing discriminations of early disease subjects from controls in both cases. Chapter 4 introduces the pattern recognition based data analysis pipeline we need for such discriminations. It describes the image preprocessing, cross-validation based classification and classifier performance evaluation, along with significance testing of classifier performance and model interpretation.

Chapter 5 puts the data and methodology of the previous two chapters to use, presenting preliminary analysis performed early in the course of the PhD. We compare cross-sectional and longitudinal biomarkers (referred to as “features” in pattern recognition) that can be extracted from structural MRI using the image preprocessing procedures described in the previous chapter. We use these structural MRI based features to discriminate both MCI subjects from controls and in pairwise discriminations between the three study groups of the Parkinson’s disease study we describe in Chapter 3. The analysis presented mostly serves to illustrate the difficulty both types of features have in discriminating early neurodegeneration⁴. Consequently it motivates the development of better methods of using longitudinal information in discrimination settings.

Chapter 6 describes our approach to combining longitudinal information with cross-sectional information to form features that are well suited for discriminating neurodegeneration. As a feature construction method (see Section 4.4) it is independent of the imaging modality, image preprocessing and subsequent pattern recognition algorithm it is applied to. This allows us to apply it to both structural MRI and diffusion MRI in later chapters. Furthermore we also demonstrate that the method can handle an arbitrary number of measurements per subject from both

⁴ However, we should point out that Chapter 5 also presents a promising EEG result that should not be overlooked.

balanced and unbalanced longitudinal designs and can be applied to new cohorts which may have limited longitudinal follow-up.

Chapter 7 is the first of two chapters that apply our method, in this case to discriminate MCI subjects from controls using the available MCI longitudinal neuroimaging dataset. To further validate the method, we also discriminate early dementia subjects from controls using a second, publicly available longitudinal dataset. In all cases we use structural MRI based features, as in the preliminary analysis of Chapter 5. Across the various discriminations we perform, we are able to showcase the method’s flexibility across various study designs, number of samples per subject and follow-up periods between samples. We also demonstrate the method’s ability to transfer longitudinal information from one set of subjects to make disease prediction in another set of subjects who may only have cross-sectional information available.

Chapter 8 applies the method to pairwise discriminations of the three study groups of the Parkinson’s disease dataset using longitudinal diffusion MRI based features. In doing so, we demonstrate that the method can be applied in a completely different disease discrimination problem, imaging modality and preprocessing procedure than in the previous chapter. We further evaluate the effect of several common choices for diffusion tensor imaging (DTI) based features on classifier performance.

Chapter 9 summarizes the conclusions of the thesis, highlights its novel contributions and suggests some promising directions for future work.

Chapter 2 Literature Review

2.1 Overview

In this chapter we review the current state of neuroimaging-based discrimination of Alzheimer's disease (AD) and Parkinson's disease (PD) using pattern recognition methods. After a brief introduction and background to pattern recognition in neuroimaging, we discuss the progress made in discriminating both prodromal AD and AD itself as well as early PD.

The literature on diagnostic approaches to both diseases is substantial and this review does not attempt to cover the wide variety of other types of biomarkers that have been investigated, ranging from genetics, blood and cerebrospinal fluid (CSF). Instead we focus on imaging based studies and multi-modal studies that have included imaging and attempt to understand how imaging fits within the diagnostic criteria of these diseases. Within dementia discrimination studies we focus on those that have used structural, diffusion and functional MRI as well as PET. Within Parkinson's discrimination studies we focus on studies using transcranial sonography (TCS), structural and diffusion MRI and ^{123}I -Ioflupane SPECT (i.e. DAT-SPECT). We also discuss multi-modal and longitudinal pattern recognition studies in both diseases, highlighting their future potential and present limitations.

An important point to note is that many of the relevant studies to date have been based on small samples and may report discrimination accuracies that may suffer from overfitting (Hawkins, 2004). In light of this, when possible, accuracies from larger sample studies will be reported.

2.2 Pattern recognition in neuroimaging: an introduction and brief history

Traditionally, neuroimaging studies have employed mass-univariate analytical techniques: the general linear model (GLM) in particular has been the dominant paradigm in the analysis of neuroimaging data for several decades (Friston et al., 1994). These approaches make group-level inferences separately for each brain voxel and have contributed significantly to our understanding of the brain's structure

and function. More recently, techniques from pattern recognition and machine learning have emerged as complimentary approaches to mass-univariate analysis. Pattern recognition and machine learning aim to “make predictions as accurately as possible and to understand the behaviour of learning algorithms” (Rasmussen, 2006). The distinction between the two terms (i.e. ‘pattern recognition’ and ‘machine learning’) is somewhat unclear. According to Bishop (2007), “Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. However, these activities can be viewed as two facets of the same field, and together they have undergone substantial development over the past ten years.” We will mostly use the terms “pattern recognition” or “pattern analysis” interchangeably here as they emphasize our focus on finding interpretable patterns within data in addition to the goal of making predictions on unseen data to assess the generalizability of a model.

Bishop, (2007) provides comprehensive coverage of pattern recognition and machine learning, covering both supervised and unsupervised learning. Supervised pattern recognition methods aim to predict class labels (i.e. classification) or continuous measures (i.e. regression) from the pattern within high-dimensional data and lend themselves naturally to making predictions at the level of individual subjects. They have proved to be particularly effective for clinical applications, where they have been used to predict disease state in a number of disorders (Klöppel et al., 2012b); Orrù et al., 2012; Wolfers et al., 2015; Arbabshirani et al., 2016). Such multivariate techniques offer increased sensitivity to spatially distributed effects that might, on their own, not be statistically significant.

Among the first studies that used pattern analysis in neuroimaging were those of Lautrup et al., (1994) and Mørch et al., (1997) who applied these methods to functional neuroimaging, in particular positron emission tomography (PET) and fMRI data. Haxby, (2012) recounts the motivation and methodology of another early and influential pattern analysis study: his 2001 paper on the representation of faces and objects in the ventral temporal cortex (Haxby et al., 2001). The study used multivariate pattern analysis to show that faces and objects are represented as patterns of neural activity in widely distributed and overlapping portions of the ventral temporal cortex rather than as distinct regions dedicated to representing particular faces or objects. A few years later came one of the earliest uses of what is

now the most common approach in neuroimaging, the support vector machine (SVM) by Cox and Savoy, (2003), who, in a similar study to Haxby's, trained a classifier to determine which of ten different objects a subject was looking at during an fMRI session using the imaging time series from various sets of voxels (whole brain or feature selected voxels) as inputs to the classifier. Haxby notes that interest in applying pattern analysis, sometimes referred to as multivariate pattern analysis (MVPA), in neuroscience increased with the discovery that these methods could be used to decode cognitive states (Haynes and Rees, 2005; Kamitani and Tong, 2005).

The earliest applications of pattern recognition to clinically oriented neuroimaging seem to have occurred largely independently of the neuroscientific neuroimaging studies described above. Among the earliest neuroimage based diagnosis studies were those of Kippenhan et al., (1992) building on early work by Kippenhan and Nagel, (1990) and separately by deFigueiredo et al., (1995). Both studies discriminated Alzheimer's (as well as vascular dementia in the latter study) using neural networks trained on region of interest (ROI) based features derived from PET or SPECT imaging. Roughly around the time of Cox and Savoy's study, Lao et al., (2004) used the SVM to separately discriminate gender and age groups using voxel based morphometry (VBM) based features derived from structural MRI. The idea of VBM is simple and intuitive: apply spatial transformations to three-dimensional (3D) structural images of subjects' brains to bring them into a common space where voxel-wise analysis can be performed across subjects. The common space is referred to as a stereotactic space in neurosurgery and neurology: a 3D frame of reference, defined by a reference template, to which individual subjects are warped.

The systematic and automated image registration techniques and analysis framework of VBM have had an enormous impact on neuroimaging, giving researchers an alternative to manually labelled ROI based analysis (Ashburner and Friston, 2000; Ashburner and Friston, 2001). A number of studies have used VBM based features to make individualized diagnostic predictions of neurological and psychiatric diseases (Davatzikos et al., 2008; Davatzikos et al., 2008; Duchesnay et al., 2007; Fan et al., 2006; Fan et al., 2007; Klöppel et al., 2008; Klöppel et al., 2009; Koutsouleris et al., 2009; Lerch et al., 2008; Teipel et al., 2007a; Teipel et al., 2007b; Vemuri et al., 2008a; Vemuri et al., 2008b). Among these studies and in many since, the SVM has emerged as the most widely used pattern recognition algorithm for

making individual diagnostic predictions using neuroimaging data (Orrù et al., 2012a; Arbabshirani et al., 2016). The success of the SVM can be attributed to the algorithm's ability to make accurate predictions using high dimensional feature vectors and small sample sizes, making it well suited for typical neuroimaging studies. The availability of good open source implementations such as LIBSVM (Chang and Lin, 2011) have contributed to its popularity. Orrù et al., (2012) provide a review of studies using the SVM and neuroimaging to discriminate neurological and psychiatric diseases. We discuss the SVM in detail in Section 4.4.2 and use it throughout this thesis. Note that although the SVM has been validated as an effective discriminative classifier, probabilistic algorithms such as the Gaussian Process Classifier (GPC) (Rasmussen, 2006) have shown competitive predictive performance while being able to quantify the uncertainty in their predictions (Marquand et al., 2010), an important property in clinical applications.

2.3 Diagnostic criteria

Before discussing the progress made in pattern recognition based diagnosis of AD and PD using neuroimaging we review the current diagnostic criteria in AD and mild cognitive impairment (MCI), the symptomatic pre-dementia phase of AD, as well as preclinical AD, its earliest asymptomatic phase. We discuss the clinical diagnostic criteria for PD which are based on motor symptoms as well as the premotor stages associated with early PD. An understanding of these clinical diagnostic criteria as well as the disease stages (and associated temporal biomarker model in AD) will help us frame the subsequent discussion of pattern recognition based biomarker studies using different modalities as well as the motivation of studies using multi-modal and longitudinal information.

2.3.1 Alzheimer's Disease

The criteria and guidelines for diagnosing AD were updated in 2011 following international workshops led by the Alzheimer's Association, the National Institute on Aging and the U.S.'s National Institutes of Health (Jack et al., 2011). The workshop identified three stages of Alzheimer's disease: dementia due to AD (McKhann et al., 2011), MCI due to AD (Albert et al., 2011) and a new stage termed

“preclinical Alzheimer’s disease” (Sperling et al., 2011). These criteria were a revision of the NINCDS-ADRDA criteria that had been in use since 1984 (McKhann et al., 1984). The 1984 criteria were used in clinical trials and clinical research for the diagnosis of probable AD with a sensitivity of 81% and specificity of 70% (Knopman et al., 2001). Consistent across both the old and new criteria are the use of clinical examinations and neuropsychological testing as the primary means of diagnosing probable AD (definite AD can only be diagnosed with histopathological evidence from a biopsy or autopsy). Probable AD must meet the criteria of all-cause dementia: interference with daily activities, a decline from previous functioning and cognitive impairment in two or more of the following domains: memory, reasoning, visuospatial abilities, language or personality. Additional criteria for AD dementia include a gradual onset of symptoms, a worsening of cognition over time and either amnesic presentations (learning or recall of information) or non-amnesic presentations (language, visuospatial or executive dysfunctions). The new criteria acknowledge the growing importance of biomarkers (both imaging and CSF) but do not recommend the use of such biomarkers as part of clinical practice due to a lack of uniform assessment and standardization.

The 2011 criteria defined a separate diagnostic entity known as MCI, the symptomatic predementia phase of AD. MCI has been a focus of intense study since the 1990s but was not part of the 1984 criteria. The MCI criteria put more emphasis on biomarkers than the AD criteria do, though in both cases biomarkers are not included in the *core clinical criteria*. The *core clinical criteria* for MCI are a change in cognition reported by the patient, an informant or a clinician plus objective evidence of impairment in one or more cognitive domains (listed above) with preservation of independence in functional abilities and no dementia. Unlike AD however, MCI does not cause interference in the activities of daily living. This somewhat vague distinction between AD and MCI means that differentiation must be made by a clinician on the basis of a description of daily activities from the patient and an informant. Importantly, unlike in AD, the diagnostic criteria note that evidence of longitudinal decline in cognition aids in the accuracy of diagnosis.

Biomarkers feature prominently in the *clinical research criteria* for MCI. A subject must meet the *core clinical criteria* to be diagnosed with MCI: biomarkers provide information as to the level of certainty in the diagnosis of “MCI due to AD”.

There are two categories of biomarkers in the criteria: those indicative of amyloid beta (“A β ”) deposition, namely CSF A β ₄₂ and amyloid PET imaging, and those indicative of neuronal injury, namely CSF tau/p-tau, MRI, FDG-PET and SPECT. A subject is said to have an intermediate likelihood of MCI due to AD in two situations: when there is a positive biomarker of A β and no available biomarker of neuronal injury or when there is a positive biomarker of neuronal injury and no available biomarker of A β , either because the unavailable biomarkers have not been or cannot be tested. A high likelihood of MCI due to AD corresponds to both A β and neuronal injury biomarkers being positive. MCI is not likely due to AD when both A β and neuronal injury biomarkers are negative; in such cases an alternative cause of MCI should be considered. The criteria do not cover the case of conflicting A β and neuronal injury biomarkers (one positive, one negative), noting that little is known about this situation.

The staging of preclinical AD, the earliest stage of AD identified by the workgroup, is based primarily on the hypothetical model of the temporal progression of biomarkers by Jack et al., (2010), shown in Figure 2.1. In this model, biomarkers of A β become abnormal first, followed by those of synaptic dysfunction, tau-mediated neuronal injury, brain structure, cognition and finally clinical function. A β accumulation, a necessary but not sufficient condition for MCI and AD, can occur a decade or more before clinical symptoms are detected. A data-driven model of biomarker changes, based on multi-modal data from the ADNI study, was developed by Young et al., (2014) and broadly confirmed this hypothetical model.

The three stages of preclinical AD identified mirror the hypothetical model: stage one is asymptomatic amyloidosis detected by CSF A β ₄₂ and amyloid PET, stage two is amyloidosis plus neurodegeneration detected by FDG-PET, (potentially) fMRI and structural MRI and stage three is amyloidosis, neurodegeneration and subtle cognitive decline detected by evidence of subtle change in cognitive function from subject’s baseline as well as poor performance on more challenging cognitive tests. Subjects in these stages do not yet meet the criteria for MCI. It is important to point out that preclinical AD is not a diagnosis. The preclinical AD stage is meant entirely for research purposes; for use in identifying study cohorts in longitudinal studies and clinical trials.

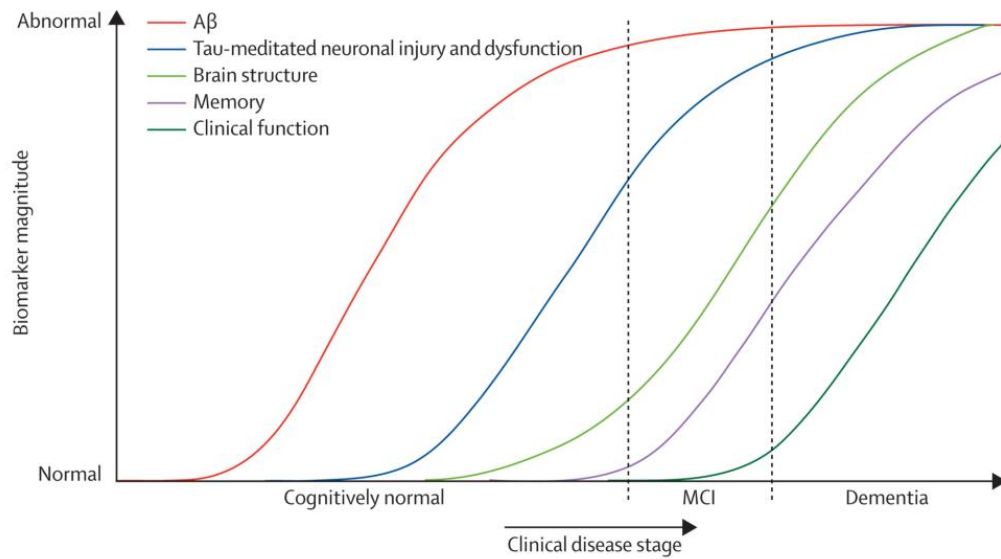


Figure 2-1 Hypothetical biomarker model of Alzheimer's disease progression. Aβ is identified by CSF Aβ₄₂ or PET amyloid imaging. Tau-mediated neuronal injury and dysfunction is identified by CSF tau or FDG-PET. Brain structure is measured by use of structural MRI. Aβ=β-amyloid. MCI=mild cognitive impairment.⁵

2.3.2 Parkinson's Disease

The diagnosis of PD is based on clinical assessment: taking a careful case history through questioning of the patient and family to ascertain which motor or pre-motor symptoms have emerged and in what sequence. A clinical examination is then conducted to detect signs of parkinsonism, a clinical syndrome with four cardinal features: bradykinesia (slowness of movement), resting tremor, rigidity and postural and gait impairment. The UK Parkinson's Disease Society Brain Bank clinical diagnostic criteria have been widely adopted as accurate means of making an objective diagnosis. This procedure consists of three steps: (i) a diagnosis of parkinsonian syndrome, consisting of bradykinesia and either rigidity, resting tremor or postural instability; (ii) a long list of exclusion criteria, such as a history of stroke, head injury, encephalitis or other neurologic or autonomic disorders; (iii) supportive positive criteria, consisting of three or more of the following: unilateral onset, resting tremor, progressive disorder, persistent asymmetry of symptoms, excellent response

⁵ Reproduced from Lancet Neurology, Vol 9, Issue 1, Jack et al., "Hypothetical model of biomarkers of the Alzheimer's pathological cascade", Pages 119 – 128, January 2010, with permission from Elsevier.

to levodopa, severe levodopa-induced chorea, levodopa response for 5 years or more, clinical course of 10 years or more, hyposmia or visual hallucinations (Massano and Bhatia, 2012).

There is a well-known temporal progression of symptoms in PD, starting from the prodromal, preclinical phase of pre-motor markers (loss of smell, REM sleep disorder, constipation, autonomic problems, depression) leading to early motor signs which eventually leads to the clinical phase of detectable motor signs that meet the diagnostic criteria (Braak et al., 2003; Lerche et al., 2014). Simultaneous with symptom progression is the gradual loss of nigrostriatal neurons, with 50% or more cells lost before motor signs appear (Hawkes, 2008). However at present there is no hypothetical model of the temporal progression of PD biomarkers, as there is in AD (Jack et al., 2010).

Studies such as the Parkinson Progression Marker Initiative (PPMI), a large multi-centre study, aim to develop a biomarker progression model by tracking de novo (i.e. newly diagnosed and untreated) Parkinson's subjects along with controls. The Progression Markers in the Premotor Phase (PMPP) study (Liepelt-Scarfone et al., 2013) aims to track a group of premotor phase subjects in addition to control and early PD groups to identify prodromal biomarkers. Reflecting the fact that the role of various biomarkers in discriminating PD disease stages has not been ascertained, both studies collect a variety of data, including blood samples, genetics, clinical assessments (motor, cognitive, autonomic, sleep, olfactory) as well as imaging (with MRI and SPECT in both cases). In Section 2.5 we review some of the pattern recognition based discrimination studies of PD using imaging, many of which have used the publicly available PPMI dataset. We use the PMPP neuroimaging data for pattern recognition based discrimination of PD, at risk subjects and controls later in this thesis.

2.4 MCI and AD discrimination

The following sections review studies discriminating MCI and AD on the basis of neuroimaging and how the approaches and research themes in such studies are related to the diagnostic criteria discussed in Section 2.3. Section 2.4.5 reviews

multi-modal studies, i.e. those that combine multiple modalities when making diagnostic predictions, while Section 2.4.6 discusses longitudinal studies that make predictions using within-subject changes over time.

2.4.1 Structural MRI

A recent extensive review of the literature identified over 500 papers that have applied pattern recognition and machine learning methods to neuroimage based neurological and psychiatric disease prediction (Arbabshirani et al., 2016). Across these studies, structural MRI is by far the most commonly used imaging modality and AD is by far the most commonly studied disease. The appeal of structural MRI as a biomarker of AD is that it is an independent and non-invasive measure of neuronal loss that provides complimentary information to a clinical diagnosis of AD, which is based on clinical measures and neuropsychological tests (Vemuri and Jack, 2010). Structural imaging biomarkers are able to detect brain atrophy that is correlated with neurofibrillary tangle deposition and neuropsychological deficits, making them suitable as markers of present AD state and the progression of AD (Frisoni et al., 2010).

In large samples, many studies have employed structural MRI to classify probable disease of Alzheimer's type (PDAT) subjects from healthy controls (HC) with high accuracy. Some example accuracies that are representative of the literature are 90% (Nho et al., 2010), 92% (Chincarini et al., 2011) and 92% (Duchesne et al., 2008). The features used in such studies have ranged from whole brain grey matter (GM) maps (Klöppel et al., 2008b), atlas based ROIs (Magnin et al., 2009), volume and thickness of cortex (Querbes et al., 2009; Desikan et al., 2009) and hippocampal shape (Gerardin et al., 2009) and volume (Chupin et al., 2009). Confirming these results, a direct and controlled comparison between radiologists and automated methods has shown that both performed comparably, achieving roughly 90% accuracy in both cases when discriminating subjects with sporadic AD from controls on the basis of structural MRI (Klöppel et al., 2008a).

Within the MCI diagnostic category, studies have tried to predict which subjects will progress to Alzheimer's disease. Subjects within the MCI group are

often further separated into those converting to Alzheimer's and those remaining stable over time. Costafreda et al., (2011) provide a summary of various approaches and their respective classification accuracies achieved on this particular problem. The studies listed used a variety of whole-brain or ROI-based features: using manually labelled hippocampal shape and volume (Ferrarini et al., 2009), multi-region semi-automatically labelled ROI's (McEvoy et al., 2009), VBM based whole brain features (Teipel et al., 2007b; Misra et al., 2009; Plant et al., 2010) and medial temporal ROI's (Duchesne et al., 2010). Costafreda's own study used automatically identified hippocampal shape, achieving 80% accuracy compared to 74% weighted accuracy across the other six studies. These studies, while promising for identifying MCI subjects at greater risk of AD who may benefit from pharmacological interventions, presuppose a subject can be accurately diagnosed with MCI. In addition, they are inherently retrospective in nature: further studies are needed that prospectively test the accuracy of prognostic biomarkers (Albert et al., 2011).

Studies have reported lower accuracy in discriminating MCI subjects from controls compared to discriminating AD subjects from controls. Part of the difficulty, in addition to the subtle pathology one would expect at an early disease stage, is the well-known heterogeneity of MCI. MCI includes within it subgroups of single-domain amnesic subjects (i.e. those with memory related impairment), single domain non-amnesic subjects (i.e. those with a non-memory impairment, such as executive control or language) and subjects with multi-domain impairment (Petersen, 2004a). Some studies have shown significant neuropsychological differences between these subgroups (Libon et al., 2010). The complex, heterogeneous nature of MCI should be kept in mind when reviewing (and planning) discrimination studies.

In a large sample, systematic study of data-driven feature selection versus whole-brain classification, Chu et al., (2012) achieved accuracies in the 65-70% range when separating MCI subjects from controls in contrast to accuracies of 80-85% when discriminating AD subjects from controls. Chu's study notes that domain-specific prior knowledge, such as the use of medial temporal lobe (MTL) based ROIs in discriminating AD, rather than data-driven feature selection, such as recursive feature elimination (RFE), is the means of achieving better classification performance. The hippocampus and, to a lesser extent, the parahippocampal region were noted as being consistently best at discriminating between groups, confirming

other studies' findings. The difficulty in accurately discriminating MCI using structural MRI derived biomarkers is consistent with Jack et al's hypothetical biomarker model, which posits the transition from normal to abnormal volumetric MRI based biomarker measures occurring during the MCI phase, with PET and CSF based biomarkers potentially transitioning sooner.

2.4.2 Diffusion MRI

In addition to neuronal loss that is detectable by structural MRI, AD is also characterized by WM changes, including loss of axons and myelin sheaths (Brun and Englund, 1986; Sjöbeck et al., 2005), WM atrophy (Hua et al., 2008; Migliaccio et al., 2012) and loss of structural integrity in WM pathways (Stebbins and Murphy, 2009). Few studies have used diffusion MRI in pattern recognition based discrimination of AD or MCI. Among the few large sample studies to date is that of Nir et al., (2015), who used features composed of fractional anisotropy (FA) and mean diffusivity (MD) measures from voxels along identified WM tracts. The study performed both statistical analysis of group differences along individual tracts and pattern recognition based classification using information from all tracts. Using the ADNI-2 dataset, the second phase of the ADNI project (Jack et al., 2015), the study investigated both group differences and the discrimination of AD subjects, "early" and "late" MCI subjects (a distinction introduced in ADNI-2) and healthy controls. They found significant group differences between AD subjects and controls along many tracts by the FA measure and all tracts by the MD measure. They also found significant group differences in MD between late MCI subjects and controls in five (out of twenty four) tracts but no group differences in FA in any tracts. No group differences were detected between early and late MCI subjects or between the overall MCI group and controls. The multivariate classifications were in line with the statistical analysis: FA and MD based features discriminated AD subjects from controls with 75-85% accuracy while MD based features discriminated late MCI subjects from controls with 68-79% accuracy, with no FA results presented for the latter problem. No results were presented for classifying early MCI subjects versus controls or all MCI subjects versus controls.

This study suggests that diffusion MRI has the potential to discriminate later stage MCI and AD but, at present, cannot discriminate early MCI. It is important to note that diffusion MRI is rapidly evolving: better imaging and analytic methods will inevitably lead to better features for pattern recognition. However, bearing the hypothetical biomarker model discussed in Section 2.3.1 in mind, biomarkers based on brain metabolism or neuronal function rather than brain structure may be intrinsically better suited to discriminating early dementia. Risacher and Saykin, (2013) provide a review of the role of various types of biomarkers in identifying different stages of AD.

2.4.3 Resting State fMRI

Studies using resting state fMRI (rs-fMRI) have shown differences in activity in the default mode network (DMN) (Raichle et al., 2001; Greicius et al., 2003) in AD subjects compared to controls (Greicius et al., 2004) and in MCI subjects relative to controls (Buckner et al., 2005). The DMN is a network of brain regions that includes the medial temporal, medial parietal and medial frontal lobes. In the last five years or so, several pattern recognition based studies have focussed on discriminating AD and MCI subjects using rs-fMRI based features. In their review Arbabshirani et al., (2016) identified six such studies, three of which discriminated Alzheimer's subjects from controls with high accuracy: 100% by Wu et al., (2013), 100% by Khazaee et al., (2015) and 92% by Zhou et al., (2010), though each study had forty or fewer samples and as such may suffer from overfitting. Challis et al., (2015) used functional connectivity based features on a slightly different discrimination problem, separating 17 AD subjects from 40 amnesic MCI subjects with 97% cross-validated accuracy on a validation data set and 90% accuracy on a completely withheld test set consisting of 10 AD subjects and 10 amnesic MCI (aMCI) subjects. Discriminating aMCI subjects from healthy controls was accomplished with up to 78% cross-validated accuracy on the validation set and 75% accuracy on the external test set of 10 aMCI subjects and 10 controls. Given these promising results, further validation of the diagnostic potential of rs-fMRI is needed in larger sample studies.

2.4.4 PET

In the biomarker progression model we have discussed, biomarkers of β -amyloid (i.e. $A\beta$) become abnormal first and can predate the onset of the clinical symptoms of AD by a decade or more. Amyloid PET imaging has increasingly been used to assess the $A\beta$ burden in those suspected of AD using positron emitting radionuclides (i.e. tracers) that specifically image $A\beta$ plaques. The earliest such tracer, 11C-Pittsburgh Compound-B (11C-PiB) (Klunk et al., 2004), is impractical for widespread clinical use, with a short half-life that necessitates an on-site cyclotron. Three tracers with longer half-lives have since been approved for clinical use by the US Food and Drug Administration: 18F-florbetapir, 18F-flutemetamol, and 18F-florbetaben.

Amyloid PET images are typically interpreted by visual inspection or using a specific quantitative measure: the standardised uptake value ratio (SUVR). An SUVR is a ratio of tracer uptake in disease-affected ROIs compared to the tracer uptake in an unaffected ROI. The use of SUVR together with visual evaluation has been shown to decrease inter-reader variability in 18F-florbetapir PET interpretation (Nayate et al., 2015). Recent studies have begun to use multivariate pattern recognition to automatically assess $A\beta$ abnormality using amyloid PET. Vandenberghe et al., (2013) used 18F-flutemetamol based images to classify scans as being amyloid positive (termed ‘Alzheimer-like’) or amyloid negative (termed ‘normal retention’). The study used whole-brain voxel based features and an SVM classifier to discriminate 25 amyloid positive scans from 24 amyloid negative scans with 100% accuracy, with training and testing class labels formed by expert visual readers.

Cattell et al., (2016) used a larger sample of subjects from across three different tracers on the same problem of classifying amyloid status. The study’s features, based on histograms of oriented three-dimensional gradients rather than voxel intensities, were independent of the choice of tracer or region of interest. As a demonstration of the robustness of their features, the study optimized the feature construction parameters on 18F-florbetapir images and applied the optimized parameters to classifying the amyloid status of a separate set of 131 18F-florbetapir

images, 209 ¹¹C-PiB images and 128 ¹⁸F-florbetaben images with over 96% accuracy across all tracers.

These studies show the promise of automated pattern recognition based approaches to classifying amyloid status. However, the role of amyloid positivity as a biomarker of AD has not been fully established. The guidelines toward defining the preclinical stage of AD (Sperling et al., 2011) note that there is a distinction between the pathophysiological process of AD (AD-P) and the clinical phase of AD (AD-C), with signs of AD-P emerging years before those of AD-C. The ability of AD-P biomarkers (such as those derived from amyloid PET) to predict the AD-C stage is not clear; although at present amyloid status is useful for stratifying subjects in therapeutic trials (O'Brien and Herholz, 2015).

In contrast to amyloid PET, ¹⁸F-2-fluoro-2-deoxyglucose (FDG) PET is able to measure cerebral metabolic rate of glucose (CMR_{glc}), which has been shown to correlate with clinical measures of dementia (Blass, 2002). FDG-PET has been used in AD research for over twenty years with studies showing CMR_{glc} deficits in parieto-temporal areas (Silverman et al., 2001), posterior cingulate cortex (PCC) (Minoshima et al., 1997) and the medial temporal lobe (MTL) (Mosconi, 2005) in AD subjects relative to controls and MCI subjects relative to controls (Buckner et al., 2005; Langbaum et al., 2009).

Given its long history in AD research (Mosconi et al., 2010; Herholz et al., 2013), there have since been surprisingly few studies performing solely FDG-PET based automated diagnosis of AD. There are several potential explanations for the present state of the literature: early studies using voxel-wise pattern recognition methods were primarily developed for and applied to MRI while PET is costly and involves exposure to ionizing radiation, making it less appealing as a clinical diagnostic tool. Following earlier work by Minoshima et al., (1995), who developed a univariate classifier using automatically registered FDG-PET images, Herholz et al., (2002) performed the first large sample automated diagnosis study, discriminating AD subjects from controls with 93% accuracy. Later studies such as that of Kerrouche et al., (2006) explored the use of multivariate data decomposition methods such as principal component analysis and canonical variates analysis to

discriminate between AD subjects, those with vascular dementia, and normal controls.

Recent discrimination studies have primarily used the ADNI dataset⁶: Hinrichs et al., (2009) discriminated AD subjects from controls with 84% accuracy using voxel based features while Haense et al., (2009) used the same method as in Herholz et al., (2002), a univariate classifier based on a sum of voxel-wise t-statistics, to achieve an 83% sensitivity and 78% specificity. Salas-Gonzalez et al., (2010) used feature selection and factor analysis with an SVM classifier to discriminate AD subjects from controls with up to 95% accuracy and MCI subjects from controls with up to 88% accuracy. Gray et al., (2012) used anatomically defined region-based features with an SVM classifier in pairwise discriminations of AD, progressive MCI, stable MCI and healthy controls. The study explored the use of both cross-sectional features (both baseline and follow-up time-points' images), longitudinal features (the within-subject change from baseline to follow-up), and features composed of the concatenation of both cross-sectional and longitudinal information. Classifiers trained using features based on the follow-up data point were shown to be better than classifiers that used baseline data. Follow-up features achieved 86% accuracy discriminating AD subjects from controls (AD/HC), 79% accuracy discriminating progressive MCI subjects from controls (pMCI/HC) and 62% accuracy discriminating progressive MCI subjects from stable MCI subjects (pMCI/sMCI). Longitudinal change-based features performed consistently worse in classifications than cross-sectional (baseline or follow-up) features. However, when concatenating follow-up and longitudinal features, the study found a consistent, though small, improvement in accuracy compared to purely follow-up features: 88% accuracy on AD/HC (a 2% improvement), 81% accuracy on pMCI/HC (a 2% improvement) and 63% accuracy on pMCI/sMCI (a 1% improvement).

Recent ADNI based studies have combined FDG-PET with structural MRI, measures of cerebrospinal fluid (CSF) and subjects' genetic information. These approaches are discussed in the following section on multi-modal classifiers.

⁶ ADNI is available at <http://adni.loni.usc.edu>.

2.4.5 Multimodal

Multimodal AD biomarkers provide complementary information that may lead to improved diagnostic predictions compared to any single modality (Fjell et al., 2010; Landau et al., 2010). Cavado et al., (2014) list three reasons for the use of multimodal biomarkers: (i) a number of pathologies are involved in AD whose causal interaction is not fully understood; (ii) the presence/detectability of these pathologies depends on disease stage; and (iii) the temporal course of pathologies is not linear, nor do the pathologies parallel each other. Their viewpoints are consistent with the biomarker progression model, which hypothesizes a sequence of temporally overlapping transitions from normal to abnormal states across different types of biomarkers.

Early multimodal pattern recognition based discrimination studies combined structural MRI with several other types of data: demographic and genetic information (Vemuri et al., 2008a) and CSF biomarkers (Vemuri et al., 2009). In the former study, the additional demographic and genetic information provided a small improvement over the structural MRI based biomarkers, which were found to be most discriminative of AD subjects from controls. In the latter study, structural MRI was found to be slightly more predictive of conversion from amnesic MCI to AD than CSF biomarkers. The study of Davatzikos et al., (2011) was consistent with Vemuri et al.'s study, showing structural MRI based features as most predictive of MCI conversion, with CSF biomarkers having comparable, though lower, predictive value. Overall, the combination of structural MRI and CSF was slightly better than structural MRI individually.

Both Hinrichs et al., (2011) and Zhang et al., (2011) used multi-kernel learning (MKL) based classifiers. MKL involves combining information from multiple between-subject similarity matrices (i.e. kernels), each formed using a single modality (Bach et al., 2004; Lanckriet et al., 2004). Hinrichs et al.'s study combined structural MRI, FDG-PET, CSF, genetics and cognitive scores to achieve a 92% accuracy in discriminating AD subjects from controls (AD/HC). Though the combination was numerically more accurate than any individual modality, simply using cognitive scores achieved 91% discrimination on the same problem. Zhang et al. combined structural MRI, FDG-PET and CSF information to achieve a 93%

accuracy on AD/HC discrimination and 76% accuracy discriminating MCI from healthy controls (MCI/HC). In comparison, both structural MRI and FDG-PET individually achieved around 86% accuracy on AD/HC and 72% on MCI/HC. Methodologically, the two studies took different approaches to finding the optimal combination of kernels: Hinrichs et al. tuned the kernel weights as part of the SVM optimization problem while Zhang et al. took the more straightforward approach of searching for optimal weights via grid search in a nested cross-validation scheme.

The MKL approach has been expanded upon in subsequent studies. For example, Young et al., (2013) used MKL in a Gaussian Process classifier (GPC), learning the kernel weights as hyperparameters using type-II maximum likelihood⁷. Their approach avoids the need for finding kernel weights via nested cross-validation while providing probabilistic predictions by virtue of the GPC's Bayesian formulation. The study achieved a 74% accuracy in discriminating MCI converters from non-converters using a combination of structural MRI, FDG-PET, CSF and genetics. In addition it showed a statistically significant improvement of the combined features over purely MRI and purely PET based features.

Though multimodal biomarkers are promising, collecting the relevant data entails a higher cost and time commitment compared to single modality biomarkers. With these trade-offs in mind, Yu et al., (2012) sought to identify the optimal combination of structural MRI, FDG-PET, CSF and genetics in predicting amnesic MCI conversion to AD within two years, for the purpose of enriching clinical trial populations. Using all biomarkers yielded an 81% accuracy while MRI emerged as the best single modality with a 78% accuracy, compared to FDG-PET at 68% and CSF at 65%. The combination of ApoE (genetics) and cognitive scores had the lowest prediction accuracy of 62%. When taking cost and time into account, the study concluded that the combination of MRI, ApoE and cognitive scores was optimal for patient recruitment in clinical trials. Taking an “all MRI” approach to multimodal classification may avoid some of these trade-offs. Dyrba et al., (2015) combined structural MRI, diffusion tensor imaging (DTI) and rs-fMRI in an MKL based classifier, discriminating 28 AD subjects from 25 controls. They found that

⁷ Type II maximum likelihood refers to maximizing the likelihood function with respect to the hyperparameters instead of computing the hyperparameters' posterior distribution. This approximation is widely used in practice.

DTI based measures of WM fibre tracts and structural MRI based GM volume resulted in the highest accuracy of 89%, while rs-fMRI did not improve predictive accuracy. Further studies are needed to understand the optimal combination of modalities for detecting early AD.

2.4.6 Longitudinal

Longitudinal information consists of repeated measurements of the same subjects over time. It is particularly important in the study of neurodegenerative diseases as they are, by definition, marked by progressive loss of structure and/or function (Jack et al., 2004; Raz et al., 2010; Risacher et al., 2010; Raz and Lindenberger, 2011). There is a long history of the use of longitudinal information in AD discrimination, often involving the comparison of a baseline to a single follow-up image, either on a whole-brain level or within a specific ROI, e.g. measuring hippocampal atrophy or ventricular dilation. Such measurements are typically performed by trained clinicians, particularly measuring hippocampal atrophy, although automated methods for quantifying whole-brain atrophy have been widely used in research. Two of the most influential automated methods are the boundary shift integral (BSI) ((P. A. Freeborough and Fox, 1997a) and FSL's SIENA algorithm (Smith et al., 2001; Smith et al., 2002). Both methods automatically coregister follow-up and baseline structural MR images and then compute a metric that summarizes longitudinal brain atrophy. Smith et al., (2007) directly compared the two methods and found they provide very similar estimates of atrophy rates in both healthy controls and AD subjects. A univariate classifier based on the atrophy rate measurement achieved an 80% cross-validated discrimination rate between AD subjects and age-matched controls. The study concluded that multivariate analysis may offer more discrimination power.

Accordingly, more recent studies have started to use high-dimensional longitudinal features in multivariate classification. Misra et al., (2009) formed a "beta" image for each subject by performing a voxel-wise linear regression on all available longitudinal structural MRI based samples. Examining rate of change images rather than simple two-sample difference images allows for the use of an arbitrary number of samples per subject and may lead to more robust estimates of

degeneration. They found few significant group differences between MCI converters and non-converters using this longitudinal information, noting a potential lack of a sufficient number of follow-up images per subject (though the study required at least three time-points' information per subject). The study concluded that subjects' baseline measurements were far better at distinguishing MCI converters from non-converters than their longitudinal information. The multi-modal study of Hinrichs et al., (2011) formed a separate kernel for each baseline imaging modality (structural MRI and FDG-PET) and a kernel for longitudinal imaging (structural MRI based Jacobian determinants of the deformation fields from baseline to 24 month follow-up). In contrast to Misra et al, they found longitudinal information to be slightly better at predicting MCI progression to AD compared to baseline information. The study did not include a longitudinal FDG-PET based kernel as it was found to weakly discriminate AD subjects from controls.

The study of Gray et al., (2012), discussed in Section 2.4.4, evaluated FDG-PET based cross-sectional (both baseline and one year follow-up) and longitudinal features in pairwise discriminations of AD, progressive and stable MCI subjects (pMCI and sMCI, respectively) and healthy controls (HC). In contrast to Hinrichs et al's MKL approach, they simply concatenated the cross-sectional and longitudinal features, finding that the combination of cross-sectional and longitudinal information (follow-up plus longitudinal in particular) led to a small but significant improvement over purely cross-sectional features. One year longitudinal information on its own was shown to be significantly less discriminative than either of the two cross-sectional features sets in the four pairwise classifications that were performed (AD/HC, AD/sMCI, pMCI/HC, pMCI/sMCI).

To broadly summarize, in many studies longitudinal information on its own appears to be approximately equally accurate or worse at discriminating stages of AD relative to purely cross-sectional information. Few studies have shown significant improvements in prediction accuracy using longitudinal information compared to cross-sectional information. A recent review of longitudinal neuroimaging in neurodegeneration came to a similar conclusion: despite the potential of longitudinal analysis in progressive diseases, the use of longitudinal biomarkers "in most cases unfortunately remains aspirational" (Schuster et al., 2015). This is somewhat at odds with the recommendations of 2011 NINCDS-

ADRDA criteria for both MCI (Albert et al., 2011) and, particularly, preclinical AD (Sperling et al., 2011), which both emphasize the importance of longitudinal information. The preclinical AD guidelines state that the change in cognition over time is likely to be more sensitive than any one-time measure. Schuster et al. point out that some longitudinal studies, particularly those studying preclinical stages, may be recruiting the wrong patients: recruitment should be based on incidence rather than prevalence, as patients who have already manifested symptoms may progress more slowly (with harder to detect within-subject changes) than pre-manifest cases.

Schuster et al. also recommend the use of more robust statistical models, such as mixed effect linear models that can handle varying numbers of measurements per subject (including a single measurement) and that can model non-linear trajectories. Non-linear trajectories of decline have been shown to be important in ageing (McDonald et al., 2009; Fjell et al., 2013). The need for better methodology extends to the multivariate classification context. Chen and DuBois Bowman, (2011) modified the SVM's optimization problem to model a linear combination of an arbitrary number of cross-sectional time-points' features. Using ADNI's PET data (baseline and one year follow-up for all subjects) they discriminated AD subjects from controls with 65-67% accuracy using baseline and concatenated baseline plus follow-up features, while their longitudinal SVM method achieved 75-78% accuracy. Another approach is to improve longitudinal registration algorithms to create better features for existing pattern recognition algorithms. Examples of such methods include that of Holland and Dale, (2011), the extension of the boundary shift integral to multiple time-points (Leung et al., 2012), and the symmetric, diffeomorphic (multiple time-point) method of Ashburner and Ridgway, (2013). Finally, there exists a conceptual space between image registration and pattern recognition, namely feature construction, that can also make use of longitudinal information. The advantage of such an approach is it does not depend on the choice of a particular image registration method or pattern recognition algorithm.

2.5 PD discrimination

Taking a similar approach to the AD sections, these sections review studies discriminating PD subjects from controls using neuroimaging with a view towards

understanding the context of such studies within the current diagnostic criteria and the known progression of disease symptoms, as discussed in Section 2.3. The pattern recognition based discrimination literature in PD is sparse, especially when compared to the corresponding literature on AD discrimination. We review some of the increasing number of structural MRI based studies despite a historically minor role in PD discrimination (Pyatigorskaya et al., 2014), along with several ^{123}I -Ioflupane SPECT based studies. We also review several studies using other modalities, such as transcranial sonography (TCS) and diffusion MRI, and point out the need for more multi-modal and longitudinal studies.

2.5.1 Structural MRI

Structural MRI has primarily been used to discriminate among Parkinsonian disorders - mirroring the diagnostic problem in clinical practice - rather than between PD subjects and controls. The problem of separating idiopathic PD (IPD) from related disorders such as progressive supranuclear palsy (PSP) and multiple system atrophy (MSA) in the early stages remains a challenge. This is important because the disorders each have a different underlying pathology, a different progression trajectory and different responses to treatment. Importantly, these disorders have very different prognoses, with median time to death of 5 years, 7-8 years and over 15 years for PSP, MSA and PD respectively (Litvan et al., 1996; Figueroa et al., 2014).

Duchesne et al., (2009) have shown that it is possible to differentially discriminate Parkinsonian disorders using volumetric data based on T1-weighted structural MR images. They discriminated IPD subjects from MSA or PSP subjects with an accuracy of 91%. Although the study contained 181 subjects, there were only 16 patients with IPD, eight patients with MSA and eight patients with PSP. As the authors note, the small number of disease subjects limited the ability to separate the individual classes further. Attempts to separate any or all of the Parkinsonian disorder subjects from healthy controls were not presented. Focke et al., (2011) presented similar accuracies separating MSA-P (the Parkinson's variant of MSA), PSP and IPD (referred to as IPS) using T1-weighted images. Some notable accuracies were: 87% discriminating PSP from IPD using grey matter and 97% using white matter, 72% and 65% discriminating MSA-P from IPD using grey

matter and white matter respectively. The study discriminated MSA-P subjects from controls with 70% accuracy using grey matter and PSP subjects from controls with 94% accuracy using white matter. The study could not discriminate IPD subjects from controls, achieving a roughly 40% accuracy that is less than chance level.

Marquand et al., (2013) discriminated multiple Parkinsonian disorders simultaneously using a probabilistic multi-class classifier, with features restricted to the subcortical motor network. Probabilistic models, such as the Gaussian Process classifier that was used, are preferable as they provide a measure of confidence with their prediction, an important point for clinical decision-making. The benefit of multi-class classification is that it makes it possible to classify a subject into a specific variant of a complex disease without much prior knowledge of disease state. Briefly, resulting accuracies were: 91% separating PSP, IPD and MSA (both cerebellar and Parkinsonian variants grouped together), 85% separating of PSP, IPD, MSA-P and MSA-C (the cerebellar variant of MSA) and, when including healthy controls (HC), 74% separating PSP, IPD, MSA and HC. On the hardest problem, separating PSP, IPD, MSA-P, MSA-C and HC simultaneously, the result was 63%, which was still statistically significant.

Few studies have been able to discriminate PD subjects from controls using T1 weighted MRI, although we found several instances of successful discrimination. Salvatore et al., (2014) performed pairwise discriminations of 28 PD subjects, 28 PSP subjects and 28 healthy control subjects using whole-brain voxels achieving an 86% accuracy when discriminating PD subjects from controls, 89% accuracy discriminating PSP subjects from controls and 89% accuracy discriminating PSP from PD subjects. The large sample study of Adeli et al., (2016) used the PPMI dataset to discriminate 374 de novo PD subjects from 169 controls on the basis of 90 ROIs from a standard whole brain atlas (AAL, Tzourio-Mazoyer et al., 2002) and an additional eight ROIs from the brainstem, left and right red nuclei and left and right SN. GM, WM and CSF tissues volumes were computed at each ROI and concatenated to form a 294 length feature vector for each subject. The authors achieved up to 81% accuracy using an optimization based method that simultaneously discards poor samples and irrelevant features during classifier training. Both studies noted that the classifier weights, which should be interpreted with caution (Haufe et al., 2014a), indicate a pattern of cerebral and subcortical

regions (including the SN and pons in both studies) are involved in discriminating patients from controls. Salvatore et al note that the spatially distributed pattern is consistent with the work of Braak et al., (2003) showing PD pathology spreading throughout the brain, starting from the brainstem and proceeding in stages to the neocortex via the basal ganglia.

2.5.2 Diffusion MRI

The influential study of Vaillancourt et al., (2009) identified reduced fractional anisotropy (FA), a measure derived from diffusion tensor imaging (DTI), in the SN of 14 de novo PD subjects compared to 14 controls. Their method achieved 100% accuracy in discriminating PD subjects from controls using manually drawn regions of interest in the SN. Péran et al., (2010) derived a small set of features using a combination of T_2^* -weighted and DTI imaging derived measures (FA and mean diffusivity, i.e. MD) of the SN, putamen and caudate. They were able to discriminate 30 PD subjects from 22 controls with 95% area under the receiver operating characteristic curve. Salamanca et al., (2015) used FA and MD within cerebellar and subcortical ROIs as features to discriminate 50 de novo PD subjects from 50 controls (a subsample of the PPMI dataset) with up to 77% accuracy.

In contrast to these promising studies, Schwarz et al., (2013) performed a DTI based study of 59 subjects (32 PD subjects and 27 matched controls) along with a meta-analysis of eleven studies to estimate the effect size of disease related nigral DTI changes. They concluded that, at present, DTI metrics are not a useful diagnostic biomarker and that unpublished negative study results are probably under-represented in the literature.

Similar to the structural MRI based studies, a whole-brain voxel based approach may better discriminate PD from controls using diffusion MRI data. While few such discrimination studies have been published, Rae et al., (2012) showed decreased FA and increased MD between PD subjects and controls in many white matter tracts in the frontal and parietal lobes, with an increased sensitivity using tract based spatial statistics (TBSS, Smith et al., 2006) compared to voxel based analysis.

2.5.3 ¹²³I-Ioflupane SPECT

¹²³I-Ioflupane SPECT (also known as DAT-SPECT or DaTSCAN) is a dopamine transporter (DAT) imaging method that is used to detect disturbed dopaminergic functioning. It has been shown to differentiate Parkinsonian syndrome (PD, MSA, PSP) subjects from those with essential tremor and healthy controls (Benamer et al., 2000). A difference between normal and abnormal scans can usually be detected by visual assessment based on the pattern of DAT binding in the caudate nucleus and putamen. However, the use of quantitative information has been recommended in addition to visual evaluation to improve diagnostic accuracy (Darcourt et al., 2009).

Several large sample studies have in recent years demonstrated semi-automated or fully automated approaches to discriminating PD using DAT-SPECT imaging derived features. Prashanth et al., (2016) developed a semi-automated approach that uses the shape and surface of the regions of high DAT activity in the caudate and putamen as features, achieving a 97% accuracy in discriminating the PPMI study's de novo PD subjects from a class that combined a healthy control group and a group of subjects with Scans Without Evidence of Dopaminergic Deficit (SWEDD). The SWEDD group subjects are those that have been diagnosed with PD (on the basis of motor symptoms) but which show normal dopaminergic activity via SPECT, raising the possibility that these subjects have been misdiagnosed with PD (Schneider et al., 2007). Illan et al., (2012) presented a fully automated approach to discriminate Parkinsonian syndrome (PS) subjects from controls using whole-brain voxel based features and an SVM classifier with 91% accuracy. Oliveira and Castelo-Branco, (2015) used voxels from the striatum as features with an SVM to discriminate de novo PD subjects from healthy controls with 98% accuracy using the PPMI dataset.

These studies demonstrate the potential of DAT-SPECT in discriminating PD or PS from controls. Based on the PPMI studies of Prashanth et al., (2016) and Oliveira and Castelo-Branco, (2015), DAT-SPECT holds promise in detecting early PD. Further to this, it has been combined with other sources of information in several multi-modal discrimination studies of early PD, as discussed in Section 2.5.5. Importantly, however, DAT-SPECT relies on detecting a significant reduction of dopamine (up to 67%), meaning there must be significant loss of dopaminergic SN

cells. This may limit its ability to detect prodromal PD. In addition, commonly used Parkinson's medications such as levodopa may alter the DAT signal, limiting longitudinal use (Fahn et al., 2004). Finally, it is an expensive and involved test for patients, taking roughly five hours from pre-medication and ligand injection to scan finish.

2.5.4 Transcranial sonography (TCS)

Transcranial sonography (TCS) is an inexpensive and non-invasive imaging technique that can be easily applied in a clinical setting. Studies have shown that there is a distinct pattern of hyperechogenicity of the SN in PD subjects compared to controls that can be detected via TCS (Berg et al., 2001; Hagenah et al., 2006). TCS based assessment of the SN is typically performed manually, with imaging quality that depends on the experience of the operator as well as the acoustic bone window of the patient. There are no known studies that have performed a fully automated discrimination of PD using TCS. Chen et al., (2012) developed a semi-automated means of extracting features from TCS images, achieving up to 78% accuracy in discriminating PD subjects from controls using an SVM classifier.

2.5.5 Multimodal and longitudinal

Recent studies such as the large, multi-centre PPMI study (Marek et al., 2011) and the single site PMPP study (Liepelt-Scarfone et al., 2013a) have begun to collect longitudinal, multi-modal biomarkers of early PD subjects alongside healthy controls. At present there is a dearth of discrimination studies that make use of either type of information. Our literature review did not identify any studies using longitudinal features to discriminate PD and only a few studies using multi-modal features. Long et al., (2012) concatenated rs-fMRI and structural MRI derived features to discriminate 19 early PD subjects from 27 controls with 87% accuracy, compared to 74% accuracy using strictly rs-fMRI based features and 80% accuracy using structural MRI features. Prashanth et al., (2016a) used a combination of non-motor clinical assessments and imaging available from the PPMI study to discriminate de novo PD subjects from controls with up to 96% accuracy. The

features used were composed of the REM sleep Behavior Disorder Screening Questionnaire (RBDSQ) score, the University of Pennsylvania Smell Identification Test (UPSIT) score, measures of CSF and striatal binding ratios (SBRs) calculated from ^{123}I -Ioflupane SPECT imaging. Combining pre-motor and imaging markers appears to be a promising approach and warrants validation by other studies and in other datasets such as the PMPP study, which also tracks an additional group of subjects identified to be at high risk of developing PD. As more studies identify and track groups of subjects in the pre-motor stage of PD there will be a greater need and opportunity for longitudinal and multi-modal discrimination methods that are able to use both subtle within-subject changes and diverse biological and clinical measures (Schuster et al., 2015).

2.6 Pitfalls and validation

There is a varying level of technical expertise needed to carry out a pattern recognition based study using neuroimaging data. In all cases, one needs to collect and collate the necessary imaging data, pre-process the images to form meaningful features that can be compared across subjects, performing some form of quality control at one or more steps along the way. Following the creation of features one can use an existing pattern recognition toolbox to train and test the performance of a classification or regression algorithm in a cross-validation based framework (Hanke et al., 2009; Schrouff et al., 2013). Such software has reduced the barriers for performing analysis, allowing non-specialists access to state-of-the-art algorithms such as the SVM and the Gaussian Process Classifier (GPC). Alternatively, one can implement some or all of the pattern recognition analysis if one desires a specialized form of feature selection, dimensionality reduction, pattern recognition algorithm or cross-validation scheme. Although custom analysis can lead to better results, it is important to emphasize that widely available toolboxes can reduce the prevalence of commonly occurring “pitfalls” in pattern recognition based analysis. Typical pitfalls include: the use of global statistics (normalizing using global mean and variance, for example), global rejection of outliers or extraction of features (using information about the test set), simultaneous selection of model parameters and evaluation of performance by performing cross-validation on the same data (rather than using

nested cross-validation, for example) (Lemm et al., 2011). Avoiding these pitfalls enforces a complete split between the training and testing data at each cross-validation fold, thereby ensuring accurate estimates of generalizability. Cross-validation is described in Section 4.4.3. Briefly, it is a systematic means of estimating the generalizability of a model using an available dataset.

In addition to the pitfalls mentioned, there is also the possibility of overfitting, i.e. training an overly complex model with subsequently poor performance in testing. Overfitting can be a particular problem when one is faced with high-dimensional features and a relatively low sample size, a situation which pertains to voxel-based neuroimaging analysis. The problem of overfitting is independent of the pitfalls we have mentioned, i.e. it is possible to overfit a model while still ensuring a complete separation between training and testing data during cross-validation. Modern pattern recognition algorithms use regularization (see Section 4.4) to minimize overfitting.

In light of these considerations, models should be validated on publicly available datasets, similar to the practice in computer science of testing pattern recognition algorithms on the MNIST dataset of handwritten digits. The ADNI dataset in AD and, increasingly, the PPMI dataset in PD have filled this role to date, with several studies performing validation of features and algorithms using ADNI in particular (Cuingnet et al., 2011; Chu et al., 2012; Sabuncu and Konukoglu, 2014). There are a number of other datasets that could potentially be useful in the study of neurodegeneration (Eickhoff et al., 2016). In addition to validation, public datasets enable the decoupling of data collection from analysis, granting researchers specializing in pattern recognition access to high quality data. As a result of ADNI, the field of AD discrimination has advanced significantly and at present is more developed than the corresponding PD literature. The recent review of Arbabshirani et al., (2016) has highlighted the need for further data sharing in neuroimaging.

In addition to validation studies, competitions can provide a valuable means of assessing the clinical applicability of automated diagnostic or prognostic systems. Competitions can reduce a subtle form of overfitting, in which the literature converges on a set of methods that perform well on a particular publicly available dataset. The recent CADDementia challenge aimed to assess the performance of

algorithms performing multi-class classification of AD, MCI and healthy controls using 354 T1-weighted structural MRI scans, with fifteen teams submitting 29 algorithms for evaluation (Bron et al., 2015). The best performing algorithm achieved 63% accuracy and 79% receiver operating characteristic area under the curve (ROC AUC). The study noted that this performance is difficult to compare to the literature as most studies perform pairwise classification (AD versus controls for instance) rather than simultaneous multi-class classification. Nevertheless, across all algorithms MCI had the consistently lowest true positive rate compared to the AD and healthy control classes, a finding that is in line with the literature and demonstrates the difficulty of discriminating this class using structural MRI and also that the true generalizability is probably lower than what is reported. Features based on VBM and those that combined volume and cortical thickness, shape and intensity generally performed best.

2.7 Conclusions

We have tried to provide an overview of some of the relevant literature in AD and PD discrimination studies, having restricted ourselves to automated, pattern recognition based studies using neuroimaging data. Across both diseases we have identified a variety of imaging modalities and preprocessing techniques to derive features from a particular modality. We have tried to frame the discussion of these modalities within the respective diagnostic criteria and disease progression models of both diseases.

Within AD this perspective may explain the difficulty of using structural MRI based features in discriminating subjects in the earlier stages of AD from controls and suggests the use of functional neuroimaging based features derived from rs-fMRI, which measures the connectivity of spatially distributed brain regions and FDG-PET, which measures the brain's metabolic activity. The automated classification of amyloid status using amyloid PET appears promising, though the role of such pathological AD biomarkers to predict the clinical stage of AD is not clear; some subjects with AD pathology may never manifest clinical symptoms. It is possible that amyloid PET can be multi-modally integrated with other functional and structural imaging modalities to detect or predict the onset of early AD.

Multi-modal AD discrimination studies are starting to show improvements in accuracy over any single constituent modality, although on the whole these improvements have been modest. The cost and complexity of collecting a variety of imaging and behavioural measures is an important issue that some studies have addressed and others will have to address in the future if multi-modal biomarkers are to be practically used. Few discrimination studies have shown significant improvements using longitudinal features compared to cross-sectional features. There is a need for better longitudinal methods that are able to make use of subtle within-subject changes in brain structure or function. The revised criteria for AD, particularly the parts pertaining to MCI and preclinical AD, along with a number of neurodegeneration focussed reviews emphasize the importance of such information in detecting the earliest stages of neurodegeneration.

Within PD, structural MRI studies are beginning to discriminate de novo subjects from controls. At present there are few large sample diffusion MRI studies in the literature, though using whole brain features in this case appears to be promising. Several studies have attempted semi-automated discrimination using TCS, though image quality, at present, ultimately depends on the skill of the operator. Several SPECT studies have shown good results in discriminating early PD from controls and this modality has been further combined with pre-motor markers in multi-modal studies, with promising results. On the whole however, there have been very few multi-modal or longitudinal PD discrimination studies to date. The PPMI study has made longitudinal information publicly available though it has been little used in PD discrimination studies.

Finally, we have tried to briefly highlight some important considerations for building predictive models: (i) the use of established pattern recognition toolboxes can reduce the number of commonly occurring modelling pitfalls; (ii) publicly available datasets and data sharing schemes have greatly improved the quality of discrimination studies and will continue to do so in the future; and (iii) competitions further validate existing methods and are perhaps the best means of assessing the clinical applicability of automated discrimination systems.

Chapter 3 Datasets

3.1 Introduction

This chapter describes two neurodegeneration studies' datasets that are used throughout the remaining thesis. These longitudinal studies aim to track a disease group or groups along with demographically matched healthy controls, with neuroimaging data available from multiple time-points for some or all study subjects in both cases. The first study we describe is the Heinz Nixdorf RECALL substudy (HNRS) of mild cognitive impairment (MCI), one of the few studies that specifically tracks subjects diagnosed with MCI, a heterogeneous diagnostic entity that represents early stage Alzheimer's disease (AD) (Morris et al., 2001; Petersen, 2004). The second study is the Progression Markers in the Premotor Phase (PMPP) study of Parkinson's disease (PD) that monitors disease progression in subjects identified to be at risk for developing Parkinson's disease (Dlugaj et al., 2010; Liepelt-Scarfone et al., 2013).

These studies aim to identify biomarkers of early disease and disease progression in their respective neurodegenerative diseases. Both studies are well suited for exploring the goals of this thesis (discussed in Chapter 1), which are (i) to evaluate and compare the ability of current cross-sectional and longitudinal imaging based features to discriminate early neurodegeneration (see the literature review of Chapter 2 and the preliminary analysis of Chapter 5) and (ii) to develop new pattern recognition based methods that make use of longitudinal information in disease discrimination (see Chapters 6, 7 and 8).

In this chapter we review the designs of the two studies, the data collected in each and the choices made with respect to which of the available imaging modalities were analysed in subsequent chapters.

3.2 Study Designs

Heinz Nixdorf RECALL substudy (HNRS)

The Heinz Nixdorf RECALL (Risk Factors, Evaluation of Coronary Calcium and Lifestyle) study is a population-based prospective cohort study with subjects randomly selected from mandatory city registries in Germany. Study methods have been previously described in detail (Schmermund et al., 2002; Stang et al., 2005). Briefly, 4814 participants between the ages of 45 and 75 were enrolled between 2000 and 2003 in the Ruhr area of Germany. After the baseline examination participants were followed over a five year period when a second examination was conducted. The second examination (response rate: 90.2%) included a short cognitive performance assessment (for details regarding participants and drop-outs see Dlugaj et al., (2010)), which was accomplished in 4086 study participants. At this follow-up time-point, a random sample of participants (aged 50-80) with impaired short cognitive performance assessment results (n=701) and age appropriate short cognitive performance assessment results (n=316) were invited to a detailed neuropsychological and neurological examination to assess mild cognitive impairment (MCI) and its subtypes for inclusion in the HNR substudy (HNRS) (Dlugaj et al., 2010).

MCI was diagnosed according to the International Working Group on MCI criteria (Petersen, 2004; Winblad et al., 2004) with the exception of the cognitive complaint criterion (the subject or the informant had to express some concern about the person's cognitive function). Thus, the following criteria were necessary for the diagnosis: (1) evidence for impairment in cognitive function on the administered objective cognitive tasks, which was not normal for age; (2) evidence for preserved basic activities of daily living/minimal impairment in complex instrumental functions and (3) exclusion of the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, fourth edition) dementia diagnosis. Participants with dementia, severe depression (ADAS depression subscale score >4), Parkinson's disease, mental retardation, severe alcohol consumption (for women > 20g/day; for men >40g/day), known brain cancer, severe problems with the German language (in foreign persons) and severe sensory impairment leading to invalid cognitive testing were excluded from the analysis.

In total, the HNRS consisted of 148 MCI cases identified and matched, at the five year follow-up of the total HNR cohort, to 148 healthy controls according to age, sex and education. Participants were examined with MRI (detailed in Section 3.3) at the substudy baseline and at the substudy 2.5 years follow-up. A second follow-up with MRI, approximately 5.5 years after substudy baseline, is being conducted, with data not yet available. All participants gave their written informed consent. The study was approved by the local institutional ethical committee and followed established guidelines of good epidemiological practice.

After exclusion of MRI contraindications, the study included a subsample of 252 subjects (85% of all 296 substudy subjects), with imaging data from 126 MCI subjects and 126 matched controls available for analysis (Mönninghoff et al., 2015).

Progression Markers in the Premotor Phase (PMPP) study

The Progression Markers in the Premotor Phase (PMPP) study of Parkinson's disease aims to monitor progression of neurodegeneration in three groups of individuals: (i) individuals assessed to be at high risk for PD (HR_{PD}) group, (ii) those diagnosed with early PD and (iii) a healthy control (HC) group composed of two subgroups (described later). This two year longitudinal study was designed to track the progression of a variety of biomarkers of PD in the prodromal phase, i.e. between the first appearance of symptoms and the full development of disease. We briefly describe the study design below; a complete description is provided in Liepelt-Scarfone et al., (2013).

Figure 3.1 depicts the design of the study. Basic and extended data, detailed in Section 3.3, were collected at baseline and two year (final) follow-up for control group 1. The HR_{PD} and PD groups had basic data collected at six month intervals from baseline to final follow-up, with extended data collected yearly. All participants gave their written informed consent. The study was approved by the local institutional ethical committee of the University of Tübingen and was conducted in compliance with the 1964 Declaration of Helsinki and its later amendments.

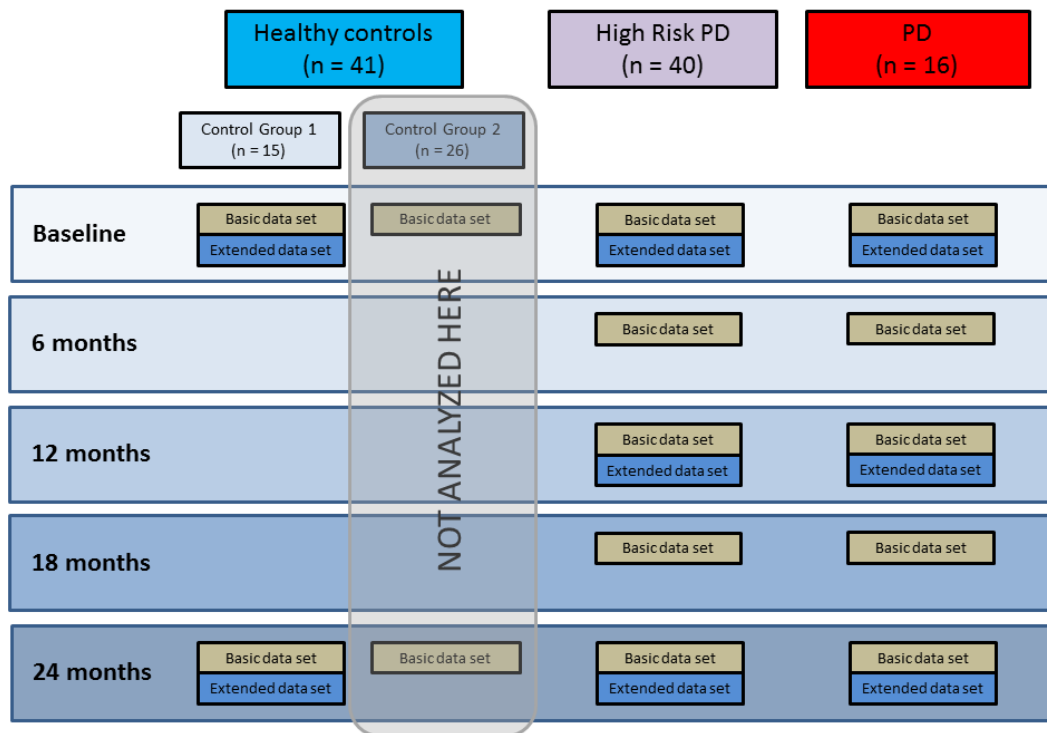


Figure 3-1 Design of the PMPP study, modified from Figure 1 of Liepelt-Scarfone et al., (2013)

The PD group consisted of 16 subjects diagnosed by established criteria (Hughes et al., 1993), chosen to be in the early stages of the disease (mean disease duration at baseline was 31 months, with Hoehn and Yahr stage ≤ 2.5). Subjects at this stage of the PD may exhibit bilateral symptoms that affect posture and gait, though overall they have minimal disability and are physically independent. Subjects were required to be older than fifty, with no history of deep brain stimulation and no genetic mutation known to cause PD. The HR_{PD} group consisted of 40 subjects with hyperechogenicity of the substantia nigra (termed SN+) as measured by TCS and one of the following: one cardinal motor sign of PD as measured by UPDRS-III or any two of the following markers: lifetime prevalence of depression, hyposmia, one-sided reduced arm swing or positive family history of PD. SN+ has been shown to be related to tissue iron content (Berg et al., 2002) and is present in 90% of PD subjects (Gaenslen et al., 2008). Importantly, about 10% of all individuals over the age of 79 years are SN+ and TCS cannot always be performed reliably or reproducibly in some individuals owing to insufficient temporal bone window. Furthermore SN+ is not

specific to PD: may be associated with other neurodegenerative disorders, such as dementia with Lewy bodies and corticobasal degeneration (Berg et al., 2011).

The healthy controls group had 41 subjects, chosen as those with normal Substantia Nigra (SN) echogenicity as measured by TCS, no signs of acute psychiatric diseases, and negative family history of PD. Controls were split into two subgroups: control group 1 consisted of 15 subjects, matched by age, gender and education to the PD group and control group 2, consisting of 26 subjects that served as a validation group for blood markers and questionnaires. Control group 2 had only basic data collected at baseline and two year follow-up. As this group did not have imaging data it was not included in any of the analyses presented in this thesis.

The exclusion criteria for all subjects were: other neurological diseases, history of drug or alcohol abuse, prior use of cholinesterase inhibitors or memantine or a Mini-Mental State Examination (MMSE) score ≤ 24 with a relevant cognitive deficit indicative of dementia.

3.3 Study Data

Heinz Nixdorf RECALL substudy (HNRS)

A standardized neuropsychological examination was conducted by a neuropsychologist using the following test assessments: Alzheimer's Disease Assessment Scale (ADAS), the number connection test of the Nuremberg Gerontopsychological Inventory (German: Nürnberger Altersinventar (NAI)) (Oswald and Fleischmann, 1994), the verbal fluency test (Aschenbrenner et al., 2000) and an Instrumental Activities of Daily Living (IADL) (Oswald and Fleischmann, 1994) scale to assess activities of daily living. For each cognitive domain age-specific test norms were administered. A cognitive domain was rated as impaired if the performance was more than one standard deviation (SD) below the age adjusted mean. A cut-off of one SD was chosen for the screening test as well as for the detailed neuropsychological assessment, because it was found to be associated with a higher relative prognostic power in predicting the development of dementia compared with a cut-off of 1.5 SD (Busse et al., 2003). Furthermore it provides a higher sensitivity, which was particularly important for the screening test

to detect participants in need of further neuropsychological assessment (Busse et al., 2003). Depression was assessed using the depression subscale of the ADAS (Ihl and Weyer, 1993). A detailed physical examination with particular focus on the neurological examination was conducted by a neurologist. Furthermore, a medical history was gathered related to cognitive functioning, duration of such symptoms, history of other medical illnesses and current treatment.

The imaging portion of the dataset consisted of structural MRI, in particular T1-weighted, fluid-attenuated inversion recovery (FLAIR) and T2*-weighted images. We used the T1-weighted images in Chapters 5 and 7.

Progression Markers in the Premotor Phase (PMPP) study

The basic data set consisted of motor assessments: Unified Parkinson's Disease Rating Scale, part III: motor evaluation, i.e. UPDRS-III and an arm swing task, a blood sample, history of medication and concomitant disease, olfaction tests, a depression assessment (the Beck Depression Inventory), a sleep disturbance assessment (PD Sleep Scale 2 and a REM sleep behaviour disorder, i.e. RBD screening) and a neuropsychological test of cognitive impairment (Mini-Mental State Examination, i.e. MMSE).

The extended data set is detailed in Appendix C. The MRI portion of the extended data set was available for analysis, consisting of T1 Weighted Structural, Diffusion Tensor Imaging (DTI), Arterial Spin Labelling (ASL), T2 Turbo Spin Echo (TSE), T2* Gradient Recalled Echo (GRE) and Magnetic Resonance Spectra (MRS) data. The 123I-FP-CIT-SPECT and Transcranial Sonography (TCS), both of which were only collected for a subset of study subjects, were not available for analysis. The electroencephalography (EEG) portion of the electrophysiology data was available. All motor, clinical, demographic, and neuropsychological data were also available for analysis.

3.4 Discussion

Both studies we have described aim to identify and track early neurodegenerative disease subjects along with matched controls. In both cases the early disease groups' subjects have been carefully selected to represent those in the prodromal phase of neurodegeneration. Within the HNRS study, the MCI subjects have been diagnosed according to widely used MCI criteria. MCI has been identified as the prodromal phase of Alzheimer's disease (AD), with subjects progressing to dementia at a rate of 10-15%, compared to 1-2% for cognitively normal elderly subjects (Petersen, 2004a). Within the PMPP study, the HR_{PD} group subjects have been selected based on a combination of presumed neuronal loss (in the substantia nigra) and specific motor or non-motor markers that the study proposes as a means of identifying individuals at increased risk for PD. These inclusion criteria are based on known and verified risk factors strongly associated with the development of PD (Braak et al., 2003; Przuntek et al., 2004; Gaenslen et al., 2008; Liepelt-Scarfone et al., 2013). In support of this view, a recent analysis of the PMPP study's clinical measures identified group differences in autonomic function between HR_{PD} subjects and controls (Liepelt-Scarfone et al., 2015). Autonomic regulatory dysfunction is widely known to occur in the early (pre-motor) phase of PD (Gaenslen et al., 2011).

As discussed in Chapter 2, T1-weighted structural MRI based biomarkers have been widely used in dementia research. Such biomarkers have shown promise in discriminating Alzheimer's disease subjects from controls and in predicting which MCI subjects will convert to Alzheimer's disease. In this thesis we attempt to discriminate MCI subjects from healthy age matched controls, a challenging problem as MCI is marked by subtle differences between disease subjects and controls along both clinical and imaging biomarker measures. Current diagnostic criteria for MCI point out that hippocampal or medial temporal volume along with measures of the rate of brain atrophy, both of which can be derived from T1-weighted MRI, may be useful for diagnosing MCI (Albert et al., 2011). As such, we use the longitudinal T1-weighted MRI that is available in the HNRS dataset when creating cross-sectional and longitudinal features for discriminating MCI. Note however that the T2*-weighted and FLAIR images from the HNRS dataset were assessed by two neuroradiologists to form measures of white matter hyperintensities (WMHs), cerebral microbleeds (CMBs) and brain atrophy (Mönninghoff et al., 2015). The

study found significant group differences in cortical atrophy and occipital periventricular WMHs between the MCI control groups, though there were no differences in CMBs between the two groups.

PET and SPECT are commonly used to detect dopaminergic dysfunction related to PD. MRI, in contrast, has had little clinical use although progress has been made in recent years towards developing biomarkers based on both structural and functional MRI (Pyatigorskaya et al., 2014). As discussed in Chapter 2, T1-weighted MRI can be used to accurately discriminate between Parkinsonian disorders on the basis of volumetric differences in the cortex, in subcortical structures and in the brainstem and cerebellum. Few studies have shown significant differences between PD subjects and controls using this modality, possibly due to its poor contrast in many of the subcortical and brainstem structures (including the substantia nigra (SN)) that play an important role in the progression of PD (Braak et al., 2003). T1-weighted images offer good grey/white matter contrast in the cortex however and automated, whole-brain methods from voxel-based morphometry (VBM) are well developed and widely used with this modality (Ashburner and Friston, 2000; Ashburner and Friston, 2001). A few studies suggest these methods may have the potential to detect the subtle differences in cortical volume and thickness that may be relevant to discriminating PD subjects from controls. We use whole brain T1-weighted MRI based features in the preliminary analysis presented in Chapter 5.

Magnetic Resonance Spectra (MRS) were available for analysis and have been shown to detect differences in metabolism within subcortical structures in PD subjects compared to controls. Unfortunately MRS data was not available for many subjects and further, what was available was assessed to not be in a usable form. Therefore the decision was made to exclude this modality from analysis. There is also T2-weighted Turbo Spin Echo (TSE) MR imaging available for many subjects, from which T2 maps can be calculated. However, the protocol was changed midway through the study and as a result some of the images are composed of 30 slices, while others have 10 slices. In addition there is reduced slice coverage, covering the area from the base of the cerebellum midway through the striatum. The decision has been made not to analyse this data. The T2* Gradient Recalled Echo (GRE) images that were available for many subjects appeared to be sensitive to susceptibility

artefacts that would prevent their use in discriminating PD and as such they were also not analysed.

The DTI and ASL data were both usable upon initial inspection. In Chapter 2 we noted that whole brain DTI based features have the potential to discriminate PD subjects from controls. We use the available longitudinal DTI data in the analysis presented in Chapter 8. Some studies have shown reduced cerebral blood flow in the cortex in PD subjects compared to controls using ASL data (Fernández-Seara et al., 2012). Though we did not perform any analysis using the ASL data in this thesis, this modality presents a promising avenue for future work.

Finally, in the preliminary analysis presented in Chapter 5 we used the EEG data, available only for the baseline time-point, to perform a multi-class discrimination of the three study groups. Although this cross-sectional analysis, performed early in the course of the PhD, was unrelated to the ultimate longitudinal analysis related goals of this thesis, the results are promising and warrant further study.

Chapter 4 Preliminary Methodology

4.1 Overview

This chapter provides the methodological background for the analyses presented throughout the remainder of the thesis. Novel methodology, related to forming features using longitudinal information, is presented in Chapter 6. Here we focus on describing the essentials of the data analysis pipeline that starts with a set of raw neuroimages (themselves the products of a combination of scanner hardware and software that makes up the “data generation pipeline” that is MRI), applies a series of preprocessing steps and subsequently trains a classifier to predict a diagnostic state. One particular pattern classification algorithm, the Support Vector Machine (SVM), is described in detail as it used in most of the discriminations we perform. We discuss the metrics used to evaluate the predictive performance of binary classifiers as well as the permutation testing procedure used to assess the statistical significance of classifier performance. Finally we discuss how one can derive useful visualizations (i.e. brain maps) from a linear classifier.

4.2 The Data Analysis Pipeline

The analysis pipeline is depicted in Figures 4.1 and 4.2. It has three basic components: image preprocessing, classification within a cross-validation scheme and performance evaluation. Preprocessing involves applying a series of operations that includes image segmentation, image registration and quality control. The purpose of these often complicated and time consuming procedures is to extract meaningful, discriminative information, termed “features” in pattern recognition, from raw imaging data. Image segmentation is used to separate a structural image into constituent tissue class images such as grey matter, white matter or cerebrospinal fluid (CSF). Image registration transforms data into a common coordinate system. It can be performed either within subject (i.e. longitudinally) or between subjects (i.e. cross-sectionally) using either raw, unsegmented images or tissue class images derived from segmentation. Other preprocessing operations are also possible using region-based and network-based approaches (Wolfers et al.,

2015). Quality control (QC), often performed by visually inspecting images, is used at one or more steps of this process. QC eliminates or adjusts for any corrupted or otherwise unsuitable images within a dataset.

Following preprocessing, classification is performed using the extracted features as training and testing data and their associated diagnostic labels (e.g. diseased or healthy at a certain point in time) as class labels. The classifier attempts to learn a rule that best separates the two or more classes of the training data. This is referred to as classifier training. The trained classifier is then used to predict the class labels of the testing data, which must be completely withheld during the model training process. Many algorithms have been devised to solve this often challenging problem; Bishop, (2007) and Hastie et al., (2011a) introduce and explain many commonly used approaches. We will describe one of the most widely used algorithms, the support vector machine (SVM), in Section 4.4.2 as we rely on it in many of the classifications we perform. Classification is typically performed within a cross-validation scheme with the generalization performance of the classifier evaluated by comparing the predicted class labels, aggregated across all cross-validation ‘folds’, to their corresponding known true class labels. Cross-validation is explained in Section 4.4.3 and depicted in Figure 4.1 while performance metrics are discussed in Section 4.7.

The analytic framework we describe is very general: it is used throughout the neuroimaging literature for pattern recognition based classification. There are many good explanations of it in the literature; see, for example, Lemm et al., (2011), Klöppel et al., (2012) and Wolfers et al., (2015). Although we do not discuss it further in this thesis, the same pipeline can also be used for regression, which involves learning a function of the input features that describes continuous measures of interest (such as age or a measure of cognitive function), rather than discrete class labels.

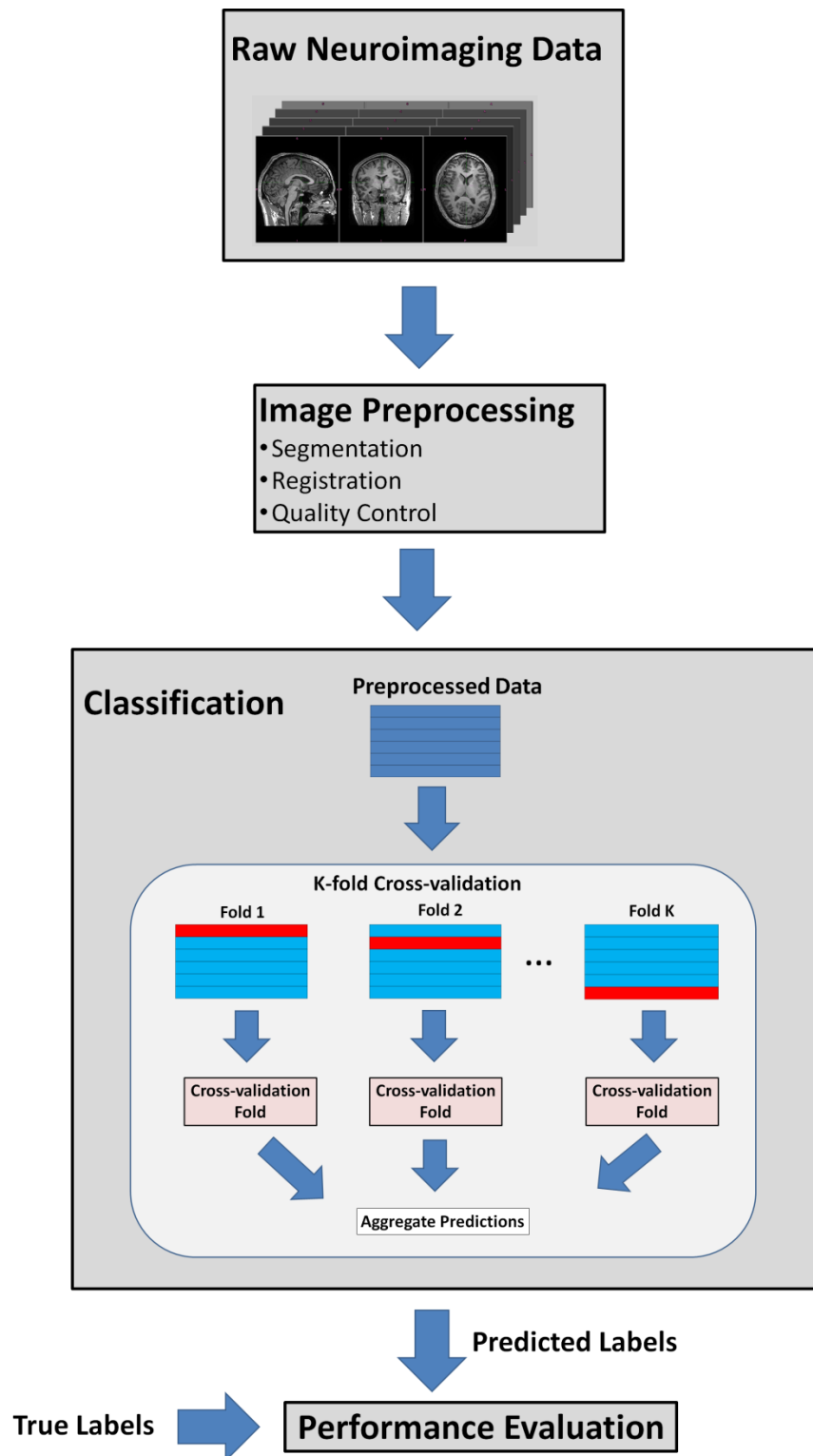


Figure 4-1 The data analysis pipeline we employ to evaluate the predictive performance of a neuroimage-based classification algorithm. All data is initially preprocessed using a combination of segmentation, registration and quality control procedures. Following this, K-fold cross-validation is performed by splitting the preprocessed data into K parts. A classifier is then trained and tested K times: for each of these cross-validation ‘folds’ one out of K parts of the data is withheld as testing data, with the remaining parts used as training data. The overall performance of the classifier is evaluated by aggregating the predicted class labels from across all cross-validation folds and comparing them to the true class labels. The classifier training and testing performed at each cross-validation fold is described in Figure 4.2.

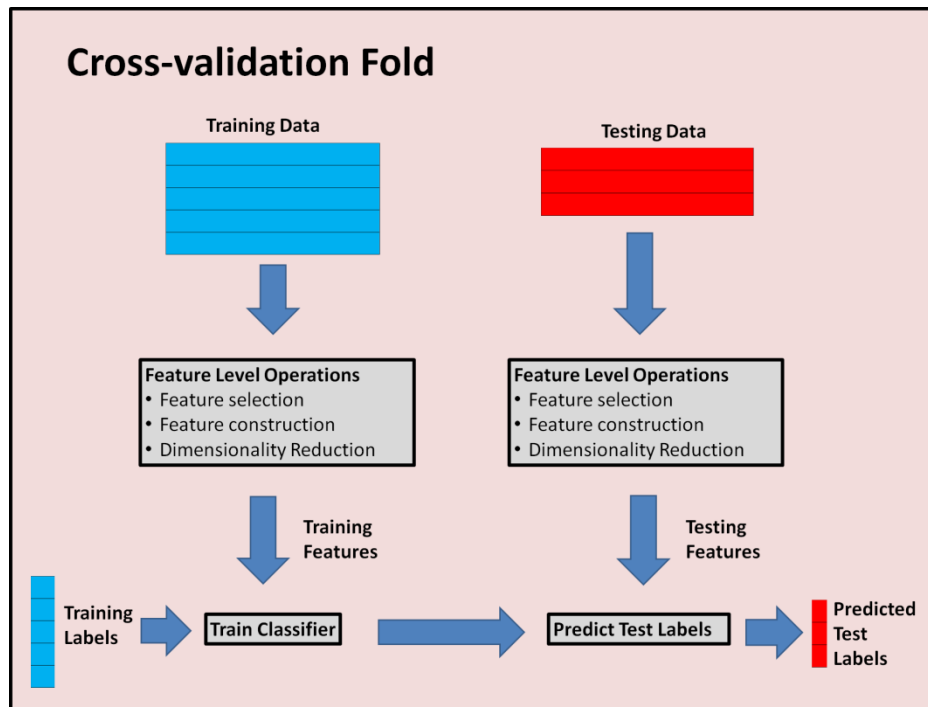


Figure 4-2 The procedure performed at each cross-validation fold (with explanation in Section 4.4).

4.3 Image Preprocessing

There are many approaches to preprocessing neuroimaging data: the procedure chosen depends on the image modality as well as the purpose of the analysis. We focus on describing the preprocessing of structural MRI here with the goal of generating features that are useful for discriminating neurodegeneration. Even within this narrowed focus, there are many design choices to be made. Cuingnet et al., (2011), for example, uses several different preprocessing methods to form a variety of structural MRI based features for discriminating Alzheimer’s disease subjects from healthy controls. We will use whole-brain Jacobian determinant based features that are derived from either between-subject or within-subject diffeomorphic image registration procedures. Whole-brain features have been shown to have comparable predictive accuracy to region of interest (ROI) based features on Alzheimer’s discrimination tasks (Cuingnet et al., 2011) while Jacobian determinant measures have shown promise in discriminating neurodegeneration by allowing a comparison of the expansion and contraction of voxels across and within subjects (Anderson et al., 2012; Hua et al., 2009; Hua et al., 2010; Studholme et al., 2004). In Chapters 5 and 7 we use cross-sectional and longitudinal Jacobian determinant features for

discriminating both early dementia and early Parkinson's disease. In the following subsections we describe and compare these two types of features.

4.3.1 Cross-sectional Jacobian determinant features

Cross-sectional Jacobian determinants quantify the expansion or contraction undergone by each voxel of an image undergoing a diffeomorphic warp to an inter-subject template. We segmented the structural images from all time-points using SPM8's "New Segment" procedure, forming grey matter, white matter, cerebrospinal fluid (CSF), soft tissue and skull images from each T1 weighted image. The grey matter and white matter tissue class images from each time-point of each subject were registered to a study specific template using SPM's DARTEL registration procedure (Ashburner, 2007). The template consists of a grey matter tissue class image and a white matter tissue class image. We formed Jacobian determinant images from the flow fields that warp each pair of grey and white matter images to the template. DARTEL is a diffeomorphic algorithm that aligns both grey and white matter images to an average template formed using all available images. The algorithm starts with a smooth template and iteratively generates a sharper template to which images are iteratively aligned. At each iteration, both the grey matter and white matter tissue class image of each subject are registered to the template's grey and white matter images, with heavier regularisation in the early iterations, resulting in a coarse-to-fine registration across iterations. Following registration, subjects' grey matter images and (separately) their white matter image are averaged to form a new grey and white matter template for the next iteration.

4.3.2 Longitudinal Jacobian determinant features

Longitudinal Jacobian determinants quantify the expansion or contraction undergone by each voxel of an image undergoing a diffeomorphic warp from an earlier time-point to later follow-up time-point, i.e. during intra-subject registration. The registration procedure used to form such features was slightly more involved than the cross-sectional version. Intra-subject registration was performed for each subject as a first step, followed by the inter-subject segmentation and registration procedure

described above. In the first step, we coregistered each subject's baseline image and follow-up images using the intra-subject diffeomorphic registration method of Ashburner and Ridgway, (2013) that is available in SPM12b. The method involves rigidly rotating and translating all available time-points of a single subject to form an average position template that approximates the exponential barycentre of all time-points. Following rigid registration, a nonlinear diffeomorphic registration is performed to warp each time-point's image to the template. In this thesis we used the pairwise version of this algorithm, always registering the baseline and follow-up images of a subject. We formed Jacobian determinant images from the flow fields warping each time-point's image to the template. In the second step we created a common inter-subject space by performing the cross-sectional segmentation and registration procedure we have described the average position brain (i.e. template) of each subject. Using the flow fields from the second step, we warped the intra-subject Jacobian determinant images to the common inter-subject template, bringing these images into a common space.

4.3.3 Cross-sectional versus longitudinal Jacobian determinants

It is important to point out that the longitudinal Jacobian determinants formed via longitudinal registration may better account for whole-brain atrophy because the longitudinal registration algorithm starts with a rigid-body transformation that rotates and translates whole head (unsegmented) serial longitudinal images without scaling them. The flow fields from the diffeomorphic registration consequently incorporate any scaling that is associated with degeneration over time. Rigidly aligning whole head images also helps to reduce the effects of scanner calibration drift, i.e. a change in MR scanner's magnetic field gradients over time. This drift can stretch the size of voxels at follow-up time-points: one study showed an artefactual 2% increase in volume between baseline and 21 month follow-up within a normal 32-year-old man (Freeborough et al., 1996). The effect can be somewhat reduced or eliminated through routine MR machine servicing, although service tolerances may permit variation of up to several percent in each dimension, which are, unfortunately, similar to rates of brain atrophy. At the analysis stage, voxel dimensions can be

rescaled by noting that, in developed adults, the size and shape of the skull usually does not change over time. Several widely-used longitudinal registration algorithms, namely the boundary shift integral (BSI) (Freeborough and Fox, 1997b) and SIENA (Smith et al., 2002b) have exploited this to derive the scaling factor necessary to serially register voxels within the brain by additionally constraining the registration to serially align the skull portion of the head. This explicit constraint is not present in the method of Ashburner and Ridgway, though it is similar in principle to their constraint that whole head images (which may include areas outside the skull and brain) be aligned.

In contrast to the longitudinal registration algorithm, the cross-sectional registration algorithm (DARTEL) starts with segmented grey and white matter brain tissue images that have been affinely registered to a standard (MNI) space. An affine transformation is a linear transformation that allows for rotation, translation, scale and shear. Consequently, the Jacobian determinants from the flow fields do not include a scaling component as the affine registration has accounted for it. When comparing deformations in brain structures across subjects it is usually preferable to use such unscaled Jacobian determinants as one is interested in comparing the relative size of brain structures across subjects. Later in this thesis (see Chapter 7) we register multiple cross-sectional images from each subject via this procedure to model the changes in Jacobian determinants over time in a common inter-subject space. By doing so we can find a space of common within-subject changes (see Chapter 6), though unfortunately, registering multiple time-points in this way may not properly account for within-subject changes in scale and may therefore underestimate whole-brain atrophy.

4.3.4 Whole-brain masks

We formed two different types of masks in the analyses presented in Chapters 5 and 7, which we will refer to as the ‘dementia mask’ and the ‘Parkinson’s mask’. The dementia mask was used to mask the Jacobian determinant images used in MCI discrimination (using the HNRS dataset, see Chapter 3) and early dementia discrimination (using the OASIS dataset, see Chapter 5). The Parkinson’s mask was

used to mask the Jacobian determinant images when discriminating early stages of PD (using the PMPP dataset, see Chapter 3).

The dementia mask was a whole brain mask that excluded extracerebral voxels. Early on in the thesis, we chose to restrict our analysis to cerebral voxels when discriminating dementia as cerebral atrophy has long been a focus of dementia research. In particular we based this decision on the discrimination study of Doyle et al., (2014) and reviews such as Fox and Schott, (2004), which emphasize the importance of atrophy in the hippocampus and the expansion of CSF in the lateral ventricles. However, other discrimination studies such as Klöppel et al., (2008c) have used whole brain features that included the brainstem and cerebellum and have shown that some areas within the cerebellum may be relevant for separating AD subjects from controls. Including these regions may have led to slightly higher discrimination accuracy.

To form this mask, we segmented the MNI152 brain with the same “New Segment” procedure, then registered the resulting grey matter image to the study template’s grey matter tissue class image using FMRIB’s Nonlinear Image Registration Tool (FNIRT) (Andersson, J.L.R et al., 2007a; Andersson, J.L.R et al., 2007b). We applied the resulting warp to the Harvard-Oxford subcortical atlas (Desikan et al., 2006) that has been affine-registered to MNI152, available in FSL (Smith et al., 2004a). We then formed the mask as a binary image consisting of all atlas regions excluding the brainstem and cerebellum. In all cases the mask retained 356,365 voxels, which were vectorised to form high-dimensional features for classification.

We chose to include the brainstem and cerebellum in the Parkinson’s mask. The brainstem has long been known to be a crucial area in the study of this disease (Braak et al., 2003) while some more recent studies have begun to point out the importance of the cerebellum (Wu and Hallett, 2013). To form a whole brain mask we used the cross-sectional template from the inter-subject registration stage, which contained grey matter and white matter tissue class images. We retained voxels that had a value of at least 0.05 intensity in either the grey matter image or white matter image, reflecting a high confidence that the voxel contains information from one of these two classes. The mask retained 416,681 voxels in this case.

4.4 Classification

The essential components necessary to perform classification are shown in Figure 4.2 for a given cross-validation ‘fold’ (see Section 4.4.3). For both the training data and the testing data some combination of feature selection, dimensionality reduction and/or feature construction are often performed to produce training and testing features, although these steps are optional. Feature selection involves reducing the number of features that are presented to the classification algorithm with the goal of retaining only those features that aid in discrimination, thereby reducing the risk of overfitting a model. Feature selection can be either hypothesis driven, as in the case of restricting features to a region of interest (ROI) such as the hippocampus or medial temporal lobe based on disease-specific prior knowledge, or automatic, via a method such as recursive feature elimination (RFE) that iteratively trains a model and discards low ranking features until a termination criterion is reached. Dimensionality reduction, a concept that is related to feature selection, involves finding a lower dimensional representation of a set of features. Finally, feature construction refers to the creation of more complex features based on simpler ones and may involve both feature selection and dimensionality reduction. Mwangi et al., (2014) provides a review of these techniques in neuroimaging.

Following these operations a classifier is trained using the training features as inputs and their associated class labels as targets. The trained classifier is used to predict the test labels with the testing features as inputs. Although somewhat mitigated by feature selection and dimensionality reduction, the features used in neuroimaging are often very high dimensional (being voxel-based) with the feature dimension often vastly exceeding the number of available samples. For such ill-posed problems, regularization based classifiers such as the SVM are necessary to avoid overfitting. Regularization, discussed in Section 4.4.2, involves imposing additional assumptions when training a classifier.

In addition to regularization, the high-dimensional features in neuroimaging have also led to the preference for algorithms that operate on the matrix of pairwise similarities between samples (termed a “kernel matrix” in pattern recognition and machine learning) rather than operating directly on the features themselves. The

advantage of such kernel algorithms is that their computational complexity depends on the number of samples rather than the number of features. Additionally, kernels allow for the mapping of a low dimensional problem into a higher dimensional feature space where classes may be better separated. In some problem domains this is an important feature, though in this thesis the space we will be working with is already of very high dimension (being vectors of hundreds of thousands of brain voxels), so we will rely exclusively on the linear kernel, i.e. the inner product between two vectors, which preserves the original dimensionality of the features. We will describe the advantages of the linear kernel in the following section.

4.4.1 Linear Kernels

There are three advantages to using a linear kernel: it introduces no additional tuning parameters, it is easily and quickly computed and, as we will see, it allows for easy interpretation of a classifier's weight vector and 'forward' map as a masked brain image. In the SVM equations, the kernel enters into the equations as a pairwise comparison of two vectors, i.e. $k(\mathbf{x}, \mathbf{x}')$, where $\mathbf{x}$ and $\mathbf{x}'$ are $d \times 1$ vectors in the input space. When using a linear kernel with row vectors, this becomes $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$ and we see that we can easily compute the similarity between sets of training or testing features matrices by forming, in a single step, a kernel matrix $\mathbf{K} = \mathbf{X}^T \mathbf{X}'$, where $\mathbf{X}$ and $\mathbf{X}'$ may be a training or testing feature matrix.

4.4.2 Support Vector Machines

This section describes one particular kernel based classification algorithm in some depth. Although the methods we develop in this thesis are generic (being at the level of feature construction) and can be used with any pattern recognition algorithm, we choose to use the support vector machine (SVM) classifier as it has been validated as an effective classifier for neuroimage based disease diagnosis (Klöppel et al., 2012; Orrù et al., 2012). In addition, the SVM classifier relies on several important ideas that are common to many other modern pattern recognition algorithms, namely kernels, regularization and sparsity. We will explain the role of each in the SVM

algorithm and hopefully provide some intuition as to why they are important in general.

SVMs are a family of supervised learning algorithms that build on a simple concept: find an optimal hyperplane that is useful for classification, regression, outlier detection or clustering. In the context of two-class classification, the optimal hyperplane is the one that maximizes the “margin” between the two labelled classes. The margin is defined as the shortest distance between the hyperplane separating the classes and the closest data points from both classes. The SVM is therefore referred to as a maximum margin classifier.

In the linear case, the SVM classifier’s decision function is described by

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (4.1)$$

where $\mathbf{w}$ is a vector that is orthogonal to the hyperplane separating the two classes, $\mathbf{x}$ is a sample and b is a constant offset term that controls the distance of the hyperplane from the origin. Samples in the first class, given a label (i.e. target value t) of $+1$, are those for which $y(\mathbf{x}) > 0$; samples in the second class, given a label of -1 , are those for which $y(\mathbf{x}) < 0$. The label of a sample is therefore simply the sign of the decision function $y(\mathbf{x})$.

The vector $\mathbf{w}$ is of central importance in this equation: its direction is orthogonal to the hyperplane separating the two classes and its magnitude is inversely proportional to the size of the margin between the two classes. The SVM’s goal of maximizing the margin between classes is equivalent to finding the vector $\mathbf{w}$ with smallest Euclidean norm $\|\mathbf{w}\|$ under the condition that the training data is perfectly separated. To understand how this is so, observe that the distance of a sample $\mathbf{x}_i$ from the separating hyperplane, defined as the points at which $y(\mathbf{x}) = 0$, is

$$\frac{t_i y(\mathbf{x}_i)}{\|\mathbf{w}\|} = \frac{t_i (\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|}. \quad (4.2)$$

The margin to be maximized is the minimum of this distance across the n available training samples, i.e.

$$\max_{\mathbf{w}, b} \left\{ \min_{i=1, \dots, n} \frac{t_i(\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} \right\}. \quad (4.3)$$

We can formulate this problem as a conventional convex optimization objective function if we constrain the absolute value of the decision function (the numerator of this term) to be unity, i.e. $\min_{i=1, \dots, n} |\mathbf{w}^T \mathbf{x}_i + b| = 1$. In other words, the points closest to the decision surface will be on the $y(\mathbf{x}) = 1$ or $y(\mathbf{x}) = -1$ line (there will be at least one training sample on each line). We can then maximize term (4.3) by minimizing the denominator, i.e. the norm of the vector $\mathbf{w}$ subject to the constraint we have imposed on the numerator. We can now state the ‘primal’ form of the SVM’s objective function as:

$$\begin{aligned} & \text{minimize}_{\mathbf{w}, b}: \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to: } t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n. \end{aligned} \quad (4.4)$$

The SVM optimization problem thus framed is a constrained quadratic programming problem that can be solved with standard mathematical optimization solvers.

Figure 4.3 provides an example of two different classification hyperplanes (here, lines) that separate the given two-dimensional data points that have been assigned to one of two classes (blue crosses or red circles). In a three dimensional problem, the surface separating the classes becomes a plane and in general it is referred to as a ‘hyperplane’ in higher dimensions. Note that the separating hyperplane is always one dimension smaller than the overall dimensionality of the feature space. In the figure there are two such hyperplanes that each perfectly separates the given training data, shown as solid lines of green and grey, with corresponding orthogonal vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ associated with each line. These vectors are referred to as the ‘weight vectors’ of a decision function of the form (4.1), because they specify the contribution of each dimension of a sample to the overall decision of which side of the decision plane (i.e. which class) the sample belongs to. In the figure, the dashed lines are the margin associated with the two decision functions. Observe that the margin of the green hyperplane is wider than that of the grey hyperplane’s. Correspondingly, the weight vector $\mathbf{w}_1$ has a smaller magnitude than $\mathbf{w}_2$, as depicted by the perpendicular arrows orthogonal to each of the

hyperplanes. In such a linearly separable problem there will be an infinite number of such hyperplanes that perfectly separate two classes. However, if we assume in this case that the green line is the maximum margin classifier, then the weight vector $\mathbf{w}_1$ associated with this hyperplane has the smallest magnitude among all possible separating hyperplanes' weight vectors. Loosely speaking, this weight vector is optimal because it is the “smoothest”, i.e. it has the smallest possible amplitude of all the separating hyperplanes.

The figure also depicts the ‘support vectors’ of the maximum margin classifier (the decision function that uses the $\mathbf{w}_1$ vector) as darkened crosses and a darkened circle. These are data points lying on the margin, i.e. where $y(\mathbf{x}) = 1$ or $y(\mathbf{x}) = -1$ for a given data point $\mathbf{x}$ and carry all information relevant to the decision problem. We will see that the SVM has the appealing property that the optimal weight vector can be expressed as a linear combination of these support vectors that lie on the margin, allowing us to ignore the other training samples when evaluating the decision of unseen (testing) data.

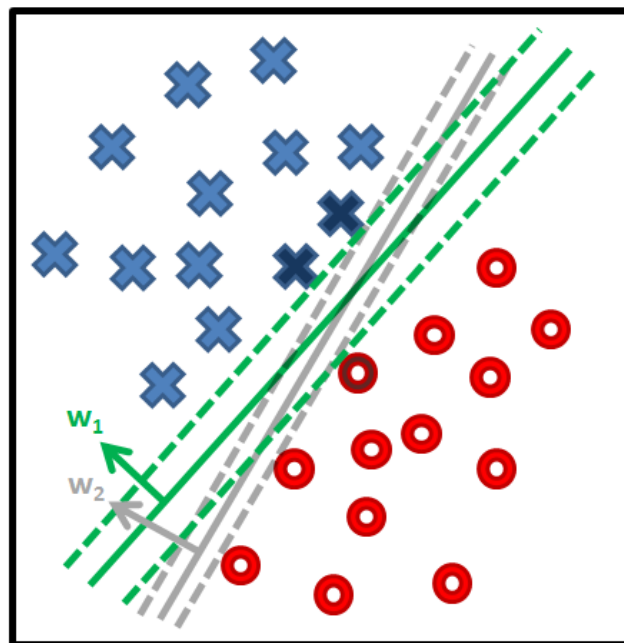


Figure 4-3 A linearly separable problem with two hyperplanes (here, lines) that perfectly separate the two classes (one marked by x's, one by o's). We see that the green hyperplane has a larger margin and a correspondingly smaller weight vector w_1 than the grey hyperplane's weight vector w_2 . In this case the green line is the maximum margin classifier. The support vectors of this classifier, the term used for the samples lying on the margin, are shown by the darkened x's and o's.

When the training samples are not linearly separable the optimization problem of (4.4) will not have a solution. In this case there will be one or more samples that prevent the clean separation of the classes with a hyperplane. This is a realistic scenario: classification is often corrupted by some form of noise, whether on the input side (uncertainty in the exact locations of samples) or, more insidiously, on the output side (uncertainty in the labels of samples that causes some to be on the “wrong” side of the decision surface, potentially deep within the other class’ territory). For such cases we can introduce ‘slack variables’ that allow samples to be misclassified by modifying the SVM’s optimization problem. Rather than the ‘hard margin’ SVM of (4.4) that does not permit misclassifications, these slack variables can be thought as creating a ‘soft margin’ that permits samples to stray within the margin or even into the other class’ territory. The primal form for the soft margin SVM is:

$$\text{minimize}_{\mathbf{w}, b}: \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (4.5)$$

$$\text{subject to: } t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \quad i = 1, \dots, n$$

where the ξ_i ’s are non-negative ‘slack’ variables. In the hard margin scenario, we imposed the constraint that $t_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$, in other words the decision function output $y(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b$ multiplied by the label t_i (either +1 or -1) must be greater or equal to one, meaning no sample can be closer than the margin marked by the $y(\mathbf{x}) = 1$ and $y(\mathbf{x}) = -1$ contour lines. Now we loosen this constraint by allowing samples to be a distance $1 - \xi_i$ away from the margin, where $\xi_i \geq 0$. When $\xi_i > 1$ we see that the distance from the margin is negative, meaning the sample is misclassified, while when $0 < \xi_i \leq 1$ the sample is within its class’ margin.

The minimization problem is now a trade-off between maximizing the margin (by minimizing the $\frac{1}{2} \|\mathbf{w}\|^2$ term) and minimizing the sum of the violations of the margin (the $C \sum_{i=1}^n \xi_i$ term). The constant C controls this trade-off (breaking notation, this scalar is capitalized to correspond with the literature); as such it is referred to as a regularization parameter. Regularization, a fundamental concept in pattern recognition, refers to the imposing of assumptions on the form of a solution

to avoid, for example, the scenario of overfitting a classifier to training data with subsequently poor performance on testing data. Here, a large C places more emphasis on punishing violations and less emphasis on maximizing the margin, i.e. less regularization, which leads to a smaller margin and tighter fit to the data. A small C emphasizes a maximization of the margin and puts less emphasis on punishing violations, i.e. more regularization. C can be thought of as an “inverse” regularization parameter and the ‘hard margin’ SVM of (4.5) can be thought of as a ‘soft margin’ SVM with C set to infinity, meaning no violations are allowed, i.e. no regularization.

Before we discuss the details of how the optimization problem of (4.5) is solved, we should emphasize that a wider margin is preferable as intuition suggests that a larger separation between classes in training (at the expense of some margin violations) should lead to a better generalization to unseen data. Several theoretical justifications for such intuition are provided in Section 7.2 of Schölkopf and Smola, (2002), the simplest of which starts by assuming that the training samples and testing samples (those unknown to the classifier at training) are drawn from the same distribution, such that the testing samples lie “close” to the training samples in the sample space. We further assume that the testing samples are generated by adding a bounded amount of noise to the training samples and that we know this bound on the noise. It follows that if we can perfectly separate the training samples with a margin at least as large as the known bound on the noise then we can still perfectly separate the testing samples. Consequently, a larger margin allows for larger amounts of uncertainty in the location of test samples while maintaining class separability. Theorem 7.3 of Schölkopf and Smola, (2002), based on the work of Bartlett and Shawe-Taylor, (1998), provides a more formal upper bound on the prediction error of a margin based classifier (i.e. the probability of misclassifying a test sample), showing that this error can be minimized by maximizing the margin (minimizing $\|\mathbf{w}\|$) while keeping the ‘margin error’ (the fraction of samples for which $t_i y(\mathbf{x}_i) < 1$) low. We can immediately recognize that these criteria are very similar to the ones being traded off in the optimization problem of (4.5), implying that the soft margin SVM formulation is seeking the optimal classifier in some strict sense.

To solve the constrained optimization problem of (4.5) we use a Lagrange multiplier for each constraint, resulting in the following Lagrange function (i.e. Lagrangian)

$$L(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \{t_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i\} - \sum_{i=1}^n \mu_i \xi_i \quad (4.6)$$

where $\alpha_i \geq 0$, $\mu_i \geq 0$ for $i = 1, \dots, n$ are the Lagrange multipliers (the Lagrangian L is a scalar). According to Theorem 6.26 of Scholkopf and Smola, (2001) we seek a saddle point of L such that the function is minimized with respect to $\mathbf{w}$ and b and maximized with respect to the Lagrange multipliers (the α_i 's and μ_i 's). At the solution the partial derivatives with respect to the primal variables must vanish, so that

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \xi, \alpha, \mu) = 0,$$

$$\frac{\partial}{\partial b} L(\mathbf{w}, b, \xi, \alpha, \mu) = 0.$$

The first partial derivative above leads to the condition that the optimal weight vector be expressed as a weighted combination of samples, with the α_i 's determining the magnitude of the contribution of each sample to the weight vector and the class label t_i (+1 or -1) determining the sign, so that

$$\mathbf{w} = \sum_{i=1}^n \alpha_i t_i \mathbf{x}_i \quad (4.7)$$

while the second partial derivate leads to the condition that the sum of these signed weights ($\alpha_i t_i$'s) be zero, so that

$$\sum_{i=1}^n \alpha_i t_i = 0. \quad (4.8)$$

Scholkopf and Smola, (2001) offer a lucid mechanical analogy for these two conditions, namely that equation (4.8) asserts that at the solution the sum of forces exerted on the hyperplane by the support vectors (i.e. those samples for $\alpha_i > 0$) must be conserved i.e. each support vector $\mathbf{x}_i$ exerts a force of magnitude α_i in the $t_i \mathbf{w} / \|\mathbf{w}\|$ direction, with the sum of these forces across support vectors equalling zero. Equation (4.7), on the other hand, asserts that the sum of the torques exerted on the hyperplane by the support vectors also sums to zero due to $\sum_i \mathbf{x}_i \times t_i \alpha_i \mathbf{w} / \|\mathbf{w}\| = \mathbf{w} \times \mathbf{w} / \|\mathbf{w}\| = 0$, where $\times$ is the cross product operation used in calculations of torque. In this way, solving the optimization problem can be viewed as finding this point of mechanical stability.

Expressions (4.7) and (4.8) allow us to eliminate the primal variables $\mathbf{w}$ and b from the optimization and state the ‘dual’ form of the optimization problem (4.5):

$$\begin{aligned} & \text{maximize}_{\alpha}: W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j \mathbf{x}_i^T \mathbf{x}_j \\ & \text{subject to: } 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n \\ & \quad \sum_{i=1}^n \alpha_i t_i = 0. \end{aligned} \tag{4.9}$$

As before, this optimization problem is framed as a standard quadratic programming problem, which, in this case, is usually solved using sequential minimal optimization (SMO) (Platt, 1998a).

We have transformed the problem of minimizing an objective function of the primal variables into maximizing an objective function of the dual variables (the support vector weights α_i). The conditions $0 \leq \alpha_i \leq C$ are required as the α_i ’s are Lagrange multipliers in the primal problem, so it must be the case that $\alpha_i \geq 0$. On the other hand, the additional condition that the derivative of (4.6) with respect to the other primal Lagrange multipliers must also vanish, i.e. $\frac{\partial}{\partial \mu_i} L(\mathbf{w}, b, \xi, \alpha, \mu) = 0$, leads to $\alpha_i = C - \mu_i$, which implies that $\alpha_i \leq C$.

It can be shown that the optimization problem in the form (4.5), with dual form (4.9), must satisfy the Karush-Kuhn-Tucker (KKT) conditions, which requires

that $\alpha_i(t_i y(\mathbf{x}_i) - 1 + \xi_i) = 0$ for every training sample at the solution (see (Bishop, 2007)). This constraint can be met in two ways: when $\alpha_i = 0$ we have that $t_i y(\mathbf{x}_i) \geq 1 - \xi_i$ and thus the sample is well classified or when $\alpha_i > 0$ we have that $t_i y(\mathbf{x}_i) = 1 - \xi_i$ and thus the sample is either on the margin ($t_i y(\mathbf{x}_i) = 1$, slack variable $\xi_i = 0$) or within the margin ($t_i y(\mathbf{x}_i) = 1 - \xi_i$, slack variable $\xi_i > 0$). Thus we see via (4.7) that all well classified samples do not contribute to the classifier's weight vector and that the classifier relies on a small set of samples, termed 'support vectors', that lie on or within the margin to make its predictions. As a result, the SVM is said to be *sparse* due to its reliance on these support vectors that are a subset of the full training data, though it should be pointed out that the entries of the weight vector that it computes are not in general sparse.

We can make two important observations with respect to this dual optimization problem. The first is that this soft margin dual formulation is the same as the hard margin dual formulation with the exception that there is no upper limit of C on the magnitude of the α_i 's in the hard margin case. The hard margin case is thus the same as the soft margin case with $C = \infty$. The soft margin will therefore saturate any large values of α_i at C , rather than letting them grow without bound to minimize the objective function. As the α_i 's are support vector weights, this corresponds to a cap on the maximum contribution of a particular training sample to the weight vector (see equation (4.7)).

The second observation is that the training samples (the $\mathbf{x}_i$'s and $\mathbf{x}_j$'s) only enter into the optimization problem via the inner product $\mathbf{x}_i^T \mathbf{x}_j$. This critical insight allows the SVM algorithm to create non-linear decision functions by replacing this inner product with a kernel function evaluation, $k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)^T \boldsymbol{\phi}(\mathbf{x}_j)$, that is equivalent to first mapping the training samples into another (potentially higher dimensional, non-linear) 'feature space' ($\mathbf{x}_i \rightarrow \boldsymbol{\phi}(\mathbf{x}_i)$ and $\mathbf{x}_j \rightarrow \boldsymbol{\phi}(\mathbf{x}_j)$) and then performing the inner product in the feature space. For problems that cannot be separated with a linear decision surface (even with the aid of a soft margin) this 'kernel trick', as it is referred to, offers the possibility of finding a higher dimensional space in which the classes can be separated. Note that the SVM is still linear in the feature space. Note as well that even if one uses a linear kernel, i.e. one

does not induce non-linearity, the kernel matrix representation is a convenient and compact representation of training and test data.

The ‘kernelized’ soft margin SVM optimization problem can now be stated as

$$\begin{aligned} & \text{maximize}_{\alpha}: W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j t_i t_j k(\mathbf{x}_i, \mathbf{x}_j) \\ & \text{subject to: } 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i t_i = 0. \end{aligned} \tag{4.10}$$

By equation (4.7) we see that the SVM’s weight vector is a linear combination of weighted support vectors. Combining (4.7) with equation (4.1) the SVM’s decision function is

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^n \alpha_i t_i \mathbf{x}_i^T \mathbf{x} + b \tag{4.11}$$

and again we observe that the function output is an inner product between the support vectors (i.e. training samples for which $\alpha_i > 0$) and the input sample, which may be a training or testing sample. The kernelized decision function used with (4.10) is therefore

$$y(\mathbf{x}) = \sum_{i=1}^n \alpha_i t_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{4.12}$$

with the corresponding weight vector defined as

$$\mathbf{w} = \sum_{i=1}^n \alpha_i t_i \boldsymbol{\phi}(\mathbf{x}_i). \tag{4.13}$$

Notice that one does not necessarily need to compute the weight vector in this way. One simply has to solve for the support vector weights via (4.10) to make predictions via (4.12), in both cases using the kernel function $k(\mathbf{x}', \mathbf{x})$ to compute the similarity between $\mathbf{x}'$ and $\mathbf{x}$ in the feature space without having to explicitly map into the

feature space (by computing $\Phi(\mathbf{x}_i)$). In this sense the SVM is able to *implicitly* map inputs into a higher dimensional space, provided the kernel used possesses some essential properties set forth by Mercer's theorem (Vapnik, 1998), such as a positive semi-definite (PSD) kernel matrix. Kernels have drawn a lot of attention in the pattern recognition and machine learning community: the books by Scholkopf and Smola, (2001) and Shawe-Taylor and Cristianini, (2004) provide comprehensive coverage.

We have shown how an SVM is a sparse (in the sense that the solution can be summarised using only a few data points, i.e. the support vectors), regularized, kernel based algorithm. Thanks to these properties the SVM is able to make predictions efficiently (via sparsity), better generalize to unseen data (via regularization), and offer the flexibility of non-linear decision surfaces (via kernels). The SVM is often very effective for separating classes though it should be noted that in the form we have presented, which is the most widely used, the algorithm is a deterministic discriminative method rather than a probabilistic generative method, such as, for example, a Gaussian Process Classifier (GPC). As such the SVM does not model the distribution of the training data (being discriminative rather than generative) and cannot quantify the uncertainty of its predictions (being deterministic rather than probabilistic).

4.4.3 Cross-validation

We have described how an SVM learns its decision function, i.e. trains its model, given a set of training data and training labels. In pattern recognition, the quality of a trained model is measured by its ability to make accurate predictions on unseen data, i.e. testing data and associated class labels that have been completely withheld from the model building process. As is customary in pattern recognition, we use K-fold cross-validation (CV) to assess the generalization ability of our trained classifiers. The procedure, depicted in Figure 4.1, involves splitting the available data into K equally sized 'folds' of the data. A classifier is then trained K times, each time a different fold of the data is used such that one of the K parts is held out as the testing data and the rest of the data (K-1 parts) are used as training data for the classifier. The overall generalization ability of the classifier is estimated by collecting the

predictions from each fold and comparing the predicted values across all folds (labels in classification or continuous values in regression) to the known true values. Five-fold or 10-fold CV is often used, particularly when there are a large number of samples. When one has a small number of samples, as is often the case in neuroimaging in general (and in this thesis in particular), one can use the maximum number of folds by setting $K = n$, where n is the number of samples. This is referred to as Leave One Out CV (LOO-CV) as at each fold just one sample is left out for testing with the rest of the $n - 1$ samples used for training. See section 7.10 of Hastie et al., (2011) for more information on CV.

Although LOO-CV is widely used in neuroimaging, where sample sizes are small, research has recently shown that it may yield highly variable estimates of generalizability. Repeated random splits, leaving out 20% of the data for testing for example, results in better estimates with less computation (Varoquaux et al., 2017). We did not perform the latter procedure in this thesis as we were not aware of it at the time of our analyses: certainly, it is an important consideration in future work.

4.5 Performance Evaluation

One can form a confusion matrix using the predicted class labels that have been aggregated across all CV folds. Figure 4.4 depicts the confusion matrix for the two class classification case: given the true labels and the predictions one can calculate the number of subjects correctly classified as positive (true positives, i.e. TP), the number of subjects incorrectly classified as positive (false positives, i.e. FP), the number of subjects correctly classified as negative (true negatives, i.e. TN) and the number of subjects incorrectly classified as negative (false negatives, i.e. FN). Using these elements, we can derive the commonly used classifier performance metrics that we report in this thesis: sensitivity, specificity, balanced accuracy, positive predictive value (PPV), negative predictive value (NPV) and a summary of the receiver operating characteristic's area under the curve (ROC AUC).

Sensitivity, also called the true positive rate or recall, is the proportion of true positives (+1 class) that were correctly identified: $TP/(TP + FN)$. Specificity, also called the true negative rate, is the proportion of true negatives (-1 class) that were

correctly identified: $TN/(TN + FP)$. Balanced accuracy, introduced by Brodersen et al., (2010), is the average of sensitivity and specificity: $(\text{sensitivity} + \text{specificity})/2$. Balanced accuracy is less sensitive to imbalanced class sizes than the overall accuracy, defined as $(TP + TN) / (TP + TN + FP + FN)$ and as such we report it throughout this thesis rather than reporting the overall accuracy. PPV, the proportion of disease state (+1 class) predictions which are correct, is computed as $TP/(TP + FP)$. NPV, the proportion of healthy state (-1 class) predictions which are correct, as $TN/(TN + FN)$. The ROC AUC plots the true positive versus false positive rate as the classifier's decision threshold is varied and is considered to be a robust measure of classifier performance in the sense that it is independent of an arbitrary choice of decision threshold.

actual condition	positive	TP	FN
	negative	FP	TN
		positive	negative
		predicted condition	

Figure 4-4 Confusion matrix for a two-class classification problem, showing total number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) predictions made by a classifier.

4.6 Significance Testing

Permutation tests can be performed to assess the statistical significance of any of these performance metrics relative to chance. Permutation testing is preferred, as parametric tests such as the binomial test, have been shown to overestimate the significance of classification accuracy when cross-validation is used whereas permutation testing is unaffected by cross-validation (Noirhomme et al., 2014). We use permutation testing to assess the statistical significance of the balanced accuracy metric in this thesis. Each permutation test involves randomly rearranging the order of the elements in the vector of n subjects' class labels $[y_1, \dots, y_n]^T$ to form a random vector of class labels that retains the original number of subjects in each class. Predicting these random class labels allows us to build up null distributions of balanced accuracies. A p-value can be estimated by dividing the number of balanced accuracies in the null distribution that exceed the true balanced accuracy by the total number of permutations. We will present results using null distributions built with 1000 permutations of class labels (Jockel, 1986; Nichols and Holmes, 2002; Mourão-Miranda et al., 2005).

4.7 Model Interpretation

4.7.1 Weight maps

Recalling equation (4.13), we have that the SVM weight vector $\mathbf{w}$ is expressed as

$$\mathbf{w} = \sum_{i=1}^n \alpha_i t_i \boldsymbol{\phi}(\mathbf{x}_i) = \sum_{i=1}^n \bar{\alpha}_i \boldsymbol{\phi}(\mathbf{x}_i)$$

where $\boldsymbol{\phi}(\mathbf{x}_i)$ is the feature vector in the space implicitly defined by the chosen kernel and the $\bar{\alpha}_i$'s are the signed versions of the support vector weights ($\bar{\alpha}_i = \alpha_i t_i$) from the solution of the SVM optimization problem in equation (4.9). As we use a linear kernel, we have that $\boldsymbol{\phi}(\mathbf{x}_i) = \mathbf{x}_i$. Equation (4.13) can then be expressed as

$$\mathbf{w} = \mathbf{X}^T \bar{\boldsymbol{\alpha}} \quad (4.14)$$

where $\mathbf{X}$ is a mean-centred version of a training feature matrix as we mean centred both training and test input data using the training data mean prior to their use as features in classification. $\bar{\boldsymbol{\alpha}}$ is the $n \times 1$ vector of $\bar{\alpha}_i$'s with entries corresponding to the rows of $\mathbf{X}$.

In this thesis the feature we use will be masked brain images. The weight vector $\mathbf{w}$, which has the same dimensionality as the features, we will therefore be referred to as a “weight map”: as a masked brain image that can be visualized to understand which parts of the brain are being relied on by the classifier when making diagnostic predictions.

4.7.2 Forward maps

We can also build a “forward map” by the method described in Haufe et al., (2014), who showed that given a linear classifier of the form $\tilde{\mathbf{y}} = \mathbf{X}\mathbf{w}$ where $\mathbf{X}$ is an $n \times d$ feature matrix, one can always find a corresponding forward map

$$\mathbf{a} = \Sigma_{\mathbf{X}} \mathbf{w} \Sigma_{\tilde{\mathbf{y}}}^{-1}. \quad (4.15)$$

where $\tilde{\mathbf{y}}$ is an $n \times 1$ vector of classifier output function values (i.e. unthresholded class label predictions), $\Sigma_{\mathbf{X}}$ is a $d \times d$ covariance matrix of features and $\Sigma_{\tilde{\mathbf{y}}}^{-1}$ is the inverse of the (scalar) covariance of the classifier output across subjects. Recalling

equation (4.1), the SVM prediction for a single input $\mathbf{x}$ is $y = \mathbf{w}^T \mathbf{x} + b$ so we can put the predictions into the form $\tilde{\mathbf{y}} = \mathbf{X}\mathbf{w}$ by subsuming the constant term b into the weight vector. We will carry on referring to $\mathbf{X}$ and $\mathbf{w}$ in this section with the understanding that we are actually referring to $[\mathbf{X} \ \mathbf{1}]$ and $\begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$ respectively in the case of the SVM, where $\mathbf{1}$ is an $n \times 1$ column of 1's.

We can observe that in (4.15) the term $\Sigma_{\tilde{\mathbf{y}}}^{-1}$ is a positive scalar value when there is a single response variable and so, assuming $\mathbf{X}$ is mean centred, we have $\mathbf{a} \propto \Sigma_{\mathbf{x}} \mathbf{w} = (\mathbf{X}^T \mathbf{X}) \mathbf{w} = \mathbf{X}^T \tilde{\mathbf{y}}$, which is equivalent to the expression

$$\mathbf{a} \propto \text{cov}(\mathbf{X}, \tilde{\mathbf{y}}) \quad (4.16)$$

where $\mathbf{a}$ is also $d \times 1$ and can be visualized as a masked brain image. We will refer to (4.16) as the forward map. The weight map defined by equation (4.14) is useful for understanding the contribution of each feature towards the prediction of a sample's diagnostic label. As discussed in Haufe et al., (2014), interpretation of such maps should be done with care as a feature may have a high weight by virtue of a group difference or as a result of high collinearity between features, e.g. features may obtain a high weight to cancel out noise in other features. In contrast, the encoding weights from the forward map in equation (4.16) represent the group differences between classes, which are often of interest when interpreting a trained classifier. If the data are standardized, the forward maps are equivalent to 'structure coefficients' widely used in multiple linear regression (Kraha et al., 2012). When discriminating a positively labelled (disease) class from a negatively labelled (control) class, stronger positive values in a forward map indicate a stronger association of a region with the positive class while negative values indicate a stronger association with the negative class (alternatively, a weaker association with the positive class).

Haufe et al. also point out that when using ordinary least squares (OLS) regression in 'backward' models (i.e. from voxels to labels), we can show, by plugging $\mathbf{w} = \Sigma_{\tilde{\mathbf{y}}}^{-1} \text{cov}(\mathbf{X}, \mathbf{y})$ into (4.15) that $\mathbf{w}$ is proportional to $\mathbf{a}$ and in this case the forward map does not improve the interpretability of the classifier. However, as the SVM's weight vector $\mathbf{w}$ is not calculated via OLS regression (there is

regularization in addition to a different type of error function), this situation does not apply.

4.7.3 t-stat Maps

In addition to this forward map we will also build a ‘t-stat map’, a particular type of statistical parametric map (SPM), by performing a two-sample t-test between disease class and healthy class subjects’ features independently at each feature dimension. These mass univariate maps should be similar to the forward maps, as in both cases we are visualizing the group differences at each voxel. We are not aware of a consensus in the research community as to whether forward maps offer any significant benefits over such mass univariate comparisons. As a result, we will provide visualizations for both t-stat maps, forward maps and weights maps where possible.

Chapter 5 Preliminary Analysis

5.1 Overview

This chapter covers preliminary analysis done early in the course of this PhD project, before the development of the analytical methodology presented in the next chapter. From a methodological perspective the analyses presented in this chapter are not in themselves novel, but this chapter instead makes a contribution from an applied perspective. Specifically, this chapter: (i) presents several attempts at discriminating early neurodegeneration (both Alzheimer’s and Parkinson’s diseases) using cross-sectional and longitudinal structural MRI based features derived from widely available registration procedures and (ii) shows the difficulty these features have in discriminating early neurodegeneration. This motivates both the methodological development and applications of this thesis and future extensions.

We make use of the two datasets described in Chapter 3: the Heinz-Nixdorf RECALL substudy (i.e. HNRS) that contains longitudinal imaging information for subjects with mild cognitive impairment (MCI) and demographically matched healthy controls and the Progression Markers in the Premotor Phase (i.e. PMPP) study of Parkinson’s disease (PD), which contains longitudinal imaging information for subjects in the early stages of PD, those identified to be at high risk of developing PD and demographically matched controls. Some of the analysis presented here appears later in Chapter 7; in particular the discrimination of MCI subjects from controls using features derived from cross-sectional and longitudinal registration will be later compared to our proposed longitudinally projected cross-sectional features. We also attempt pairwise and multi-class discriminations of the PMPP study’s three subject groups using structural MRI and EEG data.

5.2 Materials and Methods

5.2.1 Datasets

The design and data collection in both the HNRS and PMPP studies as well as the diagnostic criteria by which study subjects were grouped are described in Chapter 3. Here we briefly describe the imaging portions of these datasets as they pertain to this chapter.

HNRS dataset: structural MRI

The structural MR images in this dataset were sagittal 3D Fast Low Angle SHot (FLASH) T1 weighted images (TR = 40 ms; TE = 5 ms; flip angle = 40°; matrix size = 256 × 256 with 176 1.0 mm thick slices; FoV = 26 mm × 26 mm; bandwidth = 160 Hz/pixel), acquired on a single 1.5T MR scanner (Magnetom Avanto, Siemens Healthcare, Erlangen) equipped with a 12-channel receive-only matrix head coil provided by the vendor. In addition, fluid-attenuated inversion recovery (FLAIR) and T2 weighted sequences were also acquired for standard radiological examination.

The MRI portion of the dataset consisted of 252 subjects, with all structural MR images for all 252 images from the baseline time-point and 124 images from the first (2.5 year) follow-up time-point available for analysis⁸. We performed a manual quality control procedure that involved visually inspecting all images for distortion due to subject motion, signs of stroke or white matter abnormalities. Images that did not meet these quality control criteria were removed from further analysis. Following this procedure there were 220 baseline images and 104 follow-up images (a total of 324 images), within which there were 54 subjects that had both a baseline and follow-up image, so that we retained 54 baseline and 54 follow-up images. Within this set of 54 subjects, 24 were diagnosed as MCI at the follow-up time-point and 23 were diagnosed as healthy controls. We discarded a small set of subjects diagnosed as MCI at baseline that converted to dementia (two subjects) at first follow-up along with subjects marked as healthy controls at baseline that were subsequently labelled

⁸ The HNRS is ongoing: not all first follow-up samples had been collected at the time of our analysis.

“pre-MCI” (five subjects) at first follow-up. Discrimination was restricted to the 24 MCI subjects and 23 healthy controls for whom baseline and follow-up images were available (a total of 94 images) so that we can compare classification using either cross-sectional or longitudinal features in subjects that have the necessary longitudinal information. The analysis in Chapter 7 includes many of the remaining subjects that have images from only one time-point: those having either a baseline image but no follow-up or a follow-up image but no baseline (see Section 7.3.2). As we will show, an important benefit of the methods developed in this thesis is the ability to include subjects that have only partial or even no longitudinal follow-up in a full longitudinal analysis.

PMPP dataset: structural MRI

The structural MR images in this dataset were sagittal magnetization-prepared rapid gradient-echo (MPRAGE) T1-weighted images (TR = 2300 ms; TE = 3 ms; TI = 1100 ms; flip angle = 8°; matrix size = 256 × 224 with 176 1.0 mm thick slices; FoV = 256 × 224 mm), acquired on a Siemens MR system operating at 3T (Magnetom Trio, Siemens Healthcare, Erlangen). Baseline and second (two-year) follow-up time-points’ images from 12 PD subjects, 29 high risk subjects and 13 controls were used. All images were visually inspected for quality control purposes (following the same criteria described above for the HNRS data) and deemed acceptable for further analysis.

PMPP dataset: Electroencephalography (EEG)

We also analysed EEG data that was collected for a subset of the PMPP study’s subjects. These data, only collected at the study’s baseline time-point, were available for 15 early PD subjects, 18 at high risk of PD and 13 healthy controls. For each subject, EEG time-series were collected for each of the 21 nodes located on subjects’ scalps as shown in Figure 5.1. For each node we performed a spectral analysis and calculated the mean logarithmic power for six frequency bands, five of which we used to form features as discussed in Section 5.2.3.

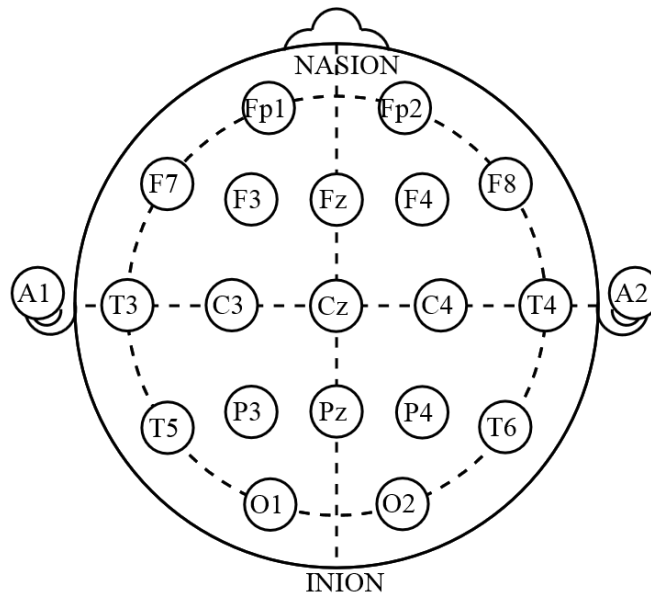


Figure 5-1 Location of scalp electrodes in EEG data that was collected, following the international 10-20 system.

5.2.2 Group Demographics

We analysed the demographic and clinical measures separately for each of the three analyses performed in this chapter.

Table 5.1 shows the demographic and clinical measures for the 47 HNRS subjects at their follow-up time-point, along with the longitudinal changes in clinical measures. There were no significant group differences in age ($F_{1,45} = 2.2$, $p = 0.14$), gender (Fisher's exact test: $p = 0.772$), years between scans (i.e. follow-up time) ($F_{1,45} = 0.1$, $p = 0.71$), but there was a significant difference in education level ($F_{1,45} = 8.4$, $p < 0.01$). As expected, there was a significant difference in cognitive function at follow-up (ADAS-cog, $F_{1,45} = 40.0$, $p < 0.001$) but no significant difference in its change over time (ADAS-cog change, $F_{1,45} = 2.0$, $p = 0.17$).

Table 5.2 provides similar information for the 54 PMPP subjects used in the structural MRI based analysis. There were no significant group differences in age ($F_{2,51} = 0.3$, $p = 0.73$), gender (Fisher's exact test: $p = 0.13$) or follow-up time ($F_{2,51} = 0.6$, $p = 0.56$). There were significant differences in UPDRS-III total sum ($F_{2,51} = 84.6$, $p < 0.001$), tremor (UPDRS-III sum of items 3.17 and 3.18, $F_{2,51} = 25.5$, $p < 0.001$).

0.001) and rigidity (UPDRS-III sum of items 3.3, $F_{2,51} = 57.8$, $p < 0.001$) across the three groups, but no significant differences in their change over time (UPDRS-III: $F_{2,51} = 2.5$, $p = 0.10$, tremor: $F_{2,51} = 0.5$, $p = 0.62$, rigidity: $F_{2,51} = 0.9$, $p = 0.41$).

Finally, Table 5.3 provides information from just the baseline time-point for the subset of 46 PMPP subjects used in the EEG analysis. The group differences are, of course, similar to those in Table 5.2: no significant differences in age ($F_{2,43} = 0.5$, $p = 0.6$), gender (Fisher's exact test: 0.33) with significant differences in UPDRS-III ($F_{2,43} = 74.7$, $p < 0.001$), tremor ($F_{2,43} = 15.2$, $p < 0.001$) and rigidity ($F_{2,43} = 29.1$, $p < 0.001$).

Table 5-1 Demographics of structural MRI based discrimination at follow-up, HNRS dataset

HNRS MRI (Follow-up + long. change)	Healthy (n = 23)	MCI (n = 24)
Male/Female	13/10	12/12
Age	72.6 ± 4.4	70.3 ± 5.8
Years Between Scans	2.6 ± 0.6	2.6 ± 0.5
Education	2.2 ± 0.7	1.6 ± 0.6
ADAS-Cog	6.1 ± 2.8	11.7 ± 3.2
ADAS-Cog change	-0.9 ± 2.5	0.3 ± 3.1

Table 5-2 Demographics of structural MRI based discrimination at follow-up, PMPP dataset

PMPP MRI (Follow-up + long. change)	Healthy (n =13)	HR_{PD} (n = 29)	PD (n =12)
Male	6 (46.2%)	22 (75.9%)	7 (58.3%)
Female	7 (53.8%)	7 (24.1%)	5 (41.7%)
Age	65.9 ± 7.3	64.2 ± 5.1	64.6 ± 8.1
Years between scans	2.0 ± 0.4	2.4 ± 1.9	1.9 ± 0.3
SN+	0 (0%)	29 (100%)	11 (91.7%)
UPDRS-III	0.8 ± 2.0	2.0 ± 2.4	24.8 ± 10.9
UPDRS-III change	0.6 ± 2.1	-1.7 ± 3.7	-4.9 ± 11.9
Tremor	0.2 ± 0.8	0.1 ± 0.4	3.3 ± 2.8
Tremor change	0.2 ± 0.8	0.0 ± 0.5	-0.2 ± 2.2
Rigidity	0.0 ± 0.0	0.0 ± 0.0	3.6 ± 2.2
Rigidity change	0.0 ± 0.0	-0.4 ± 0.8	0.2 ± 2.7

Table 5-3 Demographics of EEG based discrimination at baseline, PMPP dataset

PMPP EEG (Baseline)	Healthy (n =13)	HR_{PD} (n = 18)	PD (n =15)
Male	6 (46.2%)	13 (72.2%)	9 (60%)
Female	7 (53.8%)	5 (27.8%)	6 (40%)
Age	63.8 ± 7.2	61.8 ± 5.0	64.1 ± 9.2
SN+	0 (0%)	18 (100%)	14 (93.3%)
UPDRS-III	0.2 ± 0.6	2.8 ± 2.5	26.6 ± 11.1
Tremor	0.0 ± 0.0	0.0 ± 0.0	2.9 ± 2.9
Rigidity	0.0 ± 0.0	0.4 ± 0.7	3.1 ± 1.9

5.2.3 Image Preprocessing

Structural MRI: Jacobian Determinant features

Jacobian determinant measures extracted from the flow fields warping an image to a group or subject template are promising features for studying anatomical changes due to neurodegeneration, offering increased statistical power for detecting brain atrophy compared to segmented and registered grey matter volume (Anderson et al., 2012; Hua et al., 2009; Hua et al., 2010; Studholme et al., 2004). We formed two different types of features using the structural MRI data of both datasets: cross-sectional Jacobian determinant based features and longitudinal Jacobian determinant based features. Cross-sectional Jacobian determinants quantify the expansion or contraction undergone by each voxel of an image undergoing a diffeomorphic warp to an inter-subject template. Longitudinal Jacobian determinants quantify the expansion or contraction undergone by each voxel of an image undergoing a diffeomorphic warp from an earlier time-point to later follow-up time-point, i.e. during intra-subject registration. We discuss the preprocessing procedures used to form these features in Section 4.3 of the previous chapter. In this chapter, to form cross-sectional Jacobian determinant we registered the follow-up images from the 24 MCI subjects and 23 healthy controls, forming an inter-subject template from 47 images. To form longitudinal Jacobian determinant features we performed pairwise, longitudinal registration using the baseline and follow-up images of the same subjects (94 images in total), forming 47 within-subject templates and a single inter-subject template.

EEG features

Sensory, motor and cognitive tasks are known to depend on local and long-range communication between different areas of the brain. Synchronized neuronal oscillations across different frequency ranges are thought to be a likely mechanism for these communications. Such oscillations may be disturbed in neurodegenerative diseases such as dementia and PD (Schnitzler and Gross, 2005). EEG has been used

to study abnormal neuronal oscillations, with studies showing differences between demented PD subjects, non-demented PD subjects and healthy controls using both (Fast Fourier Transform, i.e. FFT based) frequency domain and non-linear time series analysis (de Weerd et al., 1990; Soikkeli et al., 1991; Tanaka et al., 2000; Pezard et al., 2001). These studies have mostly focussed on subjects in advanced stages of PD although a recent study by Han et al., (2013) found differences in relative power between early PD subjects and age-matched controls across several frequency bands. We do not know of any studies that have explored EEG differences between at risk (preclinical) PD subjects and controls, particularly using pattern recognition methods.

The digital EEG data made available for analysis were filtered, after collection, with a time constant of 0.3 s and a 70 Hz filter. Before further analysis, the data were visually inspected to ensure that the EEG time series were free of artefacts. For each patient the first 10 artefact-free 2-sec epochs were studied. Spectral analysis of each of the 21 electrodes' time series data was performed with a Fast Fourier Transformation (FFT). The mean logarithmic power was calculated for the following frequency bands: delta (1-3.5 Hz), theta (4.0-7.5 Hz), alpha I (8.0-10 Hz), alpha II (10.5-13 Hz), beta (13.5-20 Hz) and gamma (35-45 Hz). We did not include the gamma band in our analysis to avoid any interference from power line noise. We formed feature vectors for each subject consisting of the five other frequency bands at each of the 21 nodes, resulting in 105 features per subject.

5.2.4 Classification Algorithms

We compared the performance of two different classification algorithms on the various structural MRI based discrimination problems. The Support Vector Machine (SVM) and the Gaussian Process classifier (GPC) are both kernel based algorithms that are well suited for classifications involving high-dimensional features and low sample sizes, as is often the case in neuroimaging studies. The SVM is a particularly popular discriminative classifier whose predictive performance is competitive with state-of-the-art methods but, in its most widely used form, cannot quantify predictive uncertainty. We discuss this algorithm in detail in Section 4.4.2. Here and in the

remaining chapters of this thesis we used the LIBSVM implementation of this algorithm with a linear kernel and a fixed cost parameter.

In contrast to the SVM, the GPC can make probabilistic predictions that can quantify diagnostic uncertainty. Formally, a Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen, 2006). In the context of GP based classification, a Gaussian process models a joint distribution over latent functions describing both the training and testing outputs. This Gaussian distribution has zero mean, with covariance matrices that are functions of the training and testing data points (i.e. features). Bayes rule is used to estimate the posterior distribution of the latent functions at the test points by conditioning on the known training data. In GP classification the posterior over latent functions is passed through a sigmoidal function that maps the unbounded latent function values onto the (0, 1) interval, enabling probabilistic predictions of class membership at the test points.

As in the case of the SVM (see Section 4.4.2), the training and test samples enter into the GP equations via the inner products that make up the elements of the covariance matrices that are used. We can similarly invoke the ‘kernel trick’ by replacing these inner products by a kernel function $k(\mathbf{x}, \mathbf{x}')$ that potentially performs the inner product in a higher-dimensional feature space rather than directly in the original input space of the samples. In this way we can generalize the notion of similarity to model non-linear relationships between data points. Kernels often have one or more parameters (termed hyperparameters) that are tuned by maximizing the marginal likelihood of the training data, typically using a gradient descent based optimization procedure that avoids the need for cross-validation based tuning. Further details regarding the application of the GPC to neuroimaging data are provided by Marquand et al., (2010). The algorithm was implemented using the GPML toolbox⁹ with a linear kernel (as with the SVM).

The SVM and GPC share properties that make them well suited to neuroimaging data. As discussed in Chapter 4, both are kernel based algorithms and as such they encode the relationships between training and test data points via a kernel matrix that scales with the (typically small) number of samples rather than the

⁹ GPML toolbox is available at <http://www.gaussianprocess.org/gpml/code/matlab/doc>

dimensionality of the data points, which can be very large when using voxel-based features. Despite different formulations, both algorithms trade off model fit and model complexity. The SVM is framed as an optimization problem that maximizes the margin between classes while minimizing margin violations. It maximizes the margin via an l_2 -norm penalty on the solution it seeks, meaning it punishes the square root of the sum of squares of the solution's elements. The l_2 -norm penalty has been shown to perform competitively when the discriminative pattern relies on highly correlated features, e.g. when discriminating diseases that have a large, spatially distributed biological footprint, such as the effects of Alzheimer's disease on cortical thickness (Sabuncu and Konukoglu, 2014). In the case of the GPC the marginal likelihood similarly trades off model fit and model complexity. The model complexity term is slightly different in this case: it punishes the complexity of the covariance matrix (i.e. the kernel matrix), which is a function of the hyperparameters, rather than explicitly enforcing a particular norm on the solution. For a more detailed comparison of the two algorithms see Rasmussen and Williams, (2006), p.113 and p. 144. In practice GPC's and SVM's often have similar predictive performance on neuroimaging problems (Marquand et al., 2010).

Another commonly used regularization penalty is the l_1 -norm, which enforces sparsity, i.e. it (approximately) minimizes the number of features that contribute to discrimination by punishing the sum of the absolute values of the elements of the solution. We hypothesized that sparsity may be a desirable property for the (temporal) frequency domain based features that were derived from the PMPP study's EEG data. We used a classification algorithm that combined both l_1 -norm and l_2 -norm penalties, for situations where either a small subset of features or a set of correlated features drive classification. Relative to the L1 norm, the elastic net makes the solution more stable in the case of collinear features and is more interpretable in that it permits correlated features to enter the model. We chose the Sparse Multinomial Logistic Regression (SMLR) algorithm, first introduced by Krishnapuram et al., (2005) as a multi-class classifier with an l_1 -norm penalty and later extended to handle a combination of l_1 -norm and l_2 -norm penalties by Ryali et al., (2010). We used Dr. Marquand's MATLAB implementation of Ryali et al.'s algorithm, which maximizes a multinomial likelihood function with l_1 -norm and l_2 -norm penalties on the solution. The weights on each of these penalties were tuned

via nested cross-validation based grid search. We varied the l_1 -norm weight in logarithmic steps from 10^{-4} to 1 (higher values discarded too many features) and varied the l_2 -norm weight in logarithmic steps from 10^{-4} to 10^4 .

5.2.5 Classifier Performance Evaluation and Significance Testing

In Chapter 4 we used the confusion matrix to derive the classifier performance metrics we report here for the two class classification case (see Section 4.5), namely the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), balanced accuracy and the area under the receiver operating characteristic curve (ROC AUC). As we also perform multi-class classification in this chapter we can derive similar metrics from an m -class confusion matrix. Figure 5.2 depicts such a confusion matrix. We define the sensitivity for class i as $S_i = C_{i,i} / \sum_{j=1}^m C_{i,j}$, i.e. the number of correct class i predictions divided by the total number of class i samples. The predictive value for class i is defined as $PV_i = C_{i,i} / \sum_{j=1}^m C_{j,i}$, i.e. the number of correct class i predictions divided by the sum of all class i predictions. In the multi-class case the balanced accuracy is a generalization of the two-class version (defined as the average of the sensitivity and specificity). It defined as the average class sensitivity: $\frac{1}{m} \sum_{i=1}^m S_i$. Similarly, the overall predictive value (OPV) is the average predictive value across classes: $\frac{1}{m} \sum_{i=1}^m PV_i$.

In Chapters 7 and 8 we use permutation testing to assess the statistical significance of binary classifier balanced accuracies, as described in Section 4.6. In these preliminary analyses we perform a simpler (and significantly faster) Monte Carlo based test of the balanced accuracy metric, as described in Marquand et al., (2013). The approach involves forming a vector of predicted class labels by permuting the indices of the vector of true class labels. By comparing the predicted class labels to the true class labels we can generate a random confusion matrix with the same class distribution as the true sample. For each such matrix the statistic of interest is derived (here the balanced accuracy). The p-value is derived by computing the proportion of randomized balanced accuracies that exceeded the true balanced accuracy. This general approach is used to assess the significance of both the two-

class and multi-class (here, three-class) discriminations presented in this chapter. We generated 5000 random matrices in each case.

Although this Monte Carlo procedure is non-parametric, it may be biased as it does not account for cross-validation (Noirhomme et al., 2014). As a result, the p-values generated by this procedure may be unrealistically low, though in most cases the accuracies tested here will still not be statistically significant.

true	class 1	$C_{1,1}$	$\dots$	$C_{1,m}$
	$\dots$	$\vdots$	$\begin{smallmatrix} \cdot & & \\ & \cdot & \\ & & \cdot \end{smallmatrix}$	$\vdots$
	class m	$C_{m,1}$	$\dots$	$C_{m,m}$
		class 1	$\dots$	class m
		predicted		

Figure 5-2 Confusion matrix for the general case of an m-class classification problem, with the cell at row i, column j containing the number of samples predicted as having class j whose true class is class i.

5.3 Results

Tables 5.1 and 5.2 show classifier performance metrics for binary classifications using both cross-sectional and longitudinal Jacobian determinant based features derived from structural MRI. Based on the Monte Carlo testing of the balanced accuracies we can conclude that in most cases these classifiers do not discriminate subjects any better than chance level, which stands at 50% for binary classification. Discriminating HR_{PD} subjects from controls using cross-sectional Jacobian determinants is the only exception, where we see significant accuracies using both the SVM ($p < 0.05$) and GPC ($p < 0.01$). In no case do we see significant discrimination using longitudinal Jacobian determinant features. A paired t-test on the balanced accuracies in Table 5.5 revealed no significant differences between cross-sectional and longitudinal features across the three discriminations and two classifiers ($t_5 = -1.6$, $p = 0.18$).

Table 5.6 shows the result of the simultaneous multi-class discrimination of the three study groups in the PMPP study using spectral power based features derived from EEG, with the accompanying confusion matrix in Figure 5.3. In this case the balanced accuracy, the average of the sensitivity for the three classes in this case, is significantly higher than the 33% chance level for three-class classification ($p < 0.01$).

Table 5-4 HNRS Dataset, structural MRI based discrimination of 24 MCI subjects vs. 23 Healthy Controls (HC) subjects

Features/Classifier	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
Cross-sectional Jacobian Det.'s						
<i>SVM</i>	53.2	54.2	52.2	54.2	52.2	0.639
<i>GPC</i>	59.7	54.2	65.2	61.9	57.7	0.641
Longitudinal Jacobian Det.'s						
<i>SVM</i>	59.6	58.3	60.9	60.9	58.3	0.574
<i>GPC</i>	55.5	45.8	65.2	57.9	53.6	0.549
* Statistically significant balanced accuracy, Monte Carlo p-value < 0.05						
** Statistically significant balanced accuracy, Monte Carlo p-value < 0.01						
*** Statistically significant balanced accuracy, Monte Carlo p-value < 0.001						

Table 5-5 PMPP Dataset, structural MRI based pairwise discriminations of 3 study groups: 12 PD subjects, 29 HR_{PD} subjects, 13 HC subjects

Features/Classifier	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
PD (n=12) vs HCs (n=13)						
Cross-sectional Jacobian Det.'s						
<i>SVM</i>	35.6	25.0	46.2	30.0	40.0	0.276
<i>GPC</i>	31.7	25.0	38.5	27.3	35.7	0.269
Longitudinal Jacobian Det.'s						
<i>SVM</i>	58.7	25.0	92.3	75.0	57.1	0.506
<i>GPC</i>	59.0	33.3	84.6	66.7	57.9	0.590
HR _{PD} (n=29) vs HCs (n=13)						
Cross-sectional Jacobian Det.'s						
<i>SVM</i>	62.3*	86.2	38.5	75.8	55.6	0.666
<i>GPC</i>	67.4**	65.5	69.2	82.6	47.4	0.703
Longitudinal Jacobian Det.'s						
<i>SVM</i>	50.8	86.2	15.4	69.4	33.3	0.668
<i>GPC</i>	59.2	41.4	76.9	80.0	37.0	0.602
PD (n=12) vs HR _{PD} (n=29)						
Cross-sectional Jacobian Det.'s						
<i>SVM</i>	47.3	8.3	86.2	20.0	69.4	0.362
<i>GPC</i>	34.9	25.0	44.8	15.8	59.1	0.342
Longitudinal Jacobian Det.'s						
<i>SVM</i>	58.3	16.7	100.0	100.0	74.4	0.629
<i>GPC</i>	59.5	50.0	69.0	40.0	76.9	0.635
* Statistically significant balanced accuracy, Monte Carlo p-value < 0.05						
** Statistically significant balanced accuracy, Monte Carlo p-value < 0.01						
*** Statistically significant balanced accuracy, Monte Carlo p-value < 0.001						

Table 5-6 PMPP Dataset, EEG based discriminations of 3 study groups: 15 PD subjects, 18 HR_{PD} subjects, 13 HC subjects

Features/Classifier	Balanced Accuracy (%)	Sensitivities (%)	PVs (%)	OPV (%)
EEG nodes' spectral power				
<i>SMLR</i>	55.4**	PD: 67 HR _{PD} : 61 HC: 38	PD: 67 HR _{PD} : 50 HC: 56	57.4

* Statistically significant balanced accuracy, Monte Carlo p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo p-value < 0.01

*** Statistically significant balanced accuracy, Monte Carlo p-value < 0.001

true	PD	10	5	0
	HR _{PD}	3	11	4
	HC	2	6	5
		PD	HR _{PD}	HC
		predicted		

Figure 5-3 Confusion matrix for SMLR based classifier of EEG data based features (corresponding to Table 5.6), each cell containing the number of predictions made.

5.4 Discussion

We have attempted to discriminate early neurodegeneration using both cross-sectional and longitudinal features derived from structural MRI. We chose to use Jacobian determinant (JD) features due to their potential to detect brain atrophy associated with neurodegeneration (Hua et al., 2010). However, we did not perform a systematic comparison of JD features with other commonly used features, such as normalized grey matter volume or vertex based features from cortical surfaces. This decision was based in part on studies such as Cuingnet et al., (2011), which compared the use of ten different structural MRI based features to discriminate different stages of Alzheimer's disease (AD). The study found comparable predictive performance across many of the derived features. Furthermore, our aim was to compare the use of cross-sectional and longitudinal information in a similar manner: between subject and within subject JD features enabled such a comparison.

The low discrimination accuracies seen in Table 5.4 on the MCI versus healthy control problem are consistent with much of the AD discrimination literature discussed in Chapter 2. Figure 2 of Chu et al., (2012), for example, shows that with a similar number of samples (~50) MCI subjects were discriminated from controls with less than 60% accuracy using whole-brain (cross-sectional) features. We were somewhat surprised to see no significant discriminations using longitudinal Jacobian determinant features, which quantify subtle within-subject changes that may aid in discriminating early neurodegeneration. However, this finding is also consistent with our literature review (specifically Section 2.4.6), which observes that longitudinal information is often not better and may even be worse at discriminating neurodegeneration than cross-sectional information. Indeed, the difficulty of this discrimination inspired the development of the feature construction method we present in the next chapter; an attempt to make better use of longitudinal information to improve this discrimination.

The discriminations presented in Table 5.5 did not in general exceed chance level. There was one exception however: a significant accuracy in discriminating HR_{PD} subjects from controls using cross-sectional features across both classifiers. As noted in Section 3.2, the HR_{PD} group was selected based on substantia nigra hyperechogenicity (SN+) along with one or two markers associated with early PD. Although the rate of incidence has been shown to be 17 times higher in elderly persons with SN+ compared to those with SN- (Berg et al., 2011), only one subject within the HR_{PD} group converted to PD and overall there were few significant changes in prodromal markers over the course of this relatively short (two year) study (Liepelt-Scarfone et al., 2017). Given the evidence, it is not certain that this group truly represents prodromal PD: it may be the case that we could discriminate the HR_{PD} group using MRI because these subjects were selected based on an imaging abnormality (albeit using TCS).

We found no significant difference in accuracy between cross-sectional and longitudinal features across discriminations, which is in line with the PD discrimination literature we reviewed in Chapter 2 (see Section 2.5.1): only a few studies to date have been able to discriminate PD subjects from controls using T1-weighted MRI. In light of these findings, we concluded that the other imaging modalities available in this study may be more promising. As diffusion MRI data

were available from this study, we decided to use whole-brain diffusion MRI based features in Chapter 8, applying our proposed longitudinal feature construction method to this modality instead.

Across all pairwise discriminations presented in Tables 5.4 and 5.5 we see the SVM and GPC algorithms performing comparably. The SVM is significantly (roughly five times) faster than the GPC however: to perform the two classifications in Table 5.4 with the SVM took 13 seconds for each, while the GPC took around 70 seconds for each. The difference is due to the optimizations being performed: in the case of the SVM, thanks to Sequential Minimal Optimization (SMO) (Platt, 1998b) as implemented in LIBSVM (Chang and Lin, 2011a), a very efficient two-variable problem is solved that avoids the need to solve a more general quadratic programming problem. In the case of the GPC, a slower conjugate gradients based approach is used to find the hyperparameters that maximize the marginal likelihood of the data given the chosen model structure. In our case we kept the SVM's regularization parameter fixed ($C = 1$); optimizing such parameters via nested cross-validation may tilt the comparison of computation times in favour of the GPC, which avoids the need for nested cross-validation using the aforementioned method of finding optimal hyperparameters.

Although the computation times are reasonable for both algorithms in the analyses presented here, when performing nested cross-validation based parameter tuning (as in chapters 7 and 8) we must train significantly more models. In Chapter 7, for example, we revisit this discrimination problem using nested leave one out cross validation (LOOCV, discussed in Section 4.4.3) to tune the percentage of variance explained by the retained principal components, choosing among 19 parameter values (from 5% to 95% explained variance). Under nested cross-validation (see Section 6.12) we train $47 \times 46 \times 19 = 41,078$ models, leaving one of 47 subjects out at each outer fold, then one of the 46 subjects out in the inner fold. In this case the difference in computation time between the two algorithms, measurable in hours, led to the decision to use the SVM exclusively. It should be pointed out that nested cross-validation may not properly estimate the generalization ability of a classifier (see Section 4.4.3) and that generalization ability should be tested using separate datasets, such as the PPMI dataset.

The analysis of the EEG data we present here was performed early in the course of the PhD project, before much of the subsequent longitudinal methodology was developed. Unfortunately, we did not have longitudinal data for this modality and could not form longitudinal features. Although the EEG based discrimination of Parkinson's appears to be promising and in need of further study, we did not account for group differences in movement brought that may have been brought about by tremor in the PD subjects (see Table 5.3). Later, in Chapter 8, we will correct for movement in diffusion MRI data using motion parameters from the image pre-processing pipeline.

Chapter 6 Novel Methodology

6.1 Overview

Chapter 2 reviewed Alzheimer’s and Parkinson’s disease discrimination studies, noting that in general longitudinal features have been shown to have similar or worse predictive accuracy compared to cross-sectional features in neurodegenerative disease discrimination problems. The preliminary analysis presented in Chapter 5 is in line with this conclusion, showing no significant discrimination accuracies across both diseases using longitudinal structural MRI based features. Our literature review also noted that at present, few pattern recognition based approaches exist to make explicit use of high-dimensional longitudinal information. Here we present a novel method that attempts to use such information at the feature construction level, i.e. between the image registration and pattern recognition algorithms, making it independent of the choice of either of these parts of the data analysis pipeline¹⁰.

In this chapter we assume that we have a set of longitudinal neuroimaging data, with multiple measurements (here images) available for each subject in the data set and that we are interested in making diagnostic predictions at certain cross-sectional time-points (i.e. at baseline or follow-up) using this longitudinal information. We will start by modelling the trajectories of the image voxels as polynomial functions of time (Section 6.2). We choose a basic form for these longitudinal trajectory models without coupling. In other words, we assume that is no relationship between neighbouring voxels and no coupling between subjects. We can fit these models to each subject’s voxel time series, and build up a matrix of coefficients of linear change over time (i.e. the slope terms of these models) across all subjects and voxels. In section 6.3 we show how to build these coefficient matrices in the general case of an arbitrary number of samples per subject with varying amounts of time between scans (which we refer to as an “unbalanced longitudinal design”). In Section 6.4 we consider the special case of two samples per subject, with both a fixed time interval between scans across subjects (“balanced longitudinal design”) and varying intervals (“unbalanced longitudinal design”). For this special case we show that the intra-subject difference between samples, scaled

¹⁰ An implementation of the method we describe is available at <https://github.com/LeonAksman/lpr>.

by time differences, is equivalent to the matrix of linear change coefficients. This leads to a simple method for creating linear transformations that model longitudinal changes. Using these coefficient matrices we create a transformation into a subspace describing longitudinal change that is common across subjects (Sections 6.5 through 6.8). We then discuss how to form features for classification by projecting cross-sectional samples onto this longitudinal subspace (Section 6.10). Section 6.10 also fits our proposed feature construction method within the overall data analysis pipeline, outlined in Chapter 4, that is used to make neuroimage based predictions of disease state. Finally, Sections 6.11 and 6.12 discuss how we form a linear kernel using projected features and how we use nested cross-validation to tune parameters within a cross-validation scheme.

Throughout the chapter we will use lower case italic font when referring to scalar variables, functions or constants (e.g. x), lower case bold for vectors (e.g. $\mathbf{x}$), upper case bold for matrices (e.g. $\mathbf{X}$) and upper case “curly” italic font for sets (e.g. $\mathcal{X}$). Subscripts will be in italic (e.g. $\mathbf{x}_y$).

6.2 Basic longitudinal trajectory model

Figure 6.1 depicts a simple one dimensional example of longitudinal trajectory modelling, i.e. the fitting of models that describe the time course of a set of measurements within a subject. In the figure there are nineteen measurements of a function $x(t)$ from time $t = 1$ to $t = 10$ where $x(t)$ may represent, for example, the intensity of a magnetic field in a certain brain voxel at a particular time t . We can model the temporal trajectory of these measurements by assuming a polynomial model of the form

$$x(t) = \beta^{(0)} + \sum_{j=1}^p \beta^{(j)} t^j + \varepsilon \quad (6.1)$$

where $\beta^{(0)}$ is a constant term (zeroth order coefficient), p is the polynomial function model order, $\beta^{(j)}$ is the j^{th} order coefficient and ε is i.i.d Gaussian measurement error. In the left graph we have fitted a linear function (i.e. a first order model with

$p = 1$) to the measurements, while on the right we have fitted a quadratic function (i.e. a second order model with $p = 2$) to the same data.

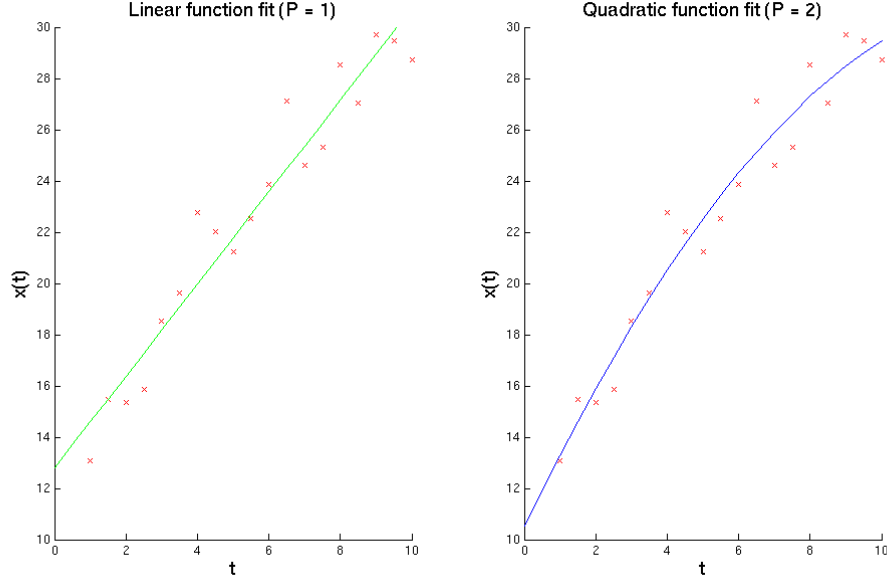


Figure 6-1 Example of fitting two different polynomial functions to a set of nineteen measurements of a function $x(t)$. The left graph fits a linear function (first order model, $p = 1$) to the data, while the right graph fits a quadratic function (second order model, $p = 2$) to the same data.

In general, we can fit such a polynomial model to the temporal trajectory of each dimension (i.e. voxel) of each subject when we have a set of high dimensional longitudinal samples (e.g. across many voxels) for a set of subjects. Formally, suppose we have a longitudinal training set $\mathcal{L}$ with l subjects such that subject i has a variable number of samples m_i over time, with total number of samples $m = \sum_{i=1}^l m_i$ across all subjects. These samples can be arranged in a temporally ascending sequence $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i})$, with each sample in the sequence being a d dimensional vector in general. A dimension in this case refers to a “channel” of measurement – each sample has d such channels which can take continuous or discrete values independently of each other.

We can build models describing the trajectories these samples follow over time, starting from the sample at the initial time, $\mathbf{x}_{i1}$, through to the sample at the last available time, $\mathbf{x}_{im_i}$. We will build these longitudinal trajectory models separately

for each dimension, assuming the same polynomial model form in Equation (6.1), which can alternatively be expressed as an inner product in the following way

$$x(t) = \mathbf{z}\mathbf{b} + \varepsilon \quad (6.2)$$

where we have formed two vectors: a longitudinal design (row) vector $\mathbf{z} = [1 \ t \ \dots \ t^p]$ and a column vector of coefficients $\mathbf{b} = [\beta^{(0)} \ \beta^{(1)} \ \dots \ \beta^{(p)}]^T$. In the Figure 6.1 example, we have $x(t) = 12.8 + 1.8t$ when we fit a linear model ($p = 1$) and $x(t) = 10.7 + 2.9t - 0.1t^2$ when we fit a quadratic model ($p = 2$), which can be expressed as $x(t) = [1 \ t][12.8 \ 1.8]^T$ in the linear case and $x(t) = [1 \ t \ t^2][10.7 \ 2.9 \ -0.1]^T$ in the quadratic case. To fit these polynomial models, we solve a system of equations with the vector $\mathbf{b}$ as the unknown variables (coefficients), the vector $\mathbf{x}$ of dependent variable samples and a longitudinal design matrix $\mathbf{Z}$ of covariate (independent variable) samples, resulting in

$$\mathbf{x} = \mathbf{Z}\mathbf{b} + \varepsilon \quad (6.3)$$

where ε is now a vector of measurement errors. We will refer to the components of the coefficient vector $\mathbf{b}$ as “trajectory coefficients” as they will be used to model voxels’ temporal trajectories.

Equation (6.3) is used to fit a polynomial model to the time course of samples along a single dimension for a given subject. However, we wish to model the time courses of all dimensions simultaneously. Fortunately, equation (6.3) can be easily generalized to all sample dimensions using a General Linear Model (GLM). Across all dimensions, for subject i with m_i samples we have a GLM of the form

$$\mathbf{X}_i = \mathbf{Z}_i\mathbf{B}_i + \mathbf{E} \quad (6.4)$$

where $\mathbf{X}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i}]^T$ is an $m_i \times d$ matrix of the subject’s samples, $\mathbf{Z}_i$ is an $m_i \times (p + 1)$ longitudinal design matrix which takes the form

$$\mathbf{Z}_i = \begin{bmatrix} 1 & t_{i1} & \dots & t_{i1}^p \\ 1 & t_{i2} & \dots & t_{i2}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \dots & t_{im_i}^p \end{bmatrix}. \quad (6.5)$$

where $t_{i1}, \dots, t_{im_i}$ are the times that correspond to samples $\mathbf{x}_{i1}, \dots, \mathbf{x}_{im_i}$ and $\mathbf{B}_i$ is a $(p + 1) \times d$ matrix of trajectory coefficients across dimensions that takes the form

$$\mathbf{B}_i = \begin{bmatrix} \mathbf{b}_i^{(0)} \\ \vdots \\ \mathbf{b}_i^{(p)} \end{bmatrix} \quad (6.6)$$

where $\mathbf{b}_i^{(j)}$ is a row vector containing subject i 's j^{th} order coefficients across d dimensions. Finally, $\mathbf{E}$ is an $m_i \times d$ matrix of i.i.d Gaussian measurement errors. In the statistics literature, the linear regression in equation (6.4) is known as a polynomial regression when $\mathbf{Z}_i$ takes the form described in equation (6.5).

In this thesis, the longitudinal samples we use are very high dimensional vectors. For example, in Chapters 5 and 7 we form samples using Jacobian determinant images that quantify the volumetric expansion or contraction of each voxel induced by a diffeomorphic warping of an image to a template. In Chapter 8, we apply this same model to fractional anisotropy (FA) images derived from diffusion tensor imaging (DTI).

6.3 Coefficient matrices in the general case

We have assumed there is no coupling between subjects, allowing us to express the model for all l subjects for whom we have the necessary longitudinal information in the form of a GLM, as

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_l \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & & & \\ & \mathbf{Z}_2 & & \\ & & \ddots & \\ & & & \mathbf{Z}_l \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_l \end{bmatrix} + \mathbf{E} \quad (6.7)$$

with $\mathbf{E}$ an $m \times d$ block diagonal matrix of subjects' i.i.d Gaussian measurement errors. As a result we can solve the GLM independently for each subject so that

$$\mathbf{B}_i = \mathbf{Z}_i^+ \mathbf{X}_i \quad (6.8)$$

where $\mathbf{Z}_i^+ = (\mathbf{Z}_i^T \mathbf{Z}_i)^{-1} \mathbf{Z}_i^T$ is the pseudo-inverse of $\mathbf{Z}_i$.

We have stated that $\mathbf{B}_i$, estimated using equation (6.8), is made up of $(p + 1)$ rows, with the $(j + 1)^{\text{th}}$ row containing the j^{th} order coefficients across all dimensions (equation (6.6)). Returning to the Figure 6.1 example with nineteen measurements from time $t = 1$ to time $t = 10$ in half unit increments, to fit a linear model ($p = 1$) we form a 19×2 dimensional longitudinal design matrix $\mathbf{Z}_i$ and a 19×1 dimensional matrix $\mathbf{X}_i$ of sample values at each corresponding time as

$$\mathbf{Z}_i = \begin{bmatrix} 1 & 1 \\ 1 & 1.5 \\ \vdots & \vdots \\ 1 & 10 \end{bmatrix}, \mathbf{X}_i = \begin{bmatrix} 10.5 \\ 11.9 \\ \vdots \\ 29.5 \end{bmatrix}.$$

We can then compute $\mathbf{B}_i$, in this case a 2×1 matrix of the subject's coefficients for the single dimension according to equation (6.8), multiplying the pseudo-inverse of $\mathbf{Z}_i$ by $\mathbf{X}_i$, resulting in $\mathbf{B}_i = \begin{bmatrix} 12.8 \\ 1.8 \end{bmatrix}$ as before. If we had two-dimensional samples instead, the longitudinal design matrix $\mathbf{Z}_i$ would not change, we would simply add the measurements along the second dimension as another column in $\mathbf{X}_i$ and compute a 2×2 matrix of coefficients $\mathbf{B}_i$ in the same way, with a column of coefficients for each dimension.

The assumed model form in (6.7) allows us to solve for the coefficients of each subject separately. We are interested, however, in understanding the nature of longitudinal changes across all available dimensions and longitudinal training set subjects so that we can find commonalities that generalize to unseen data. To do so, we will seek spatial patterns within the coefficients of the polynomial functions that we have fit. The procedure we describe in Box 6.1 shows us how to form coefficient matrices of a desired order across all individuals by solving for each individual's coefficients.

For a chosen polynomial model order p , we can form $(p + 1)$ matrices of size $l \times d$, denoted by $\mathbf{B}^{(0)}, \dots, \mathbf{B}^{(p)}$, such that $\mathbf{B}^{(j)}$ contains the coefficients of order j across all l subjects in the longitudinal training set $\mathcal{L}$. $\mathbf{B}^{(j)}$ is formed as

$$\mathbf{B}^{(j)} = \begin{bmatrix} \mathbf{b}_1^{(j)} \\ \vdots \\ \mathbf{b}_l^{(j)} \end{bmatrix} \quad (6.9)$$

where the i^{th} row contains the vector $\mathbf{b}_i^{(j)}$, corresponding to the $(j + 1)^{\text{th}}$ row of $\mathbf{B}_i$, as in equation (6.6).

Box 6.1

The procedure in Box 6.1, which merely rearranges subjects' coefficient matrices (the $\mathbf{B}_i$'s), allows us to gather coefficients of the same order across subjects (the matrices $\mathbf{B}^{(0)}, \dots, \mathbf{B}^{(p)}$) and then find patterns within a matrix of coefficients of the same order. We will focus principally on the matrix of first order coefficients across subjects, $\mathbf{B}^{(1)}$ and fit first order models for voxel trajectories ($p = 1$) throughout this thesis. This is because the datasets we have available only have a small number of time-points, which makes fitting higher order models more difficult. However, it should be clear from the foregoing that the method we propose generalises naturally to any model order. First order models are the simplest models we can fit that still allow us to understand the trend over time of a set of measurements along a dimension. The first order coefficient, i.e. the slope of the fitted model, captures this trend in a single number. In the Figure 6.1 example, when we fit a linear model, the slope of 1.8 units of output per unit of time efficiently summarizes the longitudinal trajectory of the nineteen measurements we have. In the example, the quadratic model appears to be a better fit to the samples, implying that two numbers, slope and curvature, would better summarize the longitudinal trend. Unfortunately, in neuroimaging we do not often have many measurements of a subject over time: present longitudinal studies in neuroimaging are often limited to a baseline measurement and one or two follow-up measurements. In such situations, first order models are a realistic choice that avoids overfitting a higher order model to a small number of measurements.

Following the procedure in Box 6.1, we compute $\mathbf{B}^{(1)}$ by first solving (6.8) for each subject, assuming a first order model ($p = 1$). We then take the second row of each subject's resulting coefficient matrix (the first order coefficients for that subject) and form $\mathbf{B}^{(1)}$ as in equation (6.9) when $j = 1$.

6.4 Coefficient matrices in the special case: two samples per subject

As noted, the procedure we have described above can be used in the general case of an arbitrary number of samples per subject. When the longitudinal training set is limited to a baseline and a single follow-up measurement per subject, we can derive a simpler expression for $\mathbf{B}^{(1)}$. This is an important special case in neuroimaging, where it is not unusual to have a single follow-up measurement per subject. For a longitudinal training set $\mathcal{L}$ with two time-points per subject, we can express the second time-point as $t_{i2} = t_{i1} + t_{i\Delta}$ for subject i . We can rearrange the model presented in equation (6.7) to describe the samples at a given time-point. At the baseline time-point the $l \times d$ matrix of set $\mathcal{L}$ subjects' samples, which we denote by $\mathbf{X}_{\mathcal{L}}(t_1)$, can be expressed as

$$\mathbf{X}_{\mathcal{L}}(t_1) = \mathbf{B}^{(0)} + \mathbf{T}_1 \mathbf{B}^{(1)} + \sum_{j=2}^p \mathbf{T}_1^j \mathbf{B}^{(j)} + \mathbf{E} \quad (6.10)$$

where $\mathbf{T}_1$ is an $l \times l$ matrix with subjects' first time-point values $t_{11}, \dots, t_{l1}$ along the leading diagonal and $\mathbf{E}$ is an $l \times d$ matrix of i.i.d Gaussian measurement errors. We can express set $\mathcal{L}$ subjects' samples at the second time-point as

$$\mathbf{X}_{\mathcal{L}}(t_2) = \mathbf{B}^{(0)} + (\mathbf{T}_1 + \mathbf{T}_{\Delta}) \mathbf{B}^{(1)} + \sum_{j=2}^p (\mathbf{T}_1 + \mathbf{T}_{\Delta})^j \mathbf{B}^{(j)} + \mathbf{E} \quad (6.11)$$

where $\mathbf{T}_{\Delta}$ is an $l \times l$ matrix with subjects' time intervals between samples $t_{1\Delta}, \dots, t_{l\Delta}$ along the leading diagonal. Taking the difference of equations (6.11) and (6.10) and assuming a first order model ($p = 1$) as no higher order model can be estimated in this case,

$$\mathbf{D}_{\mathcal{L}} \triangleq \mathbf{T}_{\Delta}^{-1}(\mathbf{X}_{\mathcal{L}}(t_2) - \mathbf{X}_{\mathcal{L}}(t_1)) = \mathbf{B}^{(1)} \quad (6.12)$$

where the diagonal inverse term $\mathbf{T}_{\Delta}^{-1}$ is an $l \times l$ matrix with $\frac{1}{t_{1\Delta}}, \dots, \frac{1}{t_{l\Delta}}$ along the leading diagonal. We define the matrix $\mathbf{D}_{\mathcal{L}}$ as the intra-subject (longitudinal) differences between the baseline time-point's samples and the follow-up time-point's samples, scaled by the time intervals $t_{i\Delta}$, $i = 1, \dots, l$.

If we make a stronger assumption of a balanced longitudinal design with two time-points, such that the follow-up time-point takes place a fixed time t_{Δ} after baseline for all subjects, we have that $\mathbf{T}_{\Delta} = t_{\Delta}\mathbf{I}$ with $\mathbf{I}$ as the $l \times l$ identity matrix. We can simplify the difference in equation (6.12) to

$$\mathbf{D}_{\mathcal{L}} \triangleq \frac{1}{t_{\Delta}}(\mathbf{X}_{\mathcal{L}}(t_2) - \mathbf{X}_{\mathcal{L}}(t_1)) = \mathbf{B}^{(1)}. \quad (6.13)$$

We see that in the case of a balanced longitudinal design with two samples per subject the sample difference matrix, scaled by the time interval between samples, equals $\mathbf{B}^{(1)}$.

Box 6.2

Rather than forming the $\mathbf{B}^{(1)}$ matrix via the procedure in Box 6.1, when we have two longitudinal samples per subject we can work directly with left hand side of equations (6.12) or (6.13), depending on whether the time interval between samples is fixed across subjects (i.e. a balanced design) or is subject specific (i.e. unbalanced design).

6.5 Dimensionality Reduction

We have shown how to derive the coefficient matrix $\mathbf{B}^{(1)}$ in the general case of an arbitrary number of samples per subject and the special case of two samples per subject. Each element in this matrix is the slope (first order coefficient) from the fitting of a polynomial model to a particular subject's samples along a particular dimension. It should be emphasized that each subject may have a different number of longitudinal samples yet as long as each has at least two we can fit a linear model to their voxel trajectories and add that subject as a row in the $\mathbf{B}^{(1)}$ matrix. Coefficient

matrices therefore provide a flexible means of integrating varying amounts of longitudinal information across subjects, allowing us to take the next step of finding commonalities in subjects' longitudinal changes.

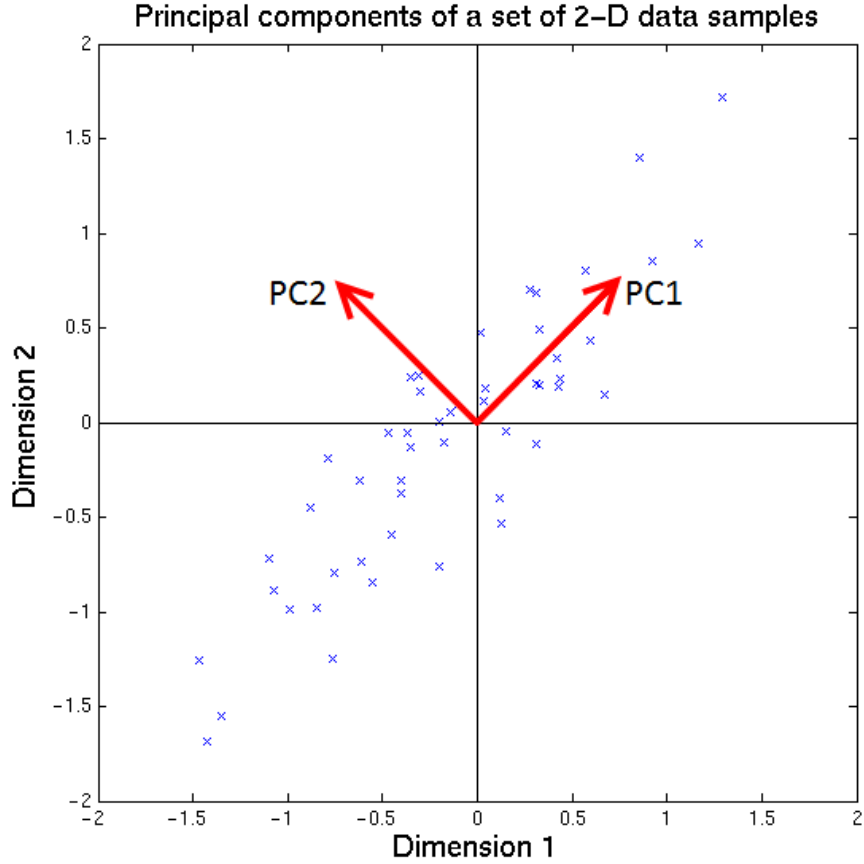


Figure 6-2 The principal components of a set of fifty two-dimensional (2-D) samples, scattered as shown.

Before proceeding, we consider a concrete example of a $\mathbf{B}^{(1)}$ matrix of size 50×2 , corresponding to the case of having estimates of linear longitudinal change (slope) across two-dimensions in a longitudinal training set of fifty subjects. We can visualize these fifty 2-D estimates as points on a plane, where each axis is a sample dimension, as in Figure 6.2. This figure was generated by assuming that each estimate is a sample drawn from a multidimensional Gaussian distribution with a given mean and covariance matrix. We are interested in the covariance matrix of this data, as it determines the “shape” of the scatter about the mean value, which we have set to a zero vector for the sake of simplicity. Each sample $\mathbf{x}$ was drawn from a 2-D Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu} = [0 \ 0]^T$ and covariance:

$$\mathbf{\Sigma} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} + 0.05 * \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.53 & 0.48 \\ 0.48 & 0.53 \end{bmatrix}.$$

The covariance matrix $\mathbf{\Sigma}$ can be thought of as a linear combination of two matrices, each a rank one covariance (sub)matrix that allows data to scatter along a particular direction. The first such matrix describes unit norm variance along the $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ direction (a diagonal in the first quadrant), the second describes 0.05 magnitude variance along the $\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ direction (a diagonal in the second quadrant). The samples we generate from this distribution have a covariance structure that is defined by the weighted combination of these two matrices. As a result, the samples in Figure 6.2 are distributed elliptically around the origin, with the major axis being the $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ direction, along which lies $\left(\frac{1}{1+0.05}\right) * 100 = 95.2\%$ of the variance and the minor axis being the $\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}^T$ direction along which the remaining $\left(\frac{0.05}{1+0.05}\right) * 100 = 4.8\%$ of the variance lies.

In general, unit norm vectors that define the principal axes (the major and minor axis in the two-dimensional case) of the scatter of a set of samples (i.e. covariance) are referred to as principal components. The unit norm vector along the major axis in this example is the first principal component (“PC1” in the figure), defined as the direction in sample space that describes the maximum amount of variance in a dataset. The unit norm vector along the minor axis is the second principal component (“PC2”), defined as the direction in sample space that describes the maximum amount of variance among all directions that are orthogonal to the first principal component.

In practice one can find this mutually orthogonal set of directions that maximally describe the variance in a set of samples of arbitrary dimension by finding the eigenvalues of the sample covariance matrix (Pearson, 1901; Hotelling, 1933). This technique, known as principal component analysis (PCA), is a cornerstone in the field of pattern recognition. In simple terms, PCA is a means of finding a set of axes that describe the scatter of a set of samples about their mean value, subject to the constraint that the axes be mutually orthogonal. It is used for dimensionality reduction, for latent variable modelling, for feature extraction and for

data visualization (Joliffe, 2002). It is also used as a first step in more advanced data decomposition methods such as independent component analysis (ICA) (Comon, 1994) and as inspiration for non-linear methods such as kernel PCA (Scholkopf et al., 1999). In Section 6.10 we use PCA to build a space of *intrinsically* lower dimension by retaining a small set of principal components that describe most of the variability of longitudinal changes across subjects.

To perform PCA on the example above, we first calculate the sample mean, which we use to compute the sample covariance matrix, which turns out to be $\begin{bmatrix} 0.43 & 0.41 \\ 0.41 & 0.49 \end{bmatrix}$, close to the true covariance matrix we sampled from. We then calculate the eigenvectors and eigenvalues of this matrix, which gives us two principal components and two values proportional to the amount of variance explained by each principal component. The eigenvectors of this sample covariance matrix are $[0.68 \ 0.74]^T$ and $[-0.74 \ 0.68]^T$ with corresponding eigenvalues of 0.87 and 0.05 respectively, so that the first eigenvector explains $\left(\frac{0.87}{0.87+0.05}\right) * 100 = 94.6\%$ of the data variance and the second explains $\left(\frac{0.05}{0.87+0.05}\right) * 100 = 5.4\%$.

Later on in this chapter we will form projections onto a lower dimensional space that explain a desired percentage of variance of a set of samples. To see how this can be done, suppose we want a linear subspace that captures at least 90% of the variance of the dataset in our example. We know that the first principal component in this example explains at least that much variance, so projecting our samples onto that one-dimensional space (a line) will accomplish our goal. To project a sample vector $\mathbf{x}$ onto the first principal component $\mathbf{u}_1$ we compute the inner product $\mathbf{x}^T \mathbf{u}_1$, resulting in a continuous scalar value that can be seen as a coordinate on the one-dimensional line represented by the principal component. The variance of all samples projected onto this line will be 94.6% of the original variance, meaning we have discarded some information but have gained a more compact representation of the data, now lying on a line instead of a plane. In this way, we reduce the dimensionality of the data from two dimensions to one, redefining the coordinate system for the projected samples (a single “new” dimension) formed as a weighted combination of the two “original” dimensions. Recall that the first principal component is the $[0.68 \ 0.74]^T$ vector, so that a unit of length along the new

dimension's direction is 0.68 units along the first original dimension and 0.74 along the second original dimension. We will refer to the lower-dimensional representations we derive from PCA as “latent spaces” or, alternatively, “linear subspaces” as, in this case, the projections from the original dimensions onto the lower dimensional representation are linear combinations of the “original” dimensions.

6.6 Eigenslopes

The application of PCA to reduce the dimensionality of images for use in pattern recognition dates back at least as far as the seminal work of Turk and Pentland, (1991), who focussed on the face recognition problem, forming features by projecting face images onto a low dimensional space that “spans the significant variations among known face images”. We will form features similarly, creating a low dimensional space spanning the significant variations among known estimates of slope (linear longitudinal change) across image dimensions, onto which we will project cross-sectional information.

We coin a similar term for the principal components of the slope matrix $\mathbf{B}^{(1)}$, calling them “eigenslopes”. These eigenslopes are the directions in (potentially high-dimensional) sample space that each describe a fraction of variation in linear longitudinal changes within the available longitudinal training data. If the longitudinal samples are masked brain images, with each sample dimension corresponding to a voxel in the brain, each eigenslope is a map of the variation, across subjects, in the longitudinal changes within each subject's brain. If we imagine each sample in the Figure 6.2 example to be a slope estimate across two voxels of a subject's images, the first eigenslope is the $[0.68 \ 0.74]^T$ vector. This direction corresponds to the situation of having roughly similar rates of change (for example, tissue progression or degeneration) along the two dimensions. The second eigenslope, the $[-0.74 \ 0.68]^T$ vector, corresponds to having roughly opposite rates of change (for example, tissue expansion along the first dimension, tissue degeneration along the second dimension or vice versa, with similar magnitudes).

It is important to point out that PCA allows for the unsupervised learning of these patterns of common longitudinal change that we have called eigenslopes. We are able to extract the eigenslopes directly from the longitudinal training data, without the need for any prior knowledge of which sample dimensions contribute to longitudinal changes.

6.7 Forming the longitudinal subspace

We have seen how we can use PCA to learn a low dimensional latent space that captures the variation in longitudinal changes across subjects. We propose that projecting cross-sectional images onto such a subspace is a natural approach to forming features for discriminating a certain class of diseases. These diseases should be marked by change over time, including (but potentially not limited to) progressive neurodegenerative disorders. The projection we will describe allows us to retain only the components of an image that lie in the space of common longitudinal changes that is encoded in the retained eigenslopes. An alternative viewpoint is that differences among subjects' brains that do not lie in the proposed space of longitudinal changes should be less discriminative of diseases that are characterized by changes over time. In simple terms, we are “projecting out” the purely cross-sectional components of subjects' images that we expect to be less discriminative of such diseases.

We will retain the k eigenslopes that explain the most variance in $\mathbf{B}^{(1)}$ to form the matrix $\mathbf{U}_k$ of size $d \times k$: a projection from d -dimensional sample space onto k -dimensional (latent) eigenslope space. Appendix A, based on Section 12.1.4 of Bishop, (2007), describes the PCA procedure we employed to compute $\mathbf{U}_k$, with the number of retained eigenslopes k chosen via nested cross-validation (described in Section 6.12). In the case of imaging features with many fewer samples than image dimensions, we will have $k \ll d$, resulting in a substantial reduction of dimensionality. In addition to the intuition we gain in thinking about this projection as being composed of set of eigenslopes that encode common longitudinal changes, using PCA has the advantages of being computationally efficient and linear, the latter property allowing for easier interpretation of results.

We refer to the PCA performed on the coefficient matrix $\mathbf{B}^{(1)}$ in the general case of an unbalanced design as *Longitudinal Trajectory Coefficient PCA (LTC-PCA)*. In this case we first solve for each subject's coefficients via (6.8) and then assemble the matrix of desired coefficients across subjects via equation (6.9). In the two-sample case, we refer to the PCA performed on the matrix $\mathbf{D}_{\mathcal{L}}$, shown by equation (6.12) to be equivalent to $\mathbf{B}^{(1)}$ for unbalanced designs and by equation (6.13) for balanced designs, as *Longitudinally Matched PCA (LM-PCA)*. In this case, we are simply differencing the two available time-points' feature matrices and scaling by fixed or subject varying time differences. In all cases, however, the projection matrix $\mathbf{U}_k$ is composed of a set of eigenslopes: the difference in terminology refers to how we form the $\mathbf{B}^{(1)}$ matrix.

6.8 Within-set prediction problem

We have described how to form projections using longitudinal information from a training set of subjects. In this and the following sections we consider two problems: making predictions using these projections when test subjects have longitudinal information and when they do not have such information. Here we describe the first problem. We will make use of several sets in this section. The longitudinal set $\bar{\mathcal{L}}$ of subjects will be defined as the set of subjects for whom the necessary number of longitudinal time-points is available. As our intention here is to use cross-validation (see Section 4.4.3) to evaluate a classifier's performance on unseen data, we define the classification set $\mathcal{C}$ as the set that is split, at each cross-validation fold, into a training set $\mathcal{N}$ of subjects used to train a classifier and a set $\mathcal{T}$ of subjects used to test the classifier's ability to generalize to unseen data. Additionally we define the longitudinal training set $\mathcal{L}$ as the set difference between $\bar{\mathcal{L}}$ and $\mathcal{T}$, $\mathcal{L} = \bar{\mathcal{L}} \setminus \mathcal{T} = \{x \in \bar{\mathcal{L}} \mid x \notin \mathcal{T}\}$, i.e. the longitudinal set subjects that are not being held out for testing in a given cross-validation fold.

When the longitudinal set $\bar{\mathcal{L}}$ and the classification set $\mathcal{C}$ are equal all subjects for whom we are interested in making predictions via cross-validation have the necessary longitudinal samples. At each cross-validation fold, we form a training set $\mathcal{N}$ and test set $\mathcal{T}$ as subsets of $\mathcal{C}$ and form the projection $\mathbf{U}_k$ using the longitudinal

information of the subjects in $\mathcal{N}$, which is equivalent to the longitudinal training set $\mathcal{L}$ in this case. This is a straightforward application of our method, which seeks to answer whether longitudinal information from a set of subjects can improve predictions for subjects within the set (Figure 6.3A).

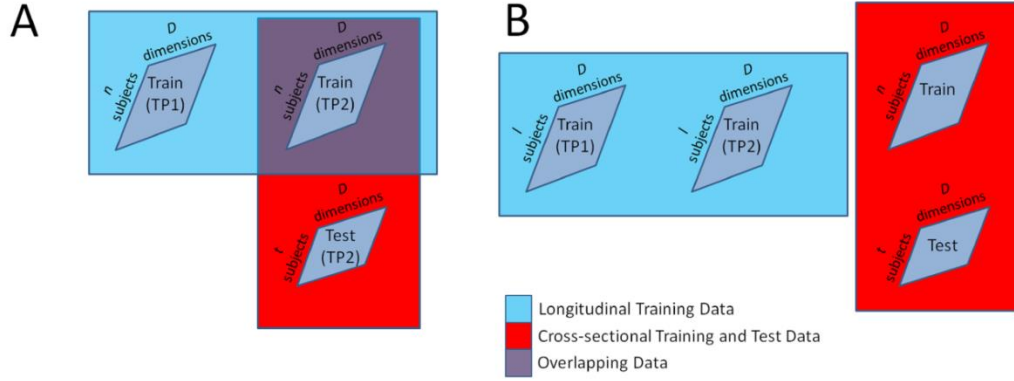


Figure 6-3 Panel A depicts the *Within-Set Prediction* problem of predicting disease state at a follow-up time-point (TP2) using longitudinal information from both two time-points. In this case the longitudinal training subject set (data in blue) matches the classification training subject set (data in the top part of the red area), with overlapping data shown in purple. Panel B depicts the *Information Transferring Prediction* problem, where the longitudinal subject set and classification subject sets are disjoint. In this case there is no overlap in data between the longitudinal data (blue area) used to form the subspace and the classification data (both training and test data, red area) being projected.

6.9 Information transferring prediction problems

In addition to the Within-Set Prediction problem our method can be used to transfer longitudinal information from one set of subjects to another. In this case the longitudinal set $\bar{\mathcal{L}}$ and the classification set $\mathcal{C}$ are disjoint, meaning they have no subjects in common. It follows that the longitudinal training set $\mathcal{L}$, used to form the longitudinal transform $\mathbf{U}_k$, is disjoint from both the training set $\mathcal{N}$ and the test set $\mathcal{T}$ at each cross-validation fold. This problem illustrates that our approach can be used to form a longitudinal subspace with one set of subjects to make diagnostic predictions for another subject set (Figure 6.3B).

6.10 Feature Formation

For both cases described above, we form training and test feature matrices, used in classification, by projecting the cross-sectional samples of the training and test subjects' samples at a particular time-point t onto a rank k longitudinal space described by $\mathbf{U}_k \mathbf{U}_k^T$. Note that the difference between the two cases described above is whether subject set $\mathcal{L}$, whose longitudinal data is used to form $\mathbf{U}_k$, is equal to or disjoint from set $\mathcal{N}$.

For either case, at each cross-validation fold the features used in classifier training and testing have the form

$$\begin{aligned}\mathbf{X}_{train} &= \mathbf{X}_{\mathcal{N}}(t) \mathbf{U}_k \mathbf{U}_k^T \\ \mathbf{X}_{test} &= \mathbf{X}_{\mathcal{T}}(t) \mathbf{U}_k \mathbf{U}_k^T.\end{aligned}\tag{6.14}$$

Here we have introduced $\mathbf{X}_{\mathcal{N}}(t)$ and $\mathbf{X}_{\mathcal{T}}(t)$ as, respectively, matrices of cross-sectional data for the training and test sets at a particular time-point t , such as at baseline or follow-up. We use $\mathbf{X}_{train}$ as an input feature matrix to train a classifier to predict $\mathbf{y}_{train} \in \{1, -1\}$, the binary class labels of the training set at a given time-point. We test the trained classifier's predictive performance using $\mathbf{X}_{test}$ as an input feature matrix to predict $\mathbf{y}_{test} \in \{1, -1\}$, the binary class labels of the test set at a given time-point. We set $\mathbf{X}_{\mathcal{N}}(t) = \mathbf{X}_{\mathcal{N}}(t_2)$ and $\mathbf{X}_{\mathcal{T}}(t) = \mathbf{X}_{\mathcal{T}}(t_2)$ when we predict the follow-up class labels using projected follow-up cross-sectional samples and $\mathbf{X}_{\mathcal{N}}(t) = \mathbf{X}_{\mathcal{N}}(t_1)$ and $\mathbf{X}_{\mathcal{T}}(t) = \mathbf{X}_{\mathcal{T}}(t_1)$ when we predict the baseline labels using projected baseline samples.

Other configurations are also possible. For example, it is possible to make prognostic predictions of the follow-up labels by projecting the baseline cross-sectional data matrices $\mathbf{X}_{\mathcal{N}}(t_1)$ and $\mathbf{X}_{\mathcal{T}}(t_1)$ onto this subspace. We have chosen to transform the features back into the original $d \times 1$ space by post-multiplying by $\mathbf{U}_k^T$, allowing an easier interpretation of the resulting model's $d \times 1$ weight vector as voxels in a masked brain image (see Sections 4.7.1 to 4.7.3). In this way the intrinsic dimensionality of the features is reduced from d to k : the features are still high-dimensional vectors but they can only vary within the small subspace spanned by the eigenslopes. Alternatively, to obtain low-dimensional feature vectors, one may simply project the baseline or follow-up features onto $\mathbf{U}_k$ instead of $\mathbf{U}_k \mathbf{U}_k^T$.

Note that each of the k vectors that make up the transform $\mathbf{U}_k$ is an eigenvector of the covariance matrix $\mathbf{\Sigma}_{\mathcal{L}} = \frac{1}{I} \sum_{i=1}^I (\mathbf{b}_i - \hat{\mathbf{b}})^T (\mathbf{b}_i - \hat{\mathbf{b}})$, where $\mathbf{b}_i$, the i^{th} row vector in the matrix $\mathbf{B}^{(1)}$, represents the slope estimates for the i^{th} subject in $\mathcal{L}$ and $\hat{\mathbf{b}}$ is the mean across all subjects in $\mathcal{L}$ (i.e. the average slope across subjects). Often this mean is subtracted from both the training and testing samples before projection by performing $(\mathbf{x} - \hat{\mathbf{b}})\mathbf{U}_k\mathbf{U}_k^T$ for every vector $\mathbf{x}$ in the training and testing sets (see e.g. Turk and Pentland, (1991)). However, we choose not to remove this mean, as in our case we would be subtracting the mean longitudinal change from cross-sectional features, which would complicate the interpretation of the features. Furthermore, as this is a linear transformation, removing this (or any other) mean in this way amounts to translating both the training and testing features in the same fixed way (by $\hat{\mathbf{b}}\mathbf{U}_k\mathbf{U}_k^T$). Such a translation does not change the samples' relative locations with respect to each other and, consequently, has no impact on subsequent pattern recognition (which is, or should be, translation invariant).

Figure 6.4 helps provide an understanding of the projection described by equation (6.14) by means of a simple two-dimensional longitudinal subspace example. The projected features can be seen as the components of cross-sectional data that lie in a space of common longitudinal change described by the retained PCs in $\mathbf{U}_k$. In the figure there are two retained PCs (i.e. eigenslopes), forming a two-dimensional plane upon which a higher dimensional cross-sectional (follow-up) sample is projected. The schematics in Figures 6.5 and 6.6 depict the formation of training and test features at each cross-validation fold for the Within-Set Prediction and Information Transferring Prediction problems respectively, with matrix $\mathbf{U}_k$ formed by performing LM-PCA on balanced longitudinal design data in both cases. The two figures can be thought of as specific forms of the generic cross-validation fold depicted in Figure 4.2, with the $\mathbf{U}_k\mathbf{U}_k^T$ projection as the particular feature level operation in this case. As this operation is performed within each cross-validation fold, the same data analysis pipeline described in Chapter 4 (see Figure 4.1 and Section 4.2) is still applicable to any analysis performed using our method.

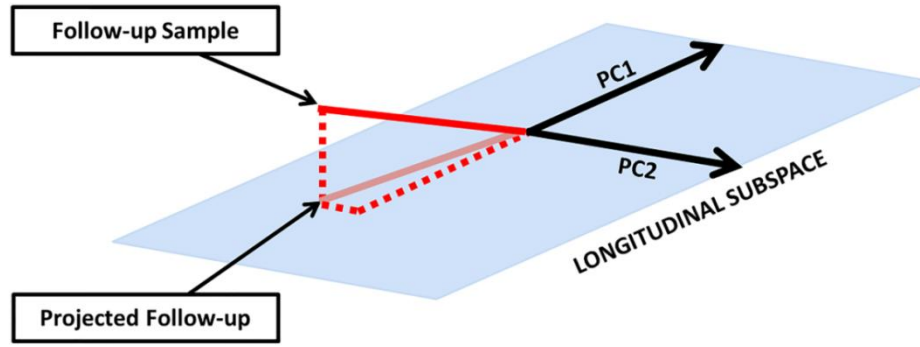


Figure 6-4 To gain intuition for the longitudinal projection described by equation (6.14), we consider a simple case of a sample vector (“Follow-up Sample”) that lies in three dimensional space (an image consisting of three voxels) being projected onto the two dimensional plane described by two principal component vectors (“PC1”, “PC2”) extracted from a hypothetical coefficient matrix describing subjects’ longitudinal changes. The projected sample (“Projected Follow-up”) is thereby composed of the components of the sample vector that lie in the space of common longitudinal changes, described by these principal components. In the case of a first order coefficient matrix, we refer to these principal components as eigenslopes.

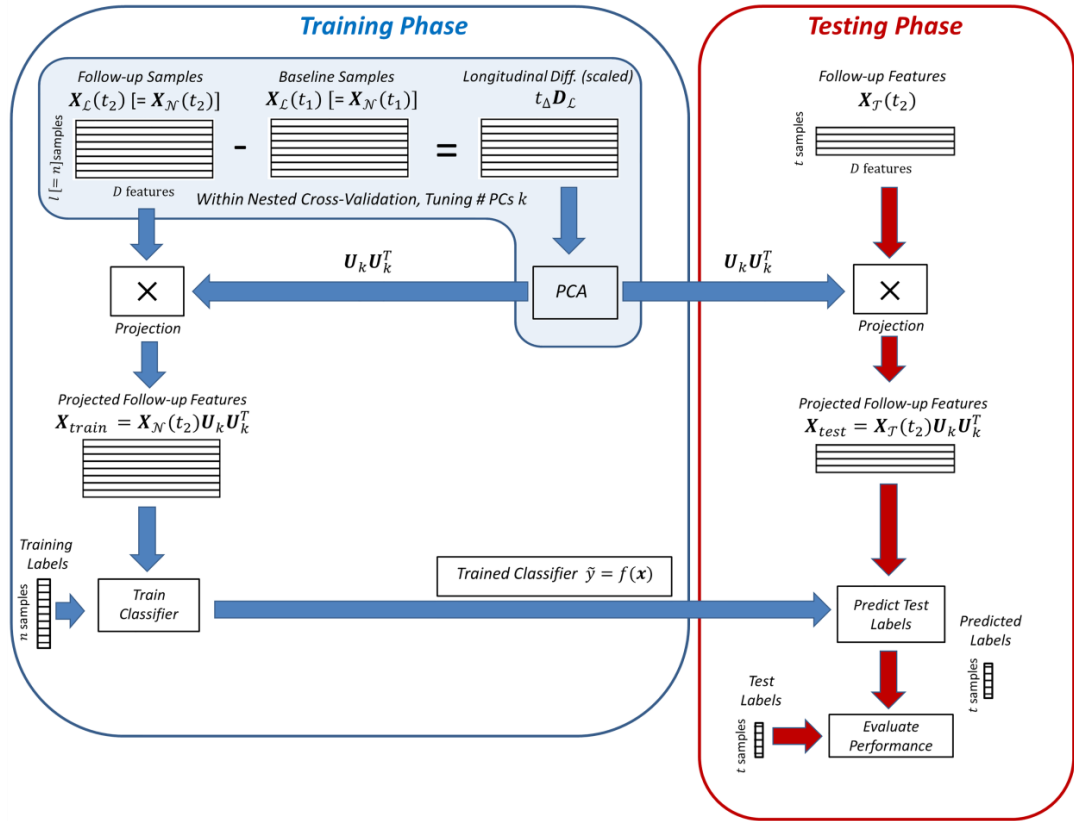


Figure 6-5 Schematic of a cross-validation fold implementing longitudinally matched PCA (LM-PCA) projected features, for the *Balanced Within-Set Prediction* problem of classifying follow-up time-point disease label. We have emphasized that the longitudinal training set is equivalent to the classification training set in this case: the same follow-up data used to form the longitudinal difference matrix (scaled by the time interval between), is projected onto the longitudinal subspace in the training step.

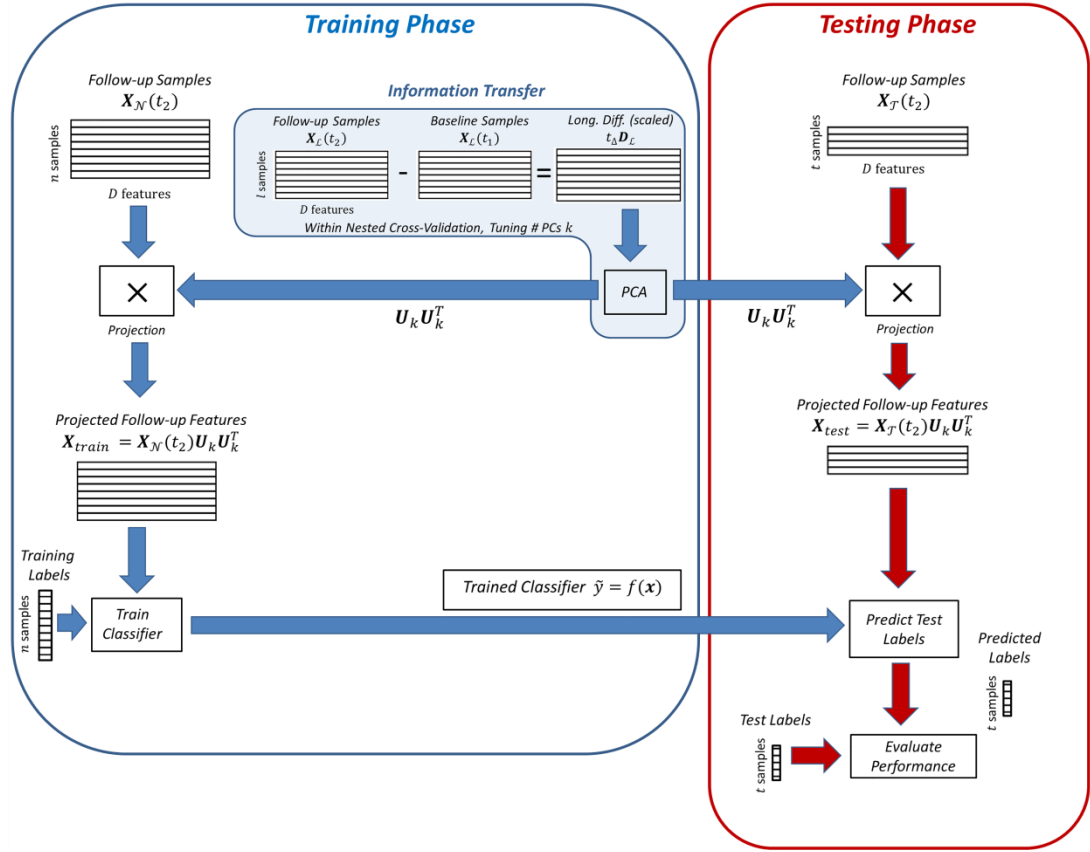


Figure 6-6 Schematic of a cross-validation fold implementing longitudinally matched PCA (LM-PCA) projected features, for the *Balanced Information Transferring Prediction* problem of classifying follow-up time-point disease label. We have emphasized that the longitudinal and classification subject sets are disjoint in this case, with the data in the “Information Transfer” area in blue being used to form the longitudinal difference matrix (scaled by the time interval between scans as in Equation (6.12)). As a result, the follow-up data that is projected onto the the longitudinal subspace differs from the follow-up data used to form the longitudinal difference matrix.

6.11 Kernel Formation

In Chapter 4 (see Section 4.4) we noted that kernel algorithms are often used in problem domains characterized by high dimensional features and low sample sizes, as is often the case with voxel-based features used in neuroimaging. To reiterate, kernel algorithms operate on the pairwise similarities between samples, allowing the computational complexity and storage requirements to depend on the low number of samples rather than the high dimensionality of the features.

In (6.14) we defined our proposed projected features as $\mathbf{X}_{train} = \mathbf{X}_N(t) \mathbf{U}_k \mathbf{U}_k^T$ and $\mathbf{X}_{test} = \mathbf{X}_T(t) \mathbf{U}_k \mathbf{U}_k^T$ so that the linear kernel (see Section 4.4.1) used during classifier training can be expressed as

$$\begin{aligned}\mathbf{K}_{train,train} &= \mathbf{X}_{train}\mathbf{X}_{train}^T = (\mathbf{X}_{\mathcal{N}}(t)\mathbf{U}_k\mathbf{U}_k^T)(\mathbf{X}_{\mathcal{N}}(t)\mathbf{U}_k\mathbf{U}_k^T)^T \\ &= (\mathbf{X}_{\mathcal{N}}(t)\mathbf{U}_k)(\mathbf{X}_{\mathcal{N}}(t)\mathbf{U}_k)^T\end{aligned}\quad (6.15)$$

as $\mathbf{U}_k$ is composed of a set of orthonormal (i.e. mutually orthogonal and having unit norm) vectors. We therefore see that the effective dimensionality of the feature vectors is reduced from d to k by this projection. We can similarly express the kernel matrix of inner products of training samples with test samples as $\mathbf{K}_{train,test} = (\mathbf{X}_{\mathcal{N}}(t)\mathbf{U}_k)(\mathbf{X}_{\mathcal{T}}(t)\mathbf{U}_k)^T$ and, if needed, that of test samples with test samples as $\mathbf{K}_{test,test} = (\mathbf{X}_{\mathcal{T}}(t)\mathbf{U}_k)(\mathbf{X}_{\mathcal{T}}(t)\mathbf{U}_k)^T$. The three matrices $\mathbf{K}_{train,train}$, $\mathbf{K}_{train,test}$ and $\mathbf{K}_{test,test}$ contain all the information needed to perform classifier training (using $\mathbf{K}_{train,train}$) and testing (using $\mathbf{K}_{train,test}$ and, for some algorithms such as the GPC, $\mathbf{K}_{test,test}$) at each cross-validation fold.

6.12 Nested Cross-Validation

In this thesis we use K-fold cross-validation (CV) to assess the generalization ability of our trained models, as discussed in Section 4.4.3. At each CV fold we must choose a value for the parameter k of retained eigenslopes (i.e. principal components of $\mathbf{B}^{(1)}$) used to form the projection $\mathbf{U}_k$ (not to be confused with the number of folds K). As we have no prior information on what this value of k should be, we adopt the popular and straightforward approach of using nested cross-validation to ‘tune’ k at each CV fold. Nested CV involves running an ‘inner’ CV procedure within the training data of a particular (‘outer’) CV fold. In this thesis, for example, we use an outer LOO-CV scheme to make predictions via the optimized parameter k and an inner LOO-CV scheme to optimize k . Within each inner CV fold we train and test the model across a range of k values corresponding to specified amounts of variance explained by the retained PCs. For a desired fraction of explained variance p_{var} , k is chosen as the minimum such that $\sum_{i=1}^k \lambda_i / \sum_{i=1}^n \lambda_i \geq p_{var}$, where λ_i is the i^{th} highest eigenvalue of the covariance matrix used in the PCA. We varied explained variance from 5% to 95% in increments of 5%. We choose the value k that maximizes the balanced accuracy (see Section 4.5) across all inner CV folds.

Chapter 7 Application to Early Dementia

7.1 Introduction

This chapter will apply the longitudinal feature construction method introduced in Chapter 4 to the problem of discriminating early dementia subjects from healthy controls using structural MRI. We will analyse two longitudinal datasets here, allowing us to demonstrate the applicability of our method in various conditions: (i) under different longitudinal designs (balanced or unbalanced); (ii) with a varying or fixed number of samples per subject; (iii) using different diagnostic criteria for early dementia (mild cognitive impairment (MCI) diagnosis or clinical dementia rating (CDR) scale) and (iv) with different follow-up periods between subjects' scans. Across both datasets we show improvements in predictive accuracy relative to cross-sectional classifiers for discriminating disease subjects from healthy controls on the basis of whole-brain structural MRI based voxels. In addition, we demonstrate our method's ability transfer longitudinal information from one set of subjects to make disease predictions in another set of subjects via Longitudinal Information Transferring predictions (see Section 6.9).

7.2 Material and Methods

7.2.1 Datasets

For the first dataset, we used clinical and imaging data from the Heinz Nixdorf RECALL substudy (HNRS) of MCI, whose design and associated data is detailed in Chapter 3. We also used the OASIS longitudinal dataset, the full details of which have been reported previously (Marcus et al., 2010). The longitudinal MRI data from the OASIS dataset consisted of 150 subjects aged 60 to 96, recruited primarily through media appeals and word of mouth. Each of the 150 study subjects had at least two separate visits in which clinical data and MRI data were collected. Clinical data consisted of dementia status, as measured by the Clinical Dementia Rating (CDR) scale. The CDR scale ranges from 0 for no cognitive impairment, to 0.5 for very mild dementia, 1 for mild dementia, 2 for moderate dementia and 3 for severe dementia. A diagnosis of MCI often corresponds to a CDR score of 0.5, although

there is heterogeneity in the rates of progression to AD within subjects at this level of dementia severity (Chang et al., 2011). Subjects with a primary cause of dementia other than AD (e.g., vascular dementia, primary progressive aphasia), active neurological or psychiatric illness (e.g. major depression), serious head injury, history of clinically meaningful stroke, and use of psychoactive drugs were excluded, as were subjects with gross anatomical abnormalities evident in their MRI images (e.g. large lesions, tumours).

7.2.2 Data Acquisition and Quality Control

The data acquisition procedure for the T1-weighted images of the HNRS dataset is detailed in Section 5.2.1. For the OASIS dataset, the sagittal 3D magnetisation prepared rapid acquisition of gradient echo (MPRAGE) images (TR = 9.7 ms; TE = 4 ms; TI = 20 ms; TD = 200 ms; flip angle = 10°; matrix size = 256 × 256 with 128 1.0 mm thick slices; FoV = 25.6 mm × 25.6 mm) were acquired, also on a Siemens MR system operating at 1.5 T (Magnetom Vision, Siemens Healthcare, Erlangen). At each visit, three or four T1-weighted images were acquired for each subject and made publicly available. For our analysis, we chose one scan for each subject that was representative of all others.

For both studies, all T1 weighted images were visually inspected. Images with artefacts or any pathology not associated with dementia, such as signs of stroke, were discarded from further analysis.

7.2.3 Image Preprocessing

We performed the same image segmentation and image registration procedures for both datasets, which were analysed separately. As in the analysis of Chapter 5, we formed both cross-sectional and longitudinal Jacobian determinant (JD) features from the structural MRI data of both datasets with the same whole-brain mask (which excluded extracerebral voxels) as before. A description of how both types of features as well as the mask were formed is provided in Chapter 4 (see Section 4.3). In order to build our longitudinal subspace and form projected and, for comparison, unprojected cross-sectional JD features, in the Within-Set Prediction problems

described in Sections 7.3.1, 7.3.3 and 7.3.4 we segmented and registered all available time-points' images (baseline and any follow-ups) to a common inter-subject template. In Section 7.3.1 we segmented and registered two images from each of the 24 MCI subject and 23 controls (total of 94 images to form template), in Section 7.3.3 we segmented and registered two images from each of the 24 mild dementia subjects and 25 controls (98 images to form template) and in Section 7.3.4 we segmented and registered three or more images from each of the 14 dementia subjects and 28 controls, with a total of 136 images used to form the template (see Section 7.3.4).

For the Between-Set Prediction problem described in Section 7.3.2 we formed the inter-subject template using the segmented baseline and follow-up images from same 24 MCI and 23 controls (same 94 images as in Section 7.3.1), then warped the segmented images of subjects that had either a baseline but no follow-up (79 MCI, 87 controls) or follow-up but no baseline (10 MCI, 20 controls) to the existing template using DARTEL.

In Sections 7.3.1, 7.3.3 and 7.3.4 we formed longitudinal JD features by longitudinally co-registering the baseline and final follow-up images of the same set of subjects used for cross-sectional registration in each respective section. In particular, we co-registered images 24 MCI subjects and 23 controls in Section 7.3.1, 24 mild dementia subjects and 25 controls in Section 7.3.3 and 14 dementia subjects and 28 controls in Section 7.3.4. Following coregistration in each case, we segmented and registered the mid-point average image of each subject to form an inter-subject template, then warped the within-subject JDs to the common space.

In Section 7.3.5 we formed two additional inter-subject templates. When transferring longitudinal information from the OASIS dataset to discriminate HNRS subjects, we formed an inter-subject template using 307 images from OASIS that met our quality control standards along with the same 47 follow-up images from the HNRS dataset that were used in Section 7.3.1 (24 MCI, 23 controls). Among the OASIS images, 25 subjects had a single image at one time-point, 73 subjects had images at two time-points, 33 subjects had images at three time-points, eight subjects had images at four time-points and one subject had images at five time-points. When transferring longitudinal information from the HNRS dataset to discriminate OASIS

subjects, we formed the template using the baseline and follow-up images from 54 HNRS subjects (47 from Section 7.3.1 plus five with a diagnosis of “Pre-MCI” at follow-up and two with a diagnosis of “Dementia” at follow-up) along with the follow-up images from the 49 OASIS subjects used in Section 7.3.3 (24 mild dementia, 25 controls).

Demographic and clinical information for these 47 HNRS subjects are provided in Chapter 5 (Table 5.1 and Section 5.2.2). Briefly, the MCI and control groups differed only terms of education level and cognitive function at follow-up. Appendix E provides the demographic and clinical information for the 49 OASIS subjects used in Sections 7.3.3 and 7.3.5: there was a significant difference in gender across groups along with a difference in MMSE score at follow-up.

7.2.4 Classification

We aim to demonstrate the applicability of our proposed feature construction method to discriminating early dementia. To do so we apply the proposed feature construction method to the cross-sectional JD samples. We will compare the proposed features to cross-sectional JD features from a single time-point and to the longitudinal JD features formed using the diffeomorphic longitudinal registration method discussed in Chapter 4 (Ashburner and Ridgway, 2013b). The cross-sectional JD features are formed using either the baseline or follow-up time-point’s images, depending on which time-point’s diagnostic labels we are predicting. In some cases (Sections 7.3.1, 7.3.3, 7.3.4) we registered all available time-points to form the inter-subject template, while in others (Sections 7.3.3, 7.3.5) we register only the time-point whose diagnosis we are predicting. Features formed using a single time-point will be termed ‘Baseline’ or ‘Follow-up’ features. The longitudinal JD features, formed using two time-points’ images, will be termed ‘Longitudinal coregistration’ features.

To briefly recapitulate our general approach for constructing features, described in detail in the previous chapter, we first model within-subject longitudinal changes in each subject’s cross-sectional JDs. To do so, we fit a polynomial function to the temporal trajectory of each voxel of each subject that has the necessary

longitudinal information. We then build up a matrix of slope coefficients from these fitted (linear, first order) models, perform principal component analysis (PCA) on the matrix, form a projection matrix from a small set of retained principal components (‘eigenslopes’) and form features for classification by projecting cross-sectional information onto the longitudinal subspace they define (see Sections 6.2 through 6.7). We will project cross-sectional JD features at a particular time-point (such as at baseline or follow-up) onto this longitudinal subspace. We will refer to these features as ‘Follow-up, LM-PCA Projected’ or ‘Follow-up, LTC-PCA Projected’, depending on which flavour of our method we use to form the slope coefficient matrix (see Section 6.7).

As in Chapter 5, we will assess the predictive performance of the various features by making cross-validated predictions using the C cost support vector classifier (SVC) described in Section 4.4.2. Cross-validation is introduced in Section 4.4.3. We used the LIBSVM (Chang and Lin, 2011) implementation of the SVC with a fixed value of $C = 1$ throughout all classifications. Figure A1 in Appendix B shows that this choice provides near optimal prediction performance for the datasets we consider here because it lies in a stable region of maximal classifier performance across the three types of features used in Table 7.I. However, in some cases it may be necessary to optimise C by cross-validation to obtain optimal performance. This may be the case if the input vectors are already relatively low-dimensional (e.g. region of interest summary measures).

We perform nested cross-validation to tune the number of retained eigenslopes via the inner cross-validation procedure and predict the label of held-out images via the outer cross-validation procedure (see Section 6.12).

7.2.5 Performance Evaluation and Significance Testing

We assess classifier performance using the same metrics used in the preliminary analysis of Chapter 5, namely the balanced accuracy, sensitivity, specificity, positive predictive value (*PPV*), negative predictive value (*NPV*) and a summary of the receiver operating characteristic’s area under the curve (*ROC AUC*). These metrics are defined in Section 4.5.

We perform permutation tests to assess the statistical significance of the balanced accuracy of each discrimination as described in Section 4.6.

7.3 Results

The results in this section make reference to the two types of classification problems described in Chapter 6, namely the *Within-Set Prediction* (Section 6.8) and *Information Transferring Prediction* (Section 6.9) problems (see Figure 6.3 for a graphical representation of these cases). In addition, this section makes reference to the two types of PCA based projections: *Longitudinal Trajectory Coefficient PCA* (LTC-PCA) and *Longitudinally Matched PCA* (LM-PCA) (See Section 6.7). Briefly, in both cases we are performing PCA on a matrix of first order coefficients; the difference is in how the coefficient matrix is computed. LTC-PCA is the general approach to forming the coefficient matrix when there is an arbitrary number of samples per subject with arbitrary time intervals between samples (based on a GLM) while LM-PCA is the two-sample special case (based on a scaled difference between time-points' features).

We will refer to *Balanced* problems when the longitudinal subject set has both fixed follow-up times and a fixed number of scans per subject (in this case two) and *Unbalanced* problems when the longitudinal subject set has either varying follow-up times or a varying number of scans per subject. We applied our proposed projection method (LM-PCA and/or LTC-PCA) as well as several comparison methods to five classification problems across two datasets (described below, with results shown in Tables 7.1-7.4). Section 7.3.7 uses a repeated measures ANOVA to compare the performance of three different types of features (our proposed longitudinal data based projection, cross-sectional data based projection and no projection) across the five discriminations being presented.

In addition to these five discriminations, each of which was done within a single dataset, Section 7.3.5 presents two additional discriminations that attempted transferred longitudinal information from one dataset to the other.

7.3.1 Balanced Within-Set Prediction (HNRS dataset)

As the HNRS has a balanced longitudinal design, with a baseline and follow-up time-point made available, we use the balanced version of LM-PCA, with the coefficient matrix formed using equation (6.13). In this *Balanced Within-Set Prediction* problem we will be predicting the follow-up time-point's diagnostic labels using projected and unprojected follow-up images. These results are derived from the 24 MCI subjects and 23 healthy controls (HCs) with data from both time-points (follow-up periods of 2.5 ± 0.2 years). To verify that the improvement we show is due to the LM-PCA projection and not only to dimensionality reduction resulting from the application of PCA, we also projected cross-sectional (follow-up) samples onto cross-sectional (baseline and follow-up) subspaces. Table 7.1 shows the classifier performance measures using unprojected follow-up features, “longitudinal coregistration” based features (i.e. longitudinal JD features formed using Ashburner and Ridgway, (2013)), and follow-up features projected onto three different subspaces: formed using baseline time-point information, formed using follow-up time-point information and formed using LM-PCA. In all cases we performed nested cross-validation to choose the optimal number of retained principal components in the projection within each outer cross-validation fold. In Figure 7.1 we see either 90% or 95% explained variances selected at each cross-validation fold: in 35 folds 95% explained variance is selected (corresponding to $k = 16$ retained PCs) while in 12 folds 90% explained variance is selected ($k = 5$ retained PCs).

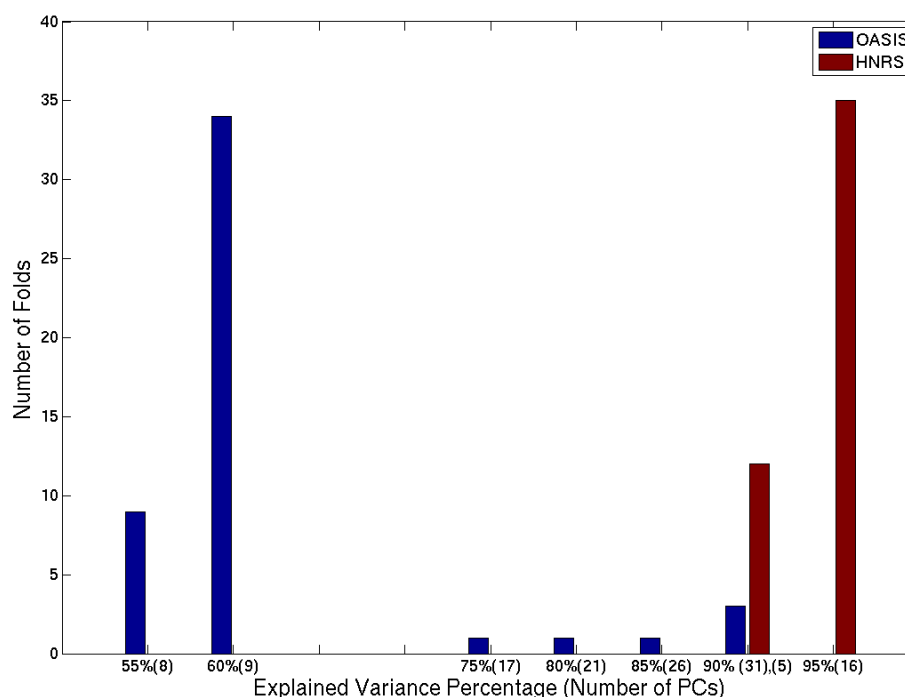


Figure 7-1 Histogram of number of cross-validation folds with given explained variance percentage and corresponding number of principal components for the *Balanced Within-Set Prediction* classification problem, discriminating MCI subjects vs. HCs using HNRS dataset (Table 7.1 result, shown in red bars) and discriminating very mild dementia from HCs using the OASIS dataset (Table 7.3 result, shown in blue bars). We see that the two distributions differ greatly: the classifier exclusively chooses 90% and 95% explained variances when discriminating MCI subjects from controls using the HNRS dataset (Table 7.1), with $k = 5$ and $k = 16$ corresponding retained PCs respectively. When discriminating very mild dementia subjects from controls using the OASIS dataset (Table 7.3), the explained variances are mostly 55% and 60% with $k = 8$ and $k = 9$ retained PCs respectively. The distributions overlap at 90% explained variance, with $k = 31$ associated PCs in the OASIS case and $k = 5$ associated PCs in the HNRS case.

Table 7-1 HNRS Dataset, *Balanced Within-Set Prediction*, discriminating MCI subjects vs. Healthy Controls (HC) subjects with two longitudinal time-points per subject for all subjects (classification subject set equal to longitudinal subject set in this case).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>Predicting Follow-up Class Label (MCI $n = 24$, HC $n = 23$) using same subjects in longitudinal subject set</i>						
Follow-up	53.2	54.2	52.2	54.2	52.2	0.639
Longitudinal coregistration ¹	59.6	58.3	60.9	60.9	58.3	0.574
Follow-up, Baseline Subspace Projected	59.7	54.2	65.2	61.9	57.7	0.578
Follow-up, Follow-up Subspace Projected	59.6	58.3	60.9	60.9	58.3	0.627
Follow-up, LM-PCA Projected	74.3**	83.3	65.2	71.4	78.9	0.774

¹ Features based on Ashburner and Ridgway, (2013)

* Statistically significant balanced accuracy, permutation test p-value < 0.05

** Statistically significant balanced accuracy, permutation test p-value < 0.01

7.3.2 Balanced Information Transferring Prediction (HNRS dataset)

Table 7.2 shows the performance measures obtained when testing the information transfer capability of our approach using the HNRS dataset. Here, we formed a balanced LM-PCA projection matrix with the subjects used in Table 7.1 (24 MCI, 23 controls), then predicted the *baseline* time-point’s diagnostic labels using unprojected and LM-PCA projected baseline features on a disjoint set of subjects that had a baseline scan but no available follow-up (79 MCI subjects and 87 controls)¹¹. Also shown in Table 7.2 is the result of predicting the *follow-up* time-point’s diagnostic labels using unprojected, follow-up subspace projected and LM-PCA projected follow-up features on a disjoint set of subjects with a follow-up scan but no corresponding baseline scan (10 MCI subjects and 20 controls) using the same projection matrix (formed with 24 MCI, 23 controls). As each subject has only one time-point’s information in these experiments, we could not form the “longitudinal coregistration” based features in this case as in the other tables. In this case, none of the balanced accuracies statistically exceeded chance under permutation testing.

¹¹ Note that the HNRS is ongoing and many of the follow-up scans have not yet been made available for analysis.

Table 7-2 HNRS Dataset, *Balanced Information Transferring Prediction*, discriminating MCI subjects vs. HC subjects with two longitudinal time-points per subject in longitudinal subject set (which is the same set of subjects used in Table 7.1). Here the classification subject set is disjoint from the longitudinal subject set in the two classification tasks considered: classifying baseline class label with subjects that have only baseline scans and classifying follow-up class label with subjects that have only follow-up scans. Note that in this case we cannot form longitudinal features using the longitudinal coregistration method of Ashburner and Ridgway, (2013) as each classification set subject has data from only baseline or follow-up time-point information.

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>Predicting Baseline Class Label (MCI n = 79, HC n = 87) using disjoint longitudinal subject set (MCI n = 24, HC n = 23) from Table 7.1</i>						
Baseline	49.1	43.0	55.2	46.6	51.6	0.474
Longitudinal coregistration ¹	–	–	–	–	–	–
Baseline, Baseline Subspace Projected	38.1	27.8	48.3	32.8	42.4	0.401
Baseline, LM-PCA Projected	48.7	34.2	63.2	45.8	51.4	0.509
<i>Predicting Follow-up Class Label (MCI n = 10, HC n = 20) using disjoint longitudinal subject set (MCI n = 24, HC n = 23) from Table 7.1</i>						
Follow-up	45.0	20.0	70.0	25.0	63.6	0.445
Longitudinal coregistration ¹	–	–	–	–	–	–
Follow-up, Follow-up Subspace Projected	47.5	20.0	75.0	28.6	65.2	0.475
Follow-up, LM-PCA Projected	60.0	50.0	70.0	45.0	73.7	0.590

¹ Features based on Ashburner and Ridgway, (2013)

* Statistically significant balanced accuracy, permutation test p-value < 0.05

** Statistically significant balanced accuracy, permutation test p-value < 0.01

7.3.3 Balanced Within-Set Prediction (OASIS dataset)

The OASIS dataset has an unbalanced longitudinal design, with each subject scanned on two or more visits. To compare the two time-point, balanced design LM-PCA method across datasets we mimicked a balanced longitudinal design with this dataset by restricting the follow-up times of subjects to be roughly similar to that of the HNRS dataset. We used a subject set composed of subjects with follow-up scans between 1.4 years and 2.5 years after baseline, resulting in follow-up periods of 1.9 ± 0.3 years (24 subjects with very mild dementia, i.e. CDR 0.5, and 25 healthy controls, i.e. CDR 0).

In Figure 7.1 we see in this case that the number of principal components explaining mostly 55% and 60% of variance were selected across cross-validation

folds to form LM-PCA projections, with $k = 8$ and $k = 9$ retained PCs respectively. This contrasts with the higher explained variances (90% and 95%) selected with the HNRS dataset (Table 7.1), highlighting the need to use nested cross-validation to tune this parameter.

Table 7-3 OASIS Dataset, *Balanced Within-Set Prediction*, discriminating very mild dementia (CDR 0.5) subjects vs. HC (CDR 0) subjects with two longitudinal time-points per subject for all subjects (classification subject set equal to longitudinal subject set in this case).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>Predicting Follow-up Class Label (very mild dementia $n = 24$, HC $n = 25$) using same subjects in longitudinal subject set</i>						
Follow-up	59.0	50.0	68.0	60.0	58.6	0.627
Longitudinal coregistration ¹	51.2	58.3	44.0	50.0	52.4	0.615
Follow-up, Baseline Subspace Projected	57.0	50.0	64.0	57.1	57.1	0.552
Follow-up, Follow-up Subspace Projected	63.4	70.8	56.0	60.7	66.7	0.700
Follow-up, LM-PCA Projected	69.4*	70.8	68.0	68.0	70.8	0.718

¹ Features based on Ashburner and Ridgway, (2013)

* Statistically significant balanced accuracy, permutation test p-value < 0.05

** Statistically significant balanced accuracy, permutation test p-value < 0.01

7.3.4 Unbalanced Within-Set Prediction (OASIS dataset)

Here we considered the problem of discriminating those subjects with three or more longitudinal time-points using the OASIS dataset, without restricting the time interval between samples. To increase the number of disease class subjects, we formed a class consisting of five subjects with mild dementia (corresponding to CDR 1) along with nine subjects with very mild dementia (CDR 0.5), for a total of 14 subjects being discriminated from 28 healthy subjects (CDR 0). There were 33 subjects with three longitudinal measurements, 8 subjects with four longitudinal measurements and one subject with five longitudinal measurements.

We compared the performance of features formed using the cross-sectional data at the final follow-up time-point for each subject (“Final Follow-up”) to projecting this data onto five different subspaces: two formed using purely cross-sectional information (“Final Follow-up, Baseline Subspace Projected” and “Final Follow-up, Final Follow-up Subspace Projected”), two formed by performing LM-PCA using two time-points’ (TPs’) scans from each subject and one by performing

LTC-PCA using all available TPs' scans. The "Final Follow-up, LM-PCA Projected (2 TPs, Short)" features were formed by creating an unbalanced LM-PCA projection (via equation (6.12)) using the last two time-points' information for each subject. In this case the time interval between scans is the shortest possible for each subject, resulting in an interval of 2.0 ± 0.8 years across subjects. The "Final Follow-up, LM-PCA Projected (2 TPs, Long)" features were formed in a similar manner by creating an LM-PCA projection using the first and last time point for each subject, i.e. with the longest time intervals between scans for each subject, resulting in an interval of 4.2 ± 1.2 years across subjects. Finally, "Final Follow-up, LTC-PCA Projected (All TPs)" were formed using all time-points for each subject. We used the more general LTC-PCA for this (with polynomial model order $p = 1$, as it is in LM-PCA), which allowed us to estimate the slope of samples' longitudinal trajectories using all (three or more) time-points for each subject rather than having to select two time-points when using LM-PCA. As in Table 7.1 and 7.3 we also compare to "longitudinal coregistration" features. In this case we coregistered the first and last time-point of each subject to best compare to the LTC-PCA projected features.

The performance metrics for all features are shown in Table 7.4. Classifiers that used the "longitudinal coregistration" and LTC-PCA projected features were the only ones that statistically exceeded chance under permutation testing. This suggests that when more than two longitudinal samples are available per subject, we can derive a better estimate of the linear coefficient matrix with the more general LTC-PCA than with the two-sample LM-PCA approach.

Table 7-4 OASIS Dataset, *Unbalanced Within-Set Prediction*, mild and very mild dementia (CDR 1, CDR 0.5) subjects vs. HC (CDR 0) subjects with a minimum of three longitudinal time-points per subject for all subjects (classification subject set equal to longitudinal subject set in this case).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>Predicting Final Follow-up Class Label (mild and very mild dementia n = 14, HC n = 28) using same subjects in longitudinal subject set</i>						
Final Follow-up	53.6	28.6	78.6	40.0	68.6	0.676
Longitudinal coregistration ¹	66.1*	42.9	89.3	66.7	75.8	0.653
Final Follow-up, Baseline Subspace Projected	58.9	35.7	82.1	50.0	71.9	0.686
Final Follow-up, Final Follow-up Subspace Projected	50.0	21.4	78.6	33.3	66.7	0.617
Final Follow-up, LM-PCA Projected (2 TPs, Short)	53.6	28.6	78.6	40.0	68.8	0.661
Final Follow-up, LM-PCA Projected (2 TPs, Long)	62.5	50.0	75.0	50.0	75.0	0.663
Final Follow-up, LTC-PCA Projected (All TPs)	67.9*	57.1	78.6	57.1	78.6	0.702

¹ Features based on Ashburner and Ridgway, (2013)

* Statistically significant balanced accuracy, permutation test p-value < 0.05

** Statistically significant balanced accuracy, permutation test p-value < 0.01

7.3.5 Longitudinal Information Transfer across datasets

Table 7.5 presents the results of two additional discriminations that attempt longitudinal information transfer. Unlike the analysis in Table 7.2, where we transferred longitudinal information one set of subjects to the other within a single dataset (specifically, the HNRS), here we form the longitudinal subspace using subjects from one dataset and then project and discriminate subjects from the other dataset. To compare the use of within and between dataset longitudinal information, we attempted the same discriminations as in Tables 7.1 (HNRS subjects, 24 MCI vs. 23 HC) and 7.3 (OASIS subjects, 24 very mild dementia vs. 25 HC) using several different longitudinal subspaces for each discrimination.

To discriminate the HNRS subjects we formed two different LTC-PCA based subspaces using OASIS images. We used 282 images from the 115 subjects that had images at two or more time-points to create the “Follow-up, 2+ LTC-PCA Subspace” features and 136 images from the 42 subjects that had images at three or more time-points to create the “Follow-up, 3+ LTC-PCA Subspace” features. To discriminate the OASIS subjects, we formed two LM-PCA based subspaces using

HNRS images using the baseline and follow-up images of the 47 subjects in Table 7.1. (24 MCI, 23 HC) to form the “Follow-up, LM-PCA Subspace #1” features. We then added the five subjects with a diagnosis of “Pre-MCI” and two subjects with a diagnosis of “Dementia” to form a second subspace of 52 subjects to form the “Follow-up, LM-PCA Subspace #2” features.

Table 7-5 Longitudinal information transfer across datasets. Discriminating 24 MCI subjects vs. 23 HC in HNRS dataset (as in Table 7.1) using several longitudinal subspaces from OASIS dataset and discriminating 24 very mild dementia subjects vs. 25 HC in OASIS dataset (as in Table 7.3) using one longitudinal subspace from HNRS dataset (as in Table 7.1).

Features	Subspace Subjects	Subspace Images	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>Predicting Follow-up Class Label (MCI n = 24, HC n = 23, HNRS dataset) using longitudinal information from OASIS dataset</i>								
Follow-up	–	–	55.3	58.3	52.2	56.0	54.5	0.670
Follow-up, 2+ LTC-PCA Subspace	115	282	63.8*	66.7	60.9	64.0	63.6	0.692
Follow-up, 3+ LTC-PCA Subspace	42	136	70.2**	70.8	69.6	70.8	69.6	0.705
<i>Predicting Follow-up Class Label (very mild dementia n = 24, HC n = 25, OASIS dataset) using longitudinal information from HNRS dataset</i>								
Follow-up	–	–	56.9	45.8	68.0	57.9	56.8	0.613
Follow-up, LM-PCA Subspace	47	94	47.0	50.0	44.0	46.2	47.8	0.498
Follow-up, LM-PCA Subspace	54	108	36.8	37.5	36.0	36.0	37.5	0.412

* Statistically significant balanced accuracy, Monte Carlo p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo p-value < 0.01

7.3.6 Visualizing and comparing eigenspaces across datasets

Figure 7.2 depicts the most explanatory eigenslopes (principal components of the slopes matrices, see Section 6.6) from two of the eigenspaces formed in Table 7.5. On the HNRS side (left), we depict the top two eigenslopes from the LM-PCA based subspace formed from the 47 HNRS subjects (also used in Table 7.1). On the OASIS side (right), we depict the second LTC-PCA based eigenspace from Table 7.5 that was formed from the 42 OASIS subjects that had three or more time-points’ images. In this case, the top two eigenslopes explain only 52% of the longitudinal variance so we additionally visualize a third composite eigenslope that is the sum of eigenslopes

#3 through #16, which cumulatively explains 35% of the variance, so that overall we are visualizing the same amount of variance as in the HNRS case (87% in both).

Figure 7.3 compares these eigenslopes using pairwise correlation and pairwise angle¹². Although the two HNRS eigenslopes are orthogonal, the two vectors have the highest pairwise correlation (0.49). These vectors therefore have the most similar pattern of variability about their mean across all feature dimensions. This comparison metric may not be as useful in this case as the closely-related angle measure, which allows us to understand whether one eigenspace would explain the variance of another (mutually orthogonal eigenspaces do not explain each other's variance). Inspecting the pairwise angles in Figure 7.3, we see that, as expected, the angle between eigenslopes within each eigenspace is always 90° as these vectors are orthonormal by construction. However, we also see that each eigenslope is nearly orthogonal to those in the other eigenspace and would therefore expect each eigenspace to explain a small amount of variance in the other dataset. Indeed, when we project the slope estimate matrix from one dataset onto the full longitudinal eigenspace of the other we explain only 18% of the HNRS data's variance and only 17% of the OASIS data's variance.

¹² The angle between two vectors is the inverse cosine of their dot product normalized (divided) by the product of their individual magnitudes.

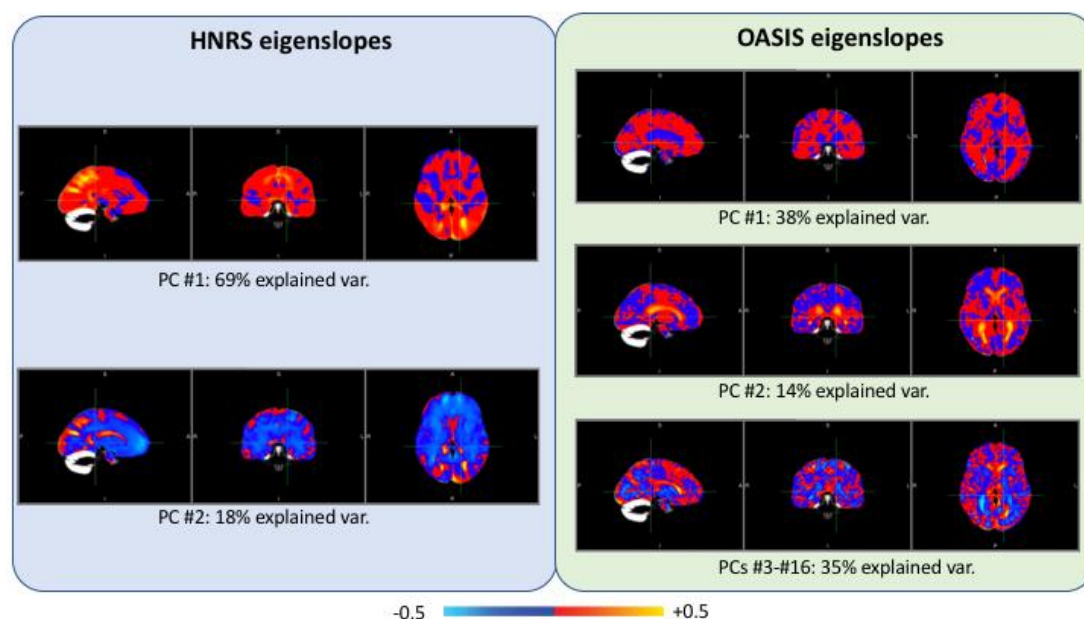


Figure 7-2 Visualization of eigenslopes from the two of the eigenspaces used in Table 7.5. Each image is scaled such that the maximum absolute value is equal to one.

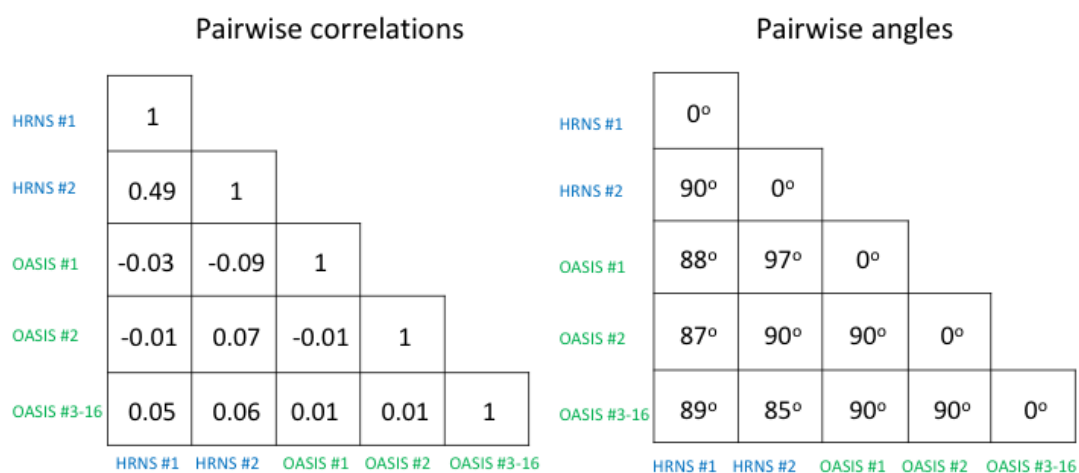


Figure 7-3 Pairwise correlations and angles between eigenslopes depicted in Figure 7.2.

7.3.7 Statistical comparison of methods

In Tables 7.1, 7.3 and 7.4 we see that LM-PCA/LTC-PCA projected features' classification accuracies achieve statistical significance while unprojected features' do not, with statistical significance bestowed upon accuracies meeting the usual criteria of $p < 0.05$ under permutation testing. The difference between a significant accuracy and a non-significant accuracy is not necessarily itself significant however

(Gelman and Stern, 2006). Indeed LM-PCA/LTC-PCA projection only improved accuracy relative to unprojected cross-sectional features or “longitudinal coregistration” based features in one case (LM-PCA projected versus unprojected features in Table 7.1, $p < 0.05$, McNemar’s test).

Because of the small sample sizes in each comparison, we have limited statistical power to detect such differences between the methods within individual discriminations. However, we are more interested in testing whether our method provided a consistent improvement overall (i.e. across all classification problems and datasets) rather than whether it improves individually for any specific classification problem. To assess this, we performed a 3×1 repeated measures analysis of variance (ANOVA) with three different methods as the “within-comparison factor” and the five balanced accuracies across comparisons as the samples for each method. The three methods compared were: (i) the most appropriate version of our method for each contrast (LM-PCA for the comparisons in Tables 7.1-7.3 or LTC-PCA for the comparisons in Table 7.4); (ii) the latest follow-up samples projected onto the latest follow-up subspace (the single follow-up in Tables 7.1 and 7.3, the baseline in the first part of Table 7.2, the follow-up in the second part of Table 7.2, the final follow-up in Table 7.4) and (iii) features formed using the (unprojected) latest follow-up samples. The results show a significant difference in the effect of the projection method, with $F(1.52, 6.08) = 10.91$, $p = 0.01$. Post-hoc pairwise t-tests showed that LM-PCA/LTC-PCA projected features performed better overall than unprojected features ($p = 0.03$) and cross-sectionally projected features ($p = 0.003$) whereas cross-sectionally projected features did not outperform unprojected features ($p = 0.74$).

Because several of the discriminations had imbalanced classes (those from Table 7.2 and 7.4), we repeated this analysis using the ROC AUC metric. The ROC AUC metric is not sensitive to the choice of the classifier’s decision threshold, which may need to be adjusted when classes are imbalanced. The results were very similar to those from the balanced accuracy based analysis: a significant difference in the effect of the projection, $F(1.87, 7.48) = 9.09$, $p = 0.01$, with post-hoc pairwise tests showing LM-PCA/LTC-PCA projected features outperforming unprojected features ($p = 0.03$) and cross-sectionally projected features ($p = 0.01$), and no difference between cross-sectionally projected and unprojected features ($p = 0.78$).

We could not include “longitudinal coregistration” based features in these ANOVAs as it was not possible to form such features in Table 7.2 due to a lack of either baseline or follow-up information in the classification subject set. Therefore, we conducted an additional paired t-test, comparing the balanced accuracies of LM-PCA in Tables 7.1 and 7.3 and LTC-PCA in Table 7.4 against “longitudinal coregistration” based features’ balanced accuracies. This showed no significant differences.

7.3.8 Discriminative and Forward Brain Maps

Figures 7.2 and 7.3 display t-statistic images, weight maps and forward maps (see Section 4.7) for the cross-sectional features and LM-PCA projected cross-sectional features derived from the *Balanced Within-Set Prediction* problem (HNRS and OASIS datasets respectively). For both types of features the figures show a good correspondence between the unthresholded two sample t-statistic maps describing group differences between disease subjects and controls at each voxel and forward maps that were generated via equation (4.15). The weight maps depicted, generated via equation (4.14), are useful for understanding which combination of regions is driving the classifier’s decisions. However they differ greatly from the other two map types, reflecting the fact that they do not represent group differences (Haufe et al., 2014b).

In Figure 7.4, which corresponds to the HNRS dataset results in Table 7.1, we see that the cross-sectional follow-up feature based forward maps and the LM-PCA projected counterparts have negative values (cold colours) across many grey and white matter regions. As we are interpreting maps of Jacobian determinant features, quantifying the expansion or contraction of voxels during diffeomorphic registration to an inter-subject template, areas of negative group differences indicate contraction (i.e. degeneration) in the positive (disease) class relative to the negative (healthy control) class. In particular, we see the LM-PCA projected cross-sectional forward maps have stronger negative values in the parietal and precuneus areas, the frontal pole and the temporal lobe relative to the cross-sectional map.

In Figure 7.5, which corresponds to the OASIS dataset result in Table 7.3, we see that both the cross-sectional and LM-PCA projected cross-sectional forward maps show strong positive values within the lateral ventricles and the Sylvian fissure, indicating a regional shrinkage of brain tissue in early dementia subjects compared to controls. In addition there are negative values in the temporal lobes in both sets of forward maps. Qualitatively, the two sets of forward maps in this figure resemble each other, with some evidence of de-noising due to the PCA procedure via a reduction in the number of negative value regions in the LM-PCA projected cross-sectional map compared to the purely cross-sectional forward map. Overall, these patterns of coefficients are highly consistent with the literature on whole brain and ventricular volume change in dementia subjects relative to controls (Freeborough and Fox, 1997; Jack et al., 2004; Nestor et al., 2008).

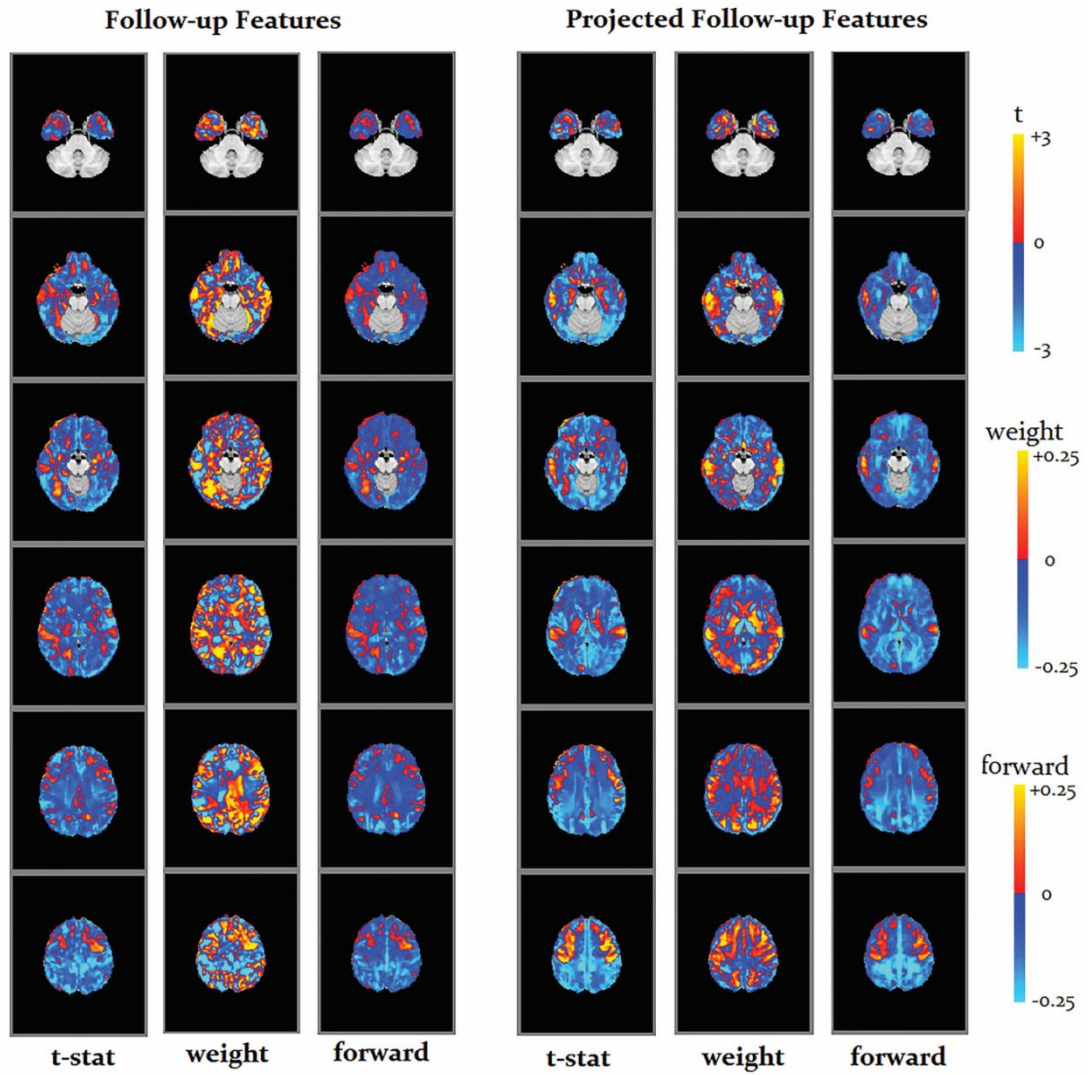


Figure 7-4 Comparison of maps for *Balanced Within-Set Prediction* problem, HNRS dataset. Mass univariate differences between groups' features via an (unthresholded) t-statistic and classifier weight and forward maps, discriminating MCI versus HC using cross-sectional (follow-up time-point) features compared to LM-PCA projected cross-sectional (same time-point) features, corresponding to Table 7.1. Weight and forward maps are scaled such that the maximum absolute value in each image is equal to one. In all maps, positive values (hot colors) depict areas of stronger association to the positive class (MCI group), while negative values (cold colors) depict areas of stronger association to the negative class (HC group).

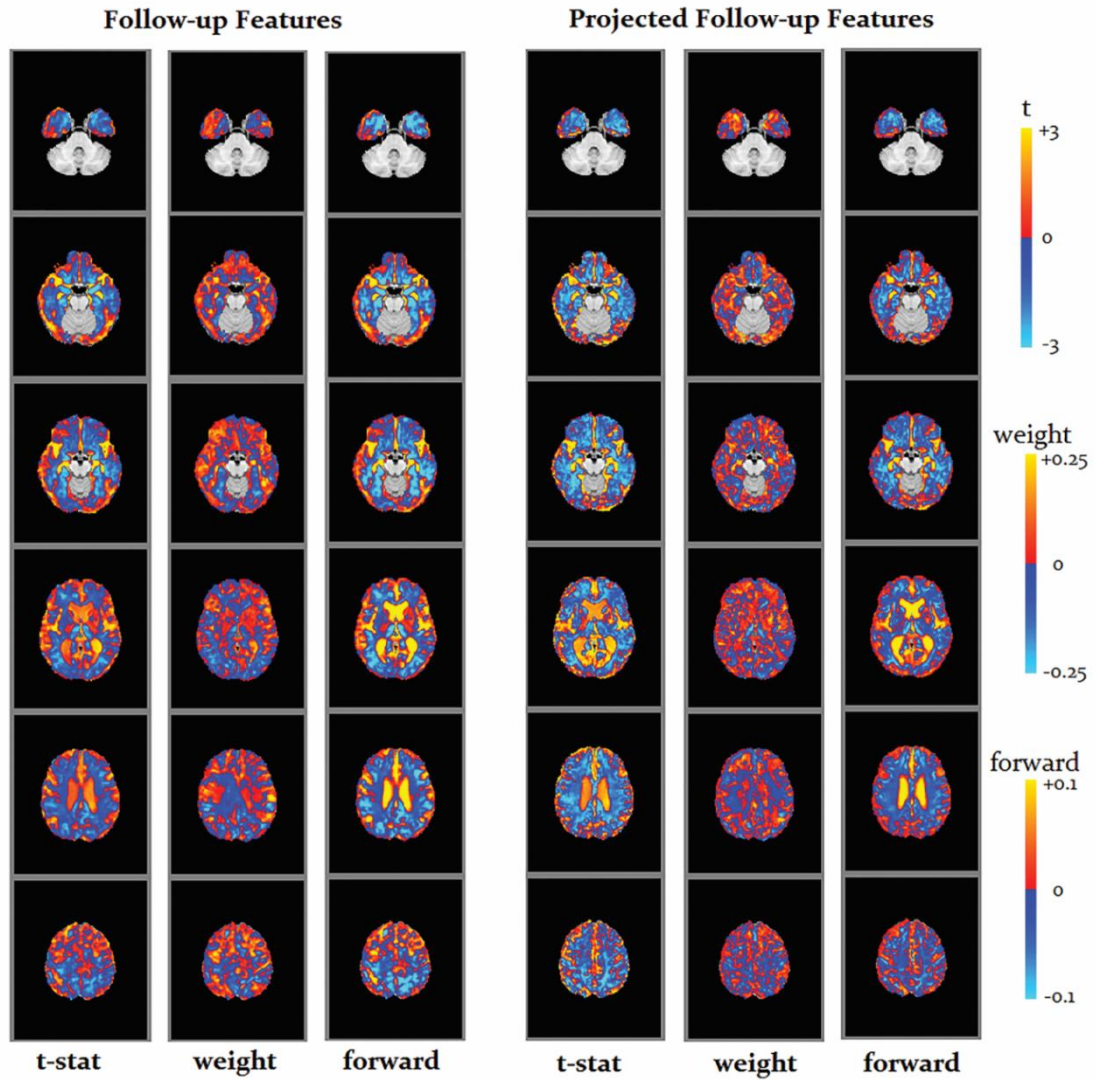


Figure 7-5 Comparison of maps for *Balanced Within-Set Prediction* problem, OASIS dataset. Mass univariate differences between groups' features via an (unthresholded) t-statistic and classifier weight and forward maps, discriminating very mild dementia (CDR 0.5) versus HC using cross-sectional (follow-up time-point) features compared to LM-PCA projected cross-sectional (same time-point) features, corresponding to Table 7.3. Weight and forward maps are scaled such that the maximum absolute value in each image is equal to one. In all maps, positive values (hot colors) depict areas of stronger association to the positive class (CDR 0.5 group), while negative values (cold colors) depict areas of stronger association to the negative class (HC group).

7.4 Discussion

In chapter 6 we introduced a new feature construction based technique for pattern recognition, which makes use of the longitudinal information available in both unbalanced and balanced longitudinal designs. In this chapter we demonstrated the effectiveness of this technique using high-dimensional neuroimaging data on the problem of discriminating both MCI subjects from healthy controls and early dementia subjects from controls, showing improvements in accuracy across two

separate datasets. The method we introduce has a single tuning parameter and, as a feature construction operation, is able to work with existing registration algorithms used in pre-processing and any pattern recognition algorithm (including regression models). In principle, the method is not limited to neuroimaging and can be applied to any high-dimensional longitudinal dataset.

We have shown that: (i) in the case of balanced longitudinal designs with two samples per subject, projected cross-sectional features have higher predictive accuracy in discriminating disease than do unprojected cross-sectional features; (ii) for unbalanced designs, higher classification accuracy can be attained using three or more samples to estimate the linear coefficient matrix used to form the necessary longitudinal projection compared to using the minimum of two samples and (iii) information transfer from one set of subjects to another is possible using the proposed method. In addition, we have shown that the proposed features, based on a linear transformation, retain the ability to visualize a linear classifier's weight maps and forward maps.

Comparing our method to others in the literature, the improvements we show on the *Balanced Within-Set Prediction* problem are similar to those achieved by Chen and DuBois Bowman, (2011). However, a direct comparison is not possible as the authors were discriminating AD from controls using PET data with a 12-month follow-up period. The authors report an approximately 10% improvement in predictive accuracy using region of interest (ROI) based voxels, similar to what we achieved on the OASIS dataset. Another comparison can be made to Wolz et al., (2010), who created non-linear manifolds using longitudinal and cross-sectional information into which they embedded both cross-sectional and longitudinal data to form features for classification. The authors also considered the problem of discriminating MCI from controls (among other contrasts between AD, MCI and controls) using T1-weighted structural MRI. They showed improved discrimination accuracy from 64% using baseline imaging features to 69% using baseline plus longitudinal imaging features.

Ziegler et al., (2015) estimated rates of change (slope) within subjects' samples in a mass-univariate context, noting that a smaller number of longitudinal samples per subject resulted in reduced parameter accuracy. This is in agreement

with the results we present in Table 7.4, where the “Final Follow-up, LTC-PCA Projected (All TPs)” features indicate that improved estimation of the linear coefficient matrix using three or more longitudinal samples per subject leads to better predictions compared to LM-PCA projected features, which were based on the minimum of two samples per subject (see below).

There appears to be an advantage in using balanced design data rather than unbalanced data when forming longitudinal projections. In Tables 7.1 and 7.3 we see that features formed using balanced LM-PCA projections have statistically significant balanced accuracies while in Table 7.4 the unbalanced LM-PCA projected features’ balanced accuracies did not achieve significance. This observation is in agreement with Ziegler et al., (2015), who note that less balanced designs led to poorer correspondence of slope estimates with ground truth as well as higher noise levels. When unbalanced design information is available, the result in Table 7.4 suggests that better predictions result from using all available time-points’ information when forming a longitudinal projection (through the use of LTC-PCA).

Tables 7.5 explored the transfer of longitudinal information across the two datasets, revisiting the discriminations of Tables 7.1 and 7.3. We see similarly low balanced accuracies using purely cross-sectional features (55.3% and 56.9% for the “Follow-up” features in Table 7.5 corresponding to 53.2% and 59.0% in Tables 7.1 and 7.3 respectively). The slight differences are attributable to differences in the set of subjects used to form each template (e.g. the template used in Table 7.1 was formed using the baseline and follow-up images of 47 HNRS subjects while the template in Table 7.5 was formed using the same subjects’ 47 follow-up images plus 307 OASIS images). We see a similar increase in balanced accuracy using a longitudinal subspace composed of 42 OASIS subjects with three or more longitudinal time-points (“Follow-up, 3+ LTC-PCA Subspace” in Table 7.5) as compared to the within-set longitudinal subspace based on the 47 subjects’ own baseline and follow-up images (“Follow-up, LM-PCA Projected” in Table 7.1). In contrast, we see only a slight increase when using all available longitudinal information, i.e. the superset of 115 OASIS subjects with two or more images (“Follow-up, 2+ LTC-PCA Subspace” in Table 7.5), implying that a smaller subject set with more follow-up information per subject generalizes better than a larger subject set with less information per subject. In contrast to these improvements,

when discriminating the 49 OASIS subjects in Table 7.5 we see the two balanced LM-PCA based subspaces (one formed using the oft-mentioned 47 HNRS subjects, the other using seven additional subjects) do not aid in discriminating the 49 OASIS subjects. There are several potential explanations: (i) the HNRS longitudinal subspace captures commonalities among healthy and mildly impaired subjects and is not as useful for discriminating early dementia subjects (e.g. the space projects out some important dementia-related areas, such as the ventricles, that do not change much within healthy or MCI subjects) and/or (ii) despite being based on balanced information these longitudinal spaces are not robust enough to generalize across datasets due to the number of subjects used or the number of samples per subject. Section 7.3.6 and corresponding Figures 7.2 and 7.3 support the latter view, showing that eigenslopes from one dataset are close to orthogonal to those in the other and that furthermore, each eigenspace only explains a small fraction of the longitudinal variance within the other dataset. Further work with multiple large longitudinal datasets is needed to disentangle the effects of longitudinal set size (both number of subjects and samples per subject) and disease severity when transferring longitudinal information.

Our method starts with the assumption of subject-specific models of longitudinal trajectories, deriving expressions for the first order coefficient (i.e. slope) matrix describing intra-subject changes over time, across all subjects and all dimensions, for the general case of an unbalanced longitudinal design. The benefit of such an approach is its conceptual and computational simplicity, as one can compute each subject's coefficients once and then assemble the desired coefficient matrix (of any model order) for a particular set of subjects at each cross-validation fold. Longitudinal trajectories are likely correlated among similarly aged subjects, however, and accounting for these inter-subject correlations may lead to better estimates of the desired coefficient matrix, potentially improving the *Information Transferring Prediction* capability of our method.

Accounting for inter-subject correlations is particularly important for neuroimaging data, where the number of longitudinal samples per subject is often small, amounting to two samples per subject in the HNRS dataset and between two and five in the OASIS dataset. Therefore, for future work we will investigate methods to model such correlations, potentially providing better estimation of

subject level coefficients by sharing information between subjects. Multi-task learning (MTL) models have recently been applied to neuroimaging to account for between-subject correlations (Zhang and Shen, 2012; Marquand et al., 2014) and may be particularly well suited for this purpose.

We have focussed on building a subspace projection using the matrix of first order coefficients. One potential limitation of this decision, imposed by a relatively small number of subjects having full follow-up data in both datasets, is that the subspace we build is most appropriate for situations where the linear term dominates temporal trajectories. There is evidence that the non-linear component of trajectories may be important, particularly in older subjects (Raz et al., 2010; Raz and Lindenberger, 2011; Fjell et al., 2014). To this end, one can estimate separate projections based on the first order and second order coefficient matrices, selecting the number of principal components retained in each projection using a common or model order specific amount of explained variance, for instance. Features could be then composed of a concatenation of cross-sectional features projected onto the linear coefficient subspace with features projected onto the quadratic coefficient subspace. Estimating the matrix of second order coefficients requires at least three time-points per subject, with the result in Table 7.4 suggesting that more than this minimum may be necessary for good estimates. Indeed, the small sample size that results from requiring that all subjects have data in all time-points is an important limitation of this study. Across the five contrasts we performed, we have shown that there is an improvement due to LM-PCA/LTC-PCA projected features relative to strictly cross-sectional features; however, due to the limited number of longitudinal samples, we could not show a consistent improvement within each contrast. Therefore, additional validation of the method on larger samples (such as ADNI¹³) and exploring the conditions under which using higher order coefficient matrices would benefit a predictive model is an important avenue for future work.

In summary, we have introduced a novel means of capitalising on longitudinal information for pattern recognition analysis of high-dimensional data. We have provided a conceptual framework for modelling longitudinal trajectories of change over time, which enables: (i) the use of balanced and unbalanced longitudinal

¹³ Alzheimer's Disease Neuroimaging Initiative (ADNI) data is available at <http://adni.loni.usc.edu>.

designs; (ii) the modelling of linear and non-linear effects and (iii) the ability to transfer longitudinal information between sets of subjects. Our results suggest that longitudinal subspace projection is a promising method for pattern recognition analysis of longitudinal neuroimaging data.

In the following chapter we apply our proposed longitudinal projection to diffusion MRI based Parkinson's disease discrimination, showing that our method can be applied to a different imaging modality and a different neurodegenerative disease discrimination problem. It also highlights the generality of LTC-PCA in the case where two study groups were scanned on three occasions (baseline, one year follow-up and two year follow-up) and one study group was only scanned twice (baseline and two year follow-up), leading to an unbalanced longitudinal design with varying time-points per subjects.

Chapter 8 Application to Early Parkinson's Disease

8.1 Introduction

In this chapter we analyse data from the PMMP longitudinal study tracking early PD subjects, subjects identified as being at risk of developing PD, along with matched controls, described in Chapter 3. Here we attempt pairwise discriminations of the three study groups using the study's longitudinal diffusion MRI data, aiming to identify patterns within measures of white matter structure that can be used to discriminate disease subjects from controls.

As in Chapter 7 we compare the performance of features based on cross-sectional information to those that incorporate longitudinal information via the feature construction method described in Chapter 6. Chapter 7 demonstrated our method's ability to handle varying longitudinal designs, i.e. with a fixed or varying number of samples per subject and/or time intervals between scans. Here we show that our method can be used on a different diagnostic problem and a different imaging modality. In particular, we will create features using fractional anisotropy (FA) samples that can be computed from diffusion tensor imaging (DTI) data. The question of which features are optimal for pattern recognition based analyses of DTI data has received little attention and remains an open question. Therefore, we investigate two common choices of FA based features to determine empirically the best choice of features for our data. In particular we explore the use of either whole brain FA features or skeletonized FA features along with the choice of two different minimum FA thresholds.

We do not present or further discuss the use of mean diffusivity (MD) based features, which are also commonly derived from DTI data. MD based features had chance level accuracy in all of the pairwise discriminations we attempt in this chapter. One possible explanation is that our DTI acquisition (described in Section 8.2.2) had only one b value, while Rae et al., (2012), who explored group differences in FA and MD between PD patients and controls, note that a DTI acquisition with multiple b values may be more sensitive to increases in MD.

8.2 Materials and Methods

8.2.1 Study Description

We used clinical and imaging data derived from the Progression Markers in the Premotor Phase (PMPP) study of Parkinson's disease, with study design and data collection described in Chapter 3.

8.2.2 Image Acquisition

Both the diffusion and structural MR images were acquired on a Siemens MR system operating at 3T (Magnetom Trio, Siemens Healthcare, Erlangen). As in Chapter 5, we used the sagittal magnetization-prepared rapid gradient-echo (MPRAGE) T1-weighted images (TR = 2300 ms; TE = 3 ms; TI = 1100 ms; flip angle = 8°; matrix size = 256 × 224 with 176 1.0 mm thick slices; in plane pixel size = 1 mm/1 mm; FoV = 256 × 224 mm). The DTI was acquired via a 2D spin-echo echo planar imaging (EPI) sequence (TR = 7500 ms; TE = 79 ms; matrix size = 128 × 116 with 61 2.0 mm thick slices; in plane pixel size = 2 mm/2 mm). Diffusion weighting was applied with $b = 800 \text{ s/mm}^2$ along six directions, with each direction acquired three times. Additional images with no diffusion weighting ($b = 0 \text{ s/mm}^2$) were acquired three times.

8.2.3 Image Preprocessing and Quality Control

Prior to forming FA images, we visually inspected the 151 raw DTI images. We found one particularly pronounced type of artefact in many images. It manifests as a loss of signal that depends on the direction of the diffusion weighting. Figure 8.1 depicts two volumes from a single image: one volume is clearly affected by the artefact, while the other, acquired in a different gradient direction, is not. These artefacts closely resemble the ones described by Gallichan et al., (2010), and as such may have been caused by systematic vibrations in the MRI machine's patient table due to low-frequency gradient switching. We also removed images that were uniformly corrupted by noise across all voxels and volumes, as shown in Figure 8.2. Altogether, we removed 60 images due to the direction-dependent artefact, four

images due to uniform noise and one image due to chemical shift artefact from insufficient fat suppression.

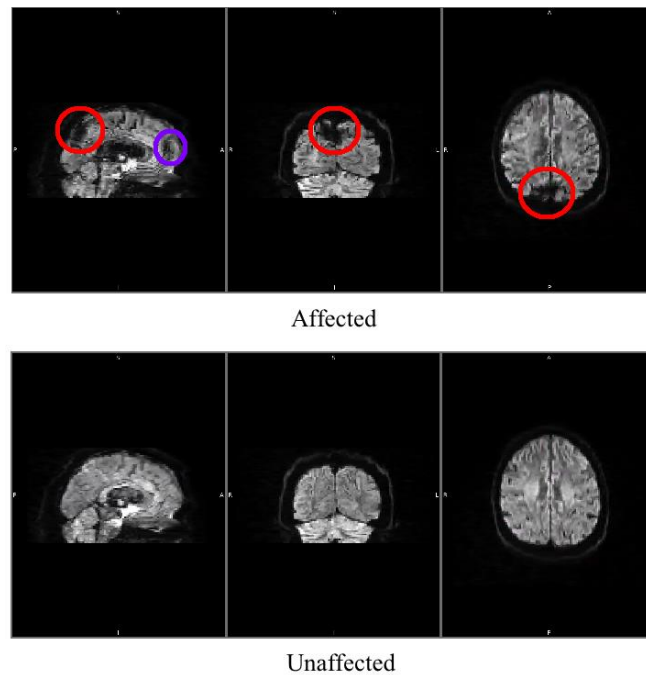


Figure 8-1 Two volumes from a single image that is affected by gradient direction specific artefacts. Artefacts often appear in the parietal lobe (circled in red) and, to a lesser extent, in the frontal lobe (circled in purple).

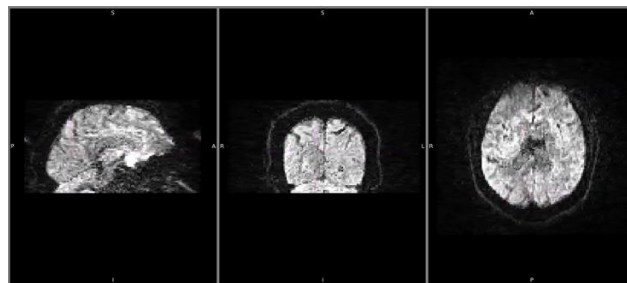


Figure 8-2 A uniformly noisy image.

Many of the remaining 86 images had consistent signal dropout across all volumes in the superior cerebral cortex and inferior cerebellum. To mitigate the effect of the signal loss, we used the ExploreDTI toolbox (Leemans et al., 2009) to preprocess these images with the aid of the relatively good quality T1-weighted images that were collected alongside each diffusion image. The preprocessing pipeline involved downsampling the T1-weighted images and rigidly registering them to the MNI T1 2mm template, with brain extraction and masking done using OptiBET (Lutkenhoff et al., 2014). ExploreDTI was then used to correct for subject motion, eddy current and EPI distortion artefacts. The diffusion images were non-

linearly coregistered to the MNI-registered T1 images during the EPI correction step. Finally, maps of fractional anisotropy (FA) were estimated in ExploreDTI using the corrected diffusion data. Fractional anisotropy is a measure of water diffusion anisotropy (i.e. directionality) at each voxel (Basser and Pierpaoli, 1996). It ranges from a value of zero in the case of unrestricted movement of water to a value of one in the case of water diffusion restricted to one direction in 3D space. FA measures of grey matter and cerebrospinal fluid (CSF) approach values of zero while FA measures of the white matter are high, approaching one in the most directionally organized areas, such as white matter tracts (Assaf and Pasternak, 2007). FA been used extensively to study the properties of white matter such as fibre density.

We visually inspected each FA image, finding 21 out of 86 FA maps that had excessive signal dropout or a bad coregistration to their respective T1 image. Examples of these problematic images, which were removed from further analysis, are provided in Figure 8.3. Following this quality control step there were 65 images out of the original 151 left for analysis.

We then tested for group differences in subjects' motion using the motion parameters saved in the ExploreDTI (*.mat) file from the subject motion/eddy current correction step. We used the decomposed measures of translation and rotation for each of the 18 non-B₀ volumes per image, finding the absolute maximum of rotation and translation separately. We then ran a one-way ANOVA to test for group differences in absolute rotation and translation, finding no significant differences in rotation ($F_{2,60} = 0.32$, $p = 0.73$) or translation ($F_{2,60} = 1.48$, $p = 0.24$). Figure 8.4 depicts the rotation and translation values across groups. We also tested for differences within each group across time-points. We found a significant difference in translation over time within the HR_{PD} group (rotation: PD: $F_{2,15} = 0.67$, $p = 0.53$, HR_{PD}: $F_{2,32} = 1.73$, $p = 0.19$, Healthy: $F_{1,8} = 0.88$, $p = 0.38$, translation: PD: $F_{2,15} = 0.81$, $p = 0.46$, HR_{PD}: $F_{2,32} = 4.23$, $p = 0.02$, Healthy: $F_{1,8} = 0.22$, $p = 0.65$). We present analysis with and without regressing out these two motion parameters from the whole-brain and TBSS based features.

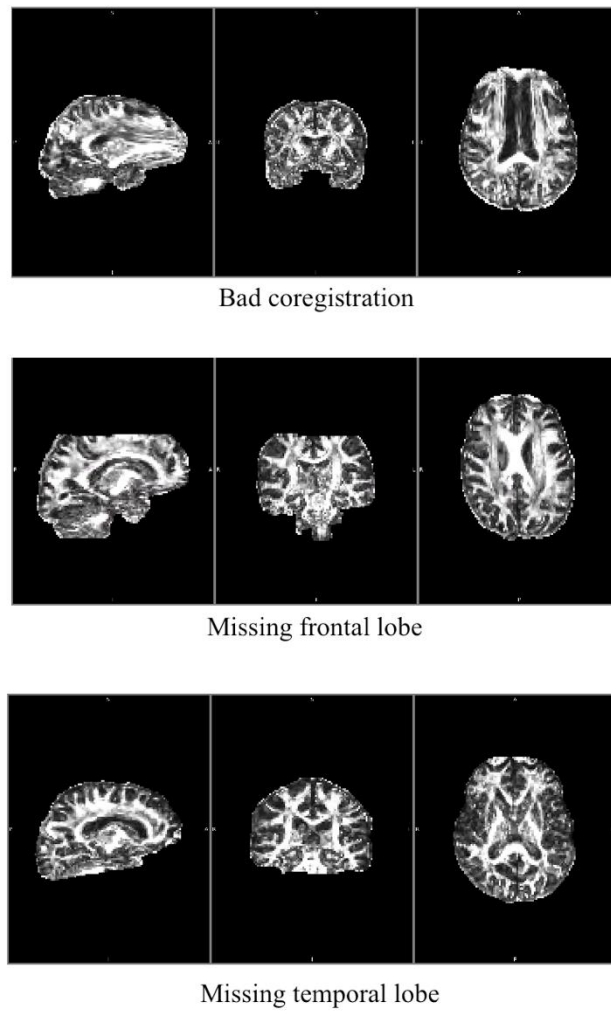


Figure 8-3 Examples of problematic FA maps.

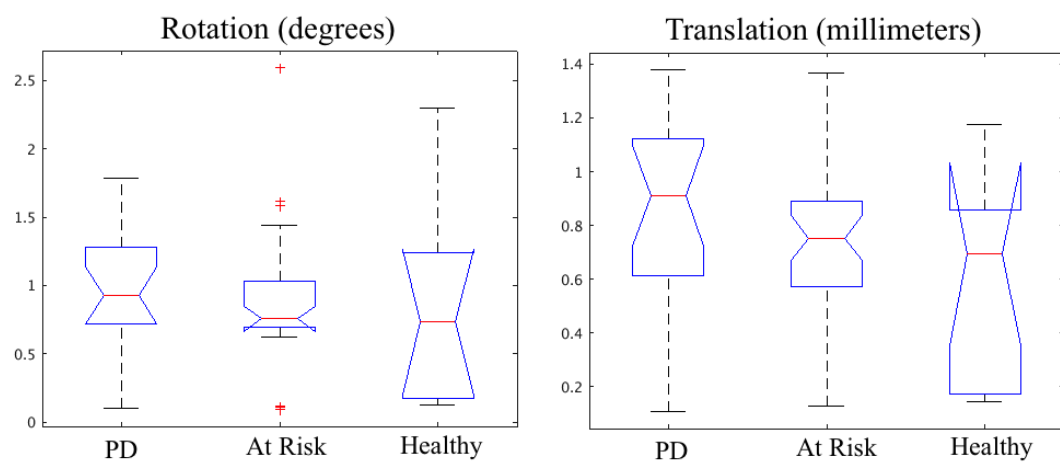


Figure 8-4 Maximum subject motion across the three groups.

8.2.4 Image Registration

We used the coregistered T1 images to register the 65 FA images that met our quality control criteria to the study template formed using DARTEL. To do so we followed part of the procedure described in Section 4.3: we first segmented the T1 images from all time points of all subjects into grey matter, white matter, cerebrospinal fluid (CSF), soft tissue and skull classes using SPM8's "New Segment" procedure. Grey matter and white matter tissue class images were then registered to a common inter-subject space, by creating a study specific template from all images using SPM8's DARTEL registration procedure (Ashburner, 2007). Finally, we applied the flow fields used to warp each individual T1 image to the template to the corresponding coregistered FA image, bringing the FA images into a common space. The resulting registered FA images were generally well registered to the white matter portion of the template: there were no significant group differences in registration quality both across the three diagnostic groups and within each group over time. Appendix D provides details of our registration quality analysis.

We used these registered FA images to form two different types of features that were used for classification separately. We formed whole-brain FA features as well as skeletonized FA features based on tract based spatial statistics (TBSS) (Smith et al., 2006). Typically, the TBSS analysis pipeline¹⁴ involves the formation of registered FA "maps" (i.e. images). In our case we performed registration separately, using T1 images as described above. There were two reasons for this: we wanted to facilitate comparison between skeletonized and whole-brain FA features¹⁵ and we were concerned that the signal dropout in portions of the DTI data would lead to an inferior FA driven registration.

Following registration we used part of the TBSS pipeline (described below) to form a mean FA skeleton, meant to represent the most likely locations of white matter tract centres across the available subjects, forming skeletonized FA features for each subject by projecting each subject's FA map onto the skeleton. The purpose of skeletonization is to reduce the variability in the comparison of FA between

¹⁴ A user guide for the TBSS analysis pipeline is available at:
<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/TBSS/UserGuide>

¹⁵ TBSS based registration is intended to align tract structure across subjects, with subsequent projection of data onto a tract skeleton, rather than aligning all voxels across subjects.

subjects by only comparing subjects in places that one is most confident is truly white matter. This feature construction procedure is intended to improve the robustness and sensitivity of any subsequent analysis performed on the skeletonized images (Smith et al., 2006). Note that the skeletonized data are then most commonly analysed using a general linear model (GLM) in a mass-univariate context, with the aim of understanding which voxels differ significantly between groups of subjects.

To form the whole-brain FA features for our multivariate analysis we used SPM to smooth the registered FA maps with an isotropic 4 mm full width at half maximum (FWHM) Gaussian kernel. Smoothing is often performed in voxel based morphometry (VBM) as a means of reducing the effect of the misalignment across subjects during registration, though the choice of kernel width is arbitrary and may affect subsequent analysis. As is customary when analysing FA images, to improve the confidence that one is analysing the same tissue type (namely white matter) across individuals using FA, we retained only those voxels in the smoothed images that had a specified minimum FA value across all subjects. This minimum value was either 0.2 or 0.3, as recommended by Smith et al., (2006), with results shown for both cases. To remove any possibility of using voxels outside the brain, we also formed a whole brain mask that retained voxels that had either at least 0.05 intensity in the DARTEL template's grey matter image or at least 0.05 in the template's white matter image as 0.05 intensity reflects a relatively high confidence that a voxel belongs to a particular tissue class.

To form skeletonized FA features we further preprocessed the (unsmoothed) registered FA maps using the DTI-TK instructions for using external registration with the TBSS procedure¹⁶, resampling them to 1 mm³ isotropic MNI152 space using FSL's command line tools (Smith et al., 2004b). As with the whole-brain features, we retained only those voxels with a minimum FA of 0.2 or 0.3 across all subjects and show results for both cases. There were no group differences in number of features discarded (see second part of Appendix D). However, we did not use the FA skeleton mask provided by TBSS (via the 'tbss_skeleton' command). Instead we formed a mask for the skeletonized features using the DARTEL template's white matter image, retaining the voxels that met a minimum threshold of at least 0.01.

¹⁶ DTI-TK's "Support for Tract-Based Spatial Statistics (TBSS)" page at: <http://dti-tk.sourceforge.net/pmwiki/pmwiki.php?n=Documentation.TBSS>

This mask excludes voxels on the outer edge of the cortex that are most likely be part of grey matter. As can be seen in Figure 8.5, the template based mask allows us to almost completely eliminate the “ring” around the edge of the brain that is a common artefact brought about by the skull stripping procedure.

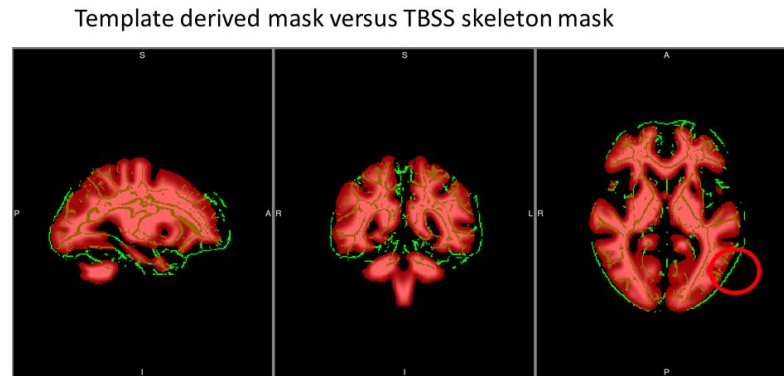


Figure 8-5 A comparison of the mask formed using the DARTEL template (white matter image, thresholded at 0.01 minimum), shown in red, versus the TBSS skeleton mask, shown in green. The two masks are overlaid onto the template's white matter image. The template based mask allows us to avoid the artefactual “ring” around the brain, circled in red.

8.2.5 Group Demographics

Because of the large number of images removed during the quality control procedure, we chose to discriminate subjects' disease state at their latest available time-point. Figure 8.6 shows the breakdown of the number of subjects that had one, two or three time-points' images per class. There are eleven PD subjects, 22 HR_{PD} subjects and eight healthy controls being discriminated with 44 images from 20 subjects (blue shaded area in figure) used to form the longitudinal subspace. Within the PD group we discriminated 2 subjects at baseline, one subject at 1st follow-up and eight subjects at 2nd follow-up. Within the HR_{PD} group we discriminated two subjects at baseline, seven subjects at 1st follow-up and 13 subjects at 2nd follow-up. Within the healthy control group, which did not have imaging at 1st follow-up, we discriminated three subjects at baseline and five subjects at 2nd follow-up.

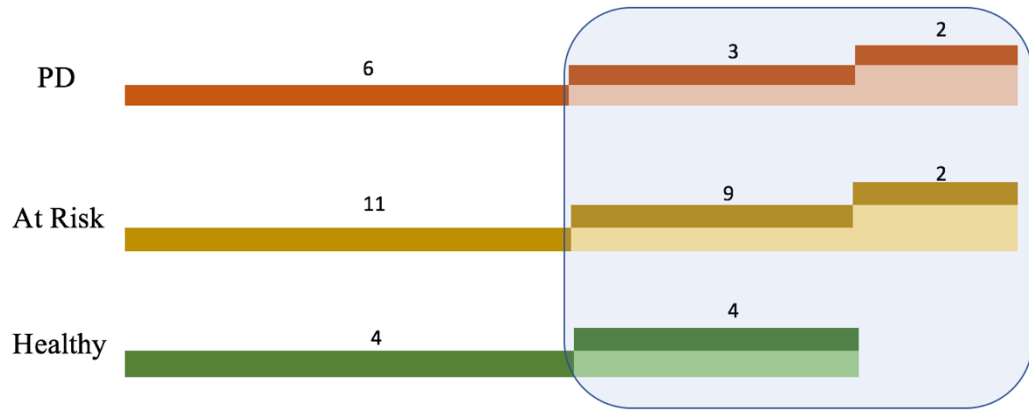


Figure 8-6 Breakdown of number of subjects per group with images at one time-point, two time-points and all three time-points. There are eleven PD subjects, 22 At Risk (i.e. HR_{PD}) subjects and eight healthy controls being discriminated at their last available time-point (as detailed in Section 8.2.5). 44 images from 20 subjects (in the blue shaded area) are used to form the longitudinal subspace.

Although we discriminated images from different time-points, to facilitate a fair comparison Table 8.1 depicts the demographic and clinical information at the final (two year) follow-up time-point for all 41 subjects being discriminated. Diagnostic groups did not differ significantly in terms of age ($F_{2,38} = 0.5$, $p = 0.63$) or sex (Fisher's exact test: 3.05, $p = 0.21$). In addition to any group differences in clinical measures, we are also potentially interested in group differences in (two year) longitudinal changes in cognitive and motor measure: e.g. the "MMSE change" row refers to the within subject (longitudinal) change between baseline and final follow-up (follow-up minus baseline) in MMSE.

There were no group differences in MMSE ($F_{2,38} = 1.35$, $p = 0.27$) or change in MMSE ($F_{2,38} = 2.7$, $p = 0.08$). There was a large group difference in motor function, measured by the total UPDRS-III score ($F_{2,38} = 67.3$, $p < 0.001$) but no significant change in UPDRS-III over time ($F_{2,38} = 1.00$, $p = 0.38$). Similarly, there were significant differences in tremor (UPDRS-III sum of items 3.17-3.18, $F_{2,38} = 12.4$, $p < 0.001$) and rigidity (UPDRS-III sum of items 3.3, $F_{2,38} = 38.7$, $p < 0.001$) but no significant changes in these over time (tremor change: $F_{2,38} = 0.9$, $p = 0.43$, rigidity change $F_{2,38} = 0.2$, $p = 0.82$). There were no significant cross-sectional or longitudinal differences in REM sleep behaviour disorder questionnaire total score (RBDQ).

Post hoc pairwise comparisons of these clinical measures between groups, all Bonferroni corrected, reveal significant differences in UPDRS-III score between the PD and both the HR_{PD} group ($p < 0.001$) and the control group ($p < 0.001$). Similarly, there were significant differences in tremor and rigidity between the PD group and both the HR_{PD} group (tremor: $p < 0.001$, rigidity: $p < 0.001$) and the control group (tremor: $p = 0.005$, rigidity: $p < 0.001$). There were no significant differences between the HR_{PD} and control groups in any measure with the exception that all HR_{PD} subjects were positive for substantia nigra hyperechogenicity (SN+), which was by design.

Table 8-1 Demographics of analysis subjects at final (two year) follow-up time-point.

		PD (n = 11)	HR _{PD} (n = 22)	HC (n = 8)
Age	(mean $\pm$ SD)	64.1 $\pm$ 8.8	62.8 $\pm$ 6.3	65.8 $\pm$ 8.6
Number of females	(n, %)	4 (36%)	6 (27%)	5 (63%)
Number of males	(n, %)	7 (64%)	16 (73%)	3 (38%)
Years between scans	(mean $\pm$ SD)	1.9 $\pm$ 0.3	1.8 $\pm$ 0.4	2.0 $\pm$ 0.1
Years since diagnosis	(mean $\pm$ SD)	4.8 $\pm$ 1.4	N/ A	N/A
SN+	(n, %)	10 (91%)	22 (100%)	0 (0%)
MMSE	(mean $\pm$ SD)	28.9 $\pm$ 0.7	29.4 $\pm$ 1.2	28.8 $\pm$ 1.4
MMSE change	(mean $\pm$ SD)	-0.5 $\pm$ 0.9	0.1 $\pm$ 1.1	-0.9 $\pm$ 1.4
UPDRS-III	(mean $\pm$ SD)	26.9 $\pm$ 11.4	1.9 $\pm$ 2.1	1.3 $\pm$ 2.4
UPDRS-III change	(mean $\pm$ SD)	-3.1 $\pm$ 11.1	-1.7 $\pm$ 3.9	1.1 $\pm$ 2.5
Tremor	(mean $\pm$ SD)	2.7 $\pm$ 2.8	0.0 $\pm$ 0.0	0.4 $\pm$ 1.1
Tremor change	(mean $\pm$ SD)	-0.5 $\pm$ 2.4	-0.1 $\pm$ 0.4	0.3 $\pm$ 1.0
Rigidity	(mean $\pm$ SD)	4.2 $\pm$ 2.5	0.2 $\pm$ 0.4	0.0 $\pm$ 0.0
Rigidity change	(mean $\pm$ SD)	-0.3 $\pm$ 3.2	-0.5 $\pm$ 0.9	0.0 $\pm$ 0.0
RBDQ	(mean $\pm$ SD)	3.8 $\pm$ 3.3	2.1 $\pm$ 2.3	1.1 $\pm$ 1.4
RBDQ change	(mean $\pm$ SD)	0.2 $\pm$ 2.7	-0.2 $\pm$ 1.2	-0.5 $\pm$ 1.1

8.2.6 Analysis Methodology

We formed purely cross-sectional features using whole brain FA and, separately, skeletonized FA samples at the final follow-up time-point. For both types of

samples, we also formed longitudinally projected cross-sectional features via Longitudinal Trajectory Coefficient PCA (LTC-PCA). Recalling Sections 6.3-6.7, LTC-PCA refers to PCA performed on the matrix of longitudinal trajectory coefficients (in this thesis: the first order coefficient matrix $\mathbf{B}^{(1)}$) when there is a varying number of longitudinal samples per subject, as here with two or three samples per subject.

As in Chapters 5 and 7, we assess the cross-validated predictive performance of these features using the C cost support vector classifier (SVC) described in Section 4.4.2 using the LIBSVM (Chang and Lin, 2011) implementation of the SVC with a fixed value of $C = 1$ throughout all classifications (see Section 7.2.4). As in the previous chapter, we perform nested leave one out cross-validation (LOO-CV) to tune the number of retained eigenslopes via an inner LOO-CV procedure and predict the label of held-out images via an outer LOO-CV procedure (see Section 6.12). We similarly report the balanced accuracy, sensitivity, specificity, positive predictive value (*PPV*), negative predictive value (*NPV*) and a summary of the receiver operating characteristic's area under the curve (*ROC AUC*) and perform Monte Carlo tests to assess the statistical significance of the balanced accuracies we report (see Section 3.2).

8.3 Results

We present four sets of results of here: Tables 8.7 and 8.8 present pairwise discriminations of the three subject groups using motion corrected and uncorrected *whole-brain* FA features while Tables 8.9 and 8.10 presents the same discriminations performed using motion corrected and uncorrected *skeletonized* FA features. In both cases we discriminated the disease class label (PD subjects vs. controls, for example) using the latest available image, with minimum FA value across subjects thresholded at either 0.2 or 0.3.

8.3.1 Whole-brain and skeletonized features

None of the features in any of the four tables discriminated HR_{PD} subjects from controls with better than chance level balanced accuracy. Similarly, PD subjects could not be discriminated from HR_{PD} subjects better than chance. LTC-PCA projection resulted in significant discriminations of PD subjects from controls in four out of eight cases (motion uncorrected whole-brain 0.2 threshold: 84.7% $p < 0.01$, motion corrected skeletonized 0.3 threshold: 80.1% $p < 0.01$, motion uncorrected skeletonized 0.2 threshold: 78.4% $p < 0.05$ and 0.3 threshold: 84.7% $p < 0.01$).

8.3.2 Statistical Comparison of Feature Construction Factors

As in Chapter 7, there is limited statistical power to detect differences between feature construction methods within comparisons due to the small samples sizes in this exploratory biomarker study. Using McNemar's test to compare classifier predictions we do not see statistically significant differences in pairwise comparisons across the three feature construction factors (effect of projection, feature style (whole-brain/skeletonized) and minimum FA threshold) in any of the three discriminations on an individual basis. However, we separately pooled the 24 balanced accuracies from the motion corrected features (Tables 8.2 and 8.4) and uncorrected (Table 8.3 and 8.5) to investigate the effect of the three factors across discriminations. We tested the two feature construction factors we were primarily interested in, namely the effect of the LTC-PCA projection and feature style, using a two-level repeated measures ANOVA. The four groups being tested each contained six samples (three discrimination accuracies from each of the two minimum FA thresholds).

In the motion corrected features there was no significant interaction between these two effects ($F_{1,5} = 0.54$, $p = 0.50$) nor was there an effect due to LTC-PCA projection ($F_{1,5} = 0.09$, $p = 0.77$) or feature style ($F_{1,5} = 0.55$, $p = 0.49$). Similarly, in the motion uncorrected features there was no significant interaction between effects ($F_{1,5} = 0.10$, $p = 0.76$) and no individually significant effects (projection: $F_{1,5} = 2.3$, $p = 0.19$, style: $F_{1,5} = 0.83$, $p = 0.41$). As we could not perform a full three-way repeated measures ANOVA due to a lack of samples per group, we performed an additional paired t-test on the effect of the FA threshold (0.2 or 0.3) across all

discriminations (both corrected and uncorrected), finding no significant effect ($t = -0.24$, $p = 0.82$).

We performed two other post hoc analyses via paired t-tests: in the first, we tested the effect of motion correction across all 24 pairs of discriminations, finding a mean decrease in balanced accuracy of $1.1 \pm 9.8\%$ due to motion correction, which was not significant ($t = -0.56$, $p = 0.58$). Finally, we tested whether the projection had an effect within the PD versus healthy controls discrimination, which was the only discrimination that exceeded chance level in some cases. Across the eight pairs (feature styles by thresholds by motion correction toggle) there was an increase in balanced accuracy of $16.4 \pm 13.2\%$, which was significant ($t = 3.5$, $p = 0.01$).

Table 8-2 Motion corrected whole-brain features used in three pairwise discriminations of the subject groups. Each element in the table shows the metrics for a minimum FA threshold of 0.2 or 0.3 across subjects (each cell shows 0.2 result/0.3 result).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>PD (n=11) vs Controls (n=8)</i>						
Cross-sectional	50.6 / 56.8	63.6 / 63.6	37.5 / 50.0	58.3 / 63.6	42.9 / 50.0	0.523 / 0.750
Cross-sectional, LTC-PCA Projected	64.8 / 67.6	54.5 / 72.7	75.0 / 62.5	75.0 / 72.7	54.5 / 62.5	0.818 / 0.750
<i>HR_{PD} (n=22) vs Controls (n=8)</i>						
Cross-sectional	51.7 / 55.7	90.9 / 86.4	12.5 / 25.0	74.1 / 76.0	33.3 / 40.0	0.739 / 0.756
Cross-sectional, LTC-PCA Projected	49.4 / 51.7	86.4 / 90.9	12.5 / 12.5	73.1 / 74.1	25.0 / 33.3	0.812 / 0.767
<i>PD (n=11) vs HR_{PD} (n=22)</i>						
Cross-sectional	54.5 / 52.3	27.3 / 27.3	81.8 / 77.3	42.9 / 37.5	69.2 / 68.0	0.579 / 0.566
Cross-sectional, LTC-PCA Projected	40.9 / 45.5	0.0 / 9.1	81.8 / 81.8	0.0 / 20.0	62.1 / 64.3	0.355 / 0.471

* Statistically significant balanced accuracy, Monte Carlo test p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo test p-value < 0.01

Table 8-3 Motion uncorrected whole brain features used in three pairwise discriminations of the subject groups. Each element in the table shows the metrics for a minimum FA threshold of 0.2 or 0.3 across subjects (each cell shows 0.2 result/0.3 result).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>PD (n=11) vs Controls (n=8)</i>						
Cross-sectional	61.4 / 59.7	72.7 / 81.8	50.0 / 37.5	66.7 / 64.3	57.1 / 60.0	0.739 / 0.807
Cross-sectional, LTC-PCA Projected	84.7** / 56.8	81.8 / 63.6	87.5 / 50.0	90.0 / 63.6	77.8 / 50.0	0.807 / 0.705
<i>HR_{PD} (n=22) vs Controls (n=8)</i>						
Cross-sectional	49.4 / 49.4	86.4 / 86.4	12.5 / 12.5	73.1 / 73.1	25.0 / 25.0	0.705 / 0.739
Cross-sectional, LTC-PCA Projected	55.7 / 64.2	86.4 / 90.9	25.0 / 37.5	76.0 / 80.0	40.0 / 60.0	0.778 / 0.812
<i>PD (n=11) vs HR_{PD} (n=22)</i>						
Cross-sectional	45.5 / 50.0	9.1 / 9.1	81.8 / 90.9	20.0 / 33.3	64.3 / 66.7	0.409 / 0.446
Cross-sectional, LTC-PCA Projected	50.0 / 50.0	0.0 / 0.0	100.0 / 100.0	0.0 / 0.0	66.7 / 66.7	0.281 / 0.554

* Statistically significant balanced accuracy, Monte Carlo test p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo test p-value < 0.01

Table 8-4 Motion corrected TBSS features used in three pairwise discriminations of the subject groups. Each element in the table shows the metrics for a minimum FA threshold of 0.2 or 0.3 across subjects (each cell shows 0.2 result/0.3 result).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>PD (n=11) vs Controls (n=8)</i>						
Cross-sectional	50.6 / 50.6	63.6 / 63.6	37.5 / 37.5	58.3 / 58.3	42.9 / 42.9	0.511 / 0.500
Cross-sectional, LTC-PCA Projected	52.3 / 80.1**	54.5 / 72.7	50.0 / 87.5	60.0 / 88.9	44.4 / 70.0	0.557 / 0.818
<i>HR_{PD} (n=22) vs Controls (n=8)</i>						
Cross-sectional	56.2 / 62.5	100 / 100	12.5 / 25.0	75.9 / 78.6	100 / 100	0.812 / 0.824
Cross-sectional, LTC-PCA Projected	54.0 / 55.7	95.5 / 86.4	12.5 / 25.0	75.0 / 76.0	50.0 / 40.0	0.710 / 0.693
<i>PD (n=11) vs HR_{PD} (n=22)</i>						
Cross-sectional	52.3 / 47.7	18.2 / 18.2	86.4 / 77.3	40.0 / 28.6	67.9 / 65.4	0.545 / 0.529
Cross-sectional, LTC-PCA Projected	47.7 / 47.7	0.0 / 9.1	95.5 / 86.4	0.0 / 25.0	65.6 / 65.5	0.256 / 0.517

* Statistically significant balanced accuracy, Monte Carlo test p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo test p-value < 0.01

Table 8-5 Motion uncorrected TBSS features used in three pairwise discriminations of the subject groups. Each element in the table shows the metrics for a minimum FA threshold of 0.2 or 0.3 across subjects (each cell shows 0.2 result/0.3 result).

Features	Bal. Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	ROC AUC
<i>PD (n=11) vs Controls (n=8)</i>						
Cross-sectional	59.7 / 48.9	81.8 / 72.7	37.5 / 25.0	64.3 / 57.1	60.0 / 40.0	0.693 / 0.716
Cross-sectional, LTC-PCA Projected	78.4* / 84.7**	81.8 / 81.8	75.0 / 87.5	81.8 / 90.0	75.0 / 77.8	0.875 / 0.898
<i>HR_{PD} (n=22) vs Controls (n=8)</i>						
Cross-sectional	54.0 / 51.7	95.5 / 90.9	12.5 / 12.5	75.0 / 74.1	50.0 / 33.3	0.744 / 0.756
Cross-sectional, LTC-PCA Projected	49.4 / 44.9	86.4 / 77.3	12.5 / 12.5	73.1 / 70.8	25.0 / 16.7	0.716 / 0.568
<i>PD (n=11) vs HR_{PD} (n=22)</i>						
Cross-sectional	47.7 / 47.7	0.0 / 0.0	95.5 / 95.5	0.0 / 0.0	65.6 / 65.6	0.455 / 0.376
Cross-sectional, LTC-PCA Projected	45.5 / 36.4	0.0 / 0.0	90.9 / 72.7	0.0 / 0.0	64.5 / 59.3	0.521 / 0.326

* Statistically significant balanced accuracy, Monte Carlo test p-value < 0.05

** Statistically significant balanced accuracy, Monte Carlo test p-value < 0.01

8.3.3 Discriminative and Forward Brain Maps

Figures 8.7-8.10 display t-statistic images, weight maps and forward maps (discussed in Section 4.7) for cross-sectional features and LTC-PCA projected cross-sectional features used in discriminating PD subjects from healthy controls (Figures 8.7 and 8.8) and HR_{PD} subjects from healthy controls (Figures 8.9 and 8.10). In all cases we display maps from motion corrected features as these were similar to the uncorrected versions.

Figures 8.7 and 8.8 show similar patterns of FA differences between PD subjects and controls in both the whole-brain and skeletonized versions of the projected maps. As noted in the previous section, this discrimination was the only one that achieved statistical significance in some cases. The three types of maps in Figure 8.7 show a decrease in FA in PD subjects relative controls in the posterior and superior areas of the brainstem along with an increase in the mid-anterior of the brainstem. Additionally, there is some decrease in FA in the corpus collosum and strong FA increase in an area of the right cortical white matter. There are similar patterns in the corresponding skeletonized maps in Figure 8.8.

The maps in Figures 8.9 and 8.10, from the discrimination of HR_{PD} subjects from controls, show a diffuse and widespread pattern of differences across the four feature variants, reflecting the difficulty of this discrimination. Importantly, the brainstem does not appear to play a prominent role in this discrimination.

PD vs HC

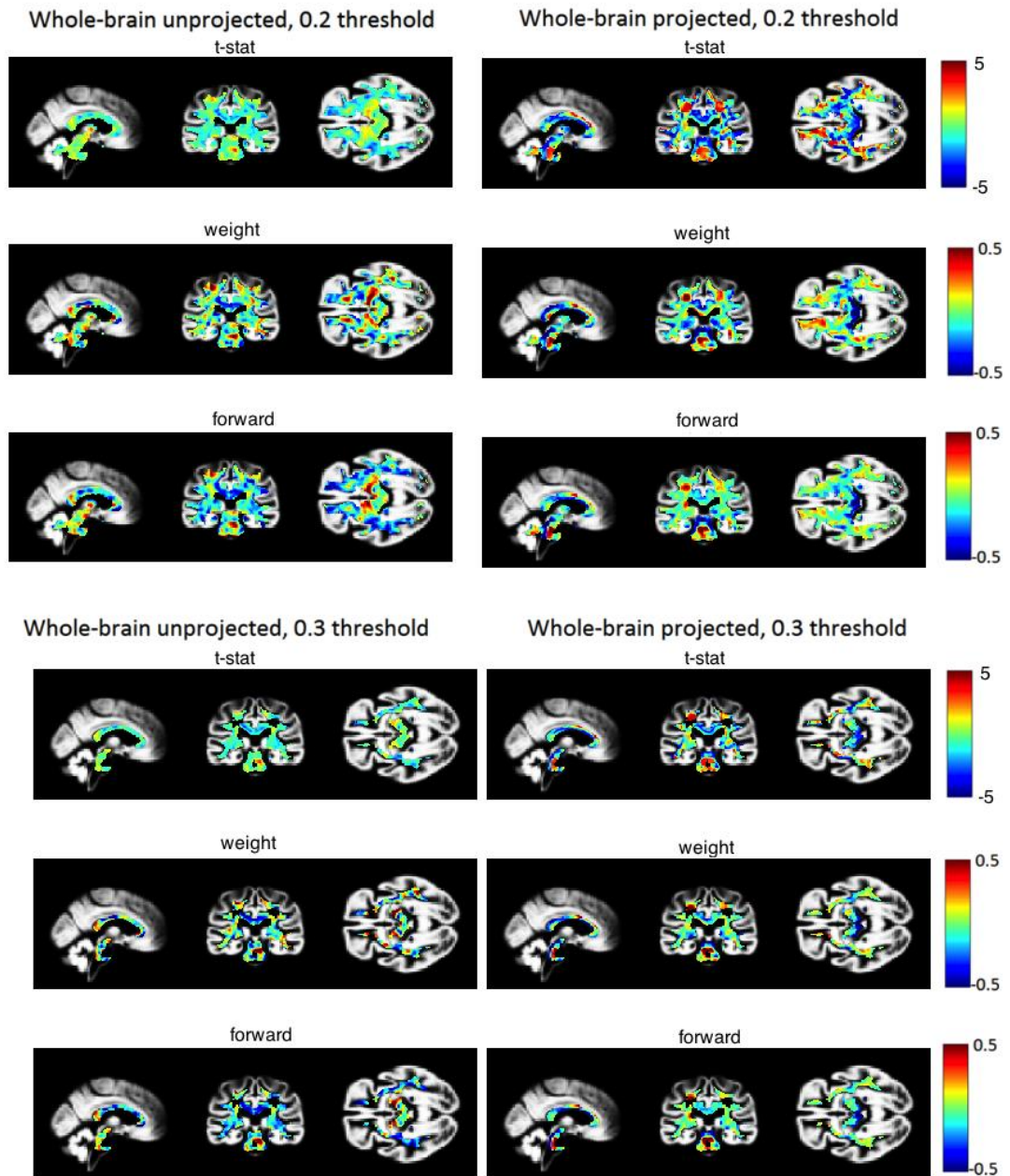


Figure 8-7 Discriminating PD subjects from controls. Sets of t-statistic, weight and forward maps from classifiers built using whole-brain, motion corrected FA features varying the two other feature construction factors (projection, minimum FA).

PD vs HC

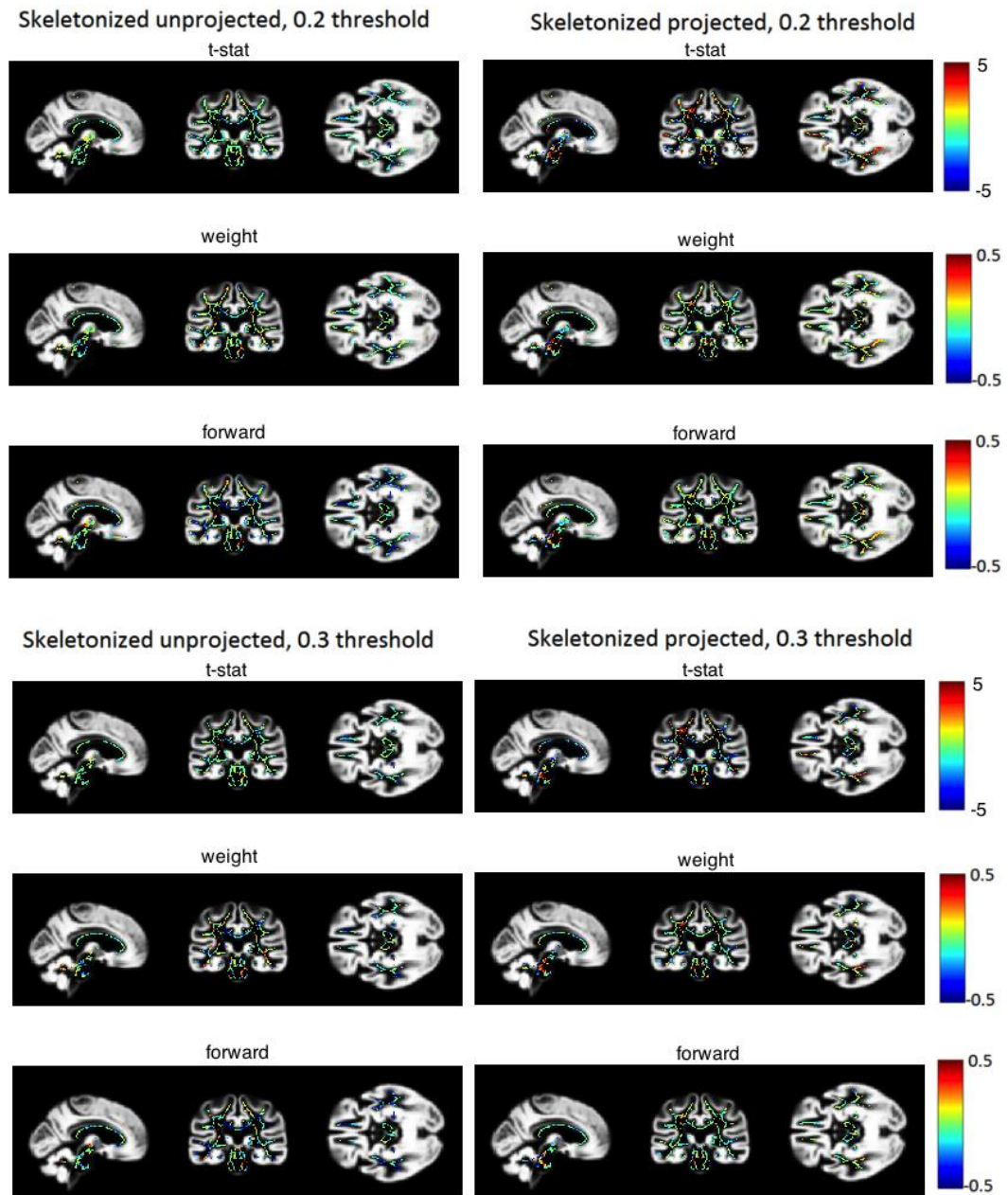


Figure 8-8 Discriminating PD subjects from controls. Sets of t-statistic, weight and forward maps from classifiers built using skeletonized, motion corrected FA features varying the two other feature construction factors (projection, minimum FA).

HR_{PD} vs HC

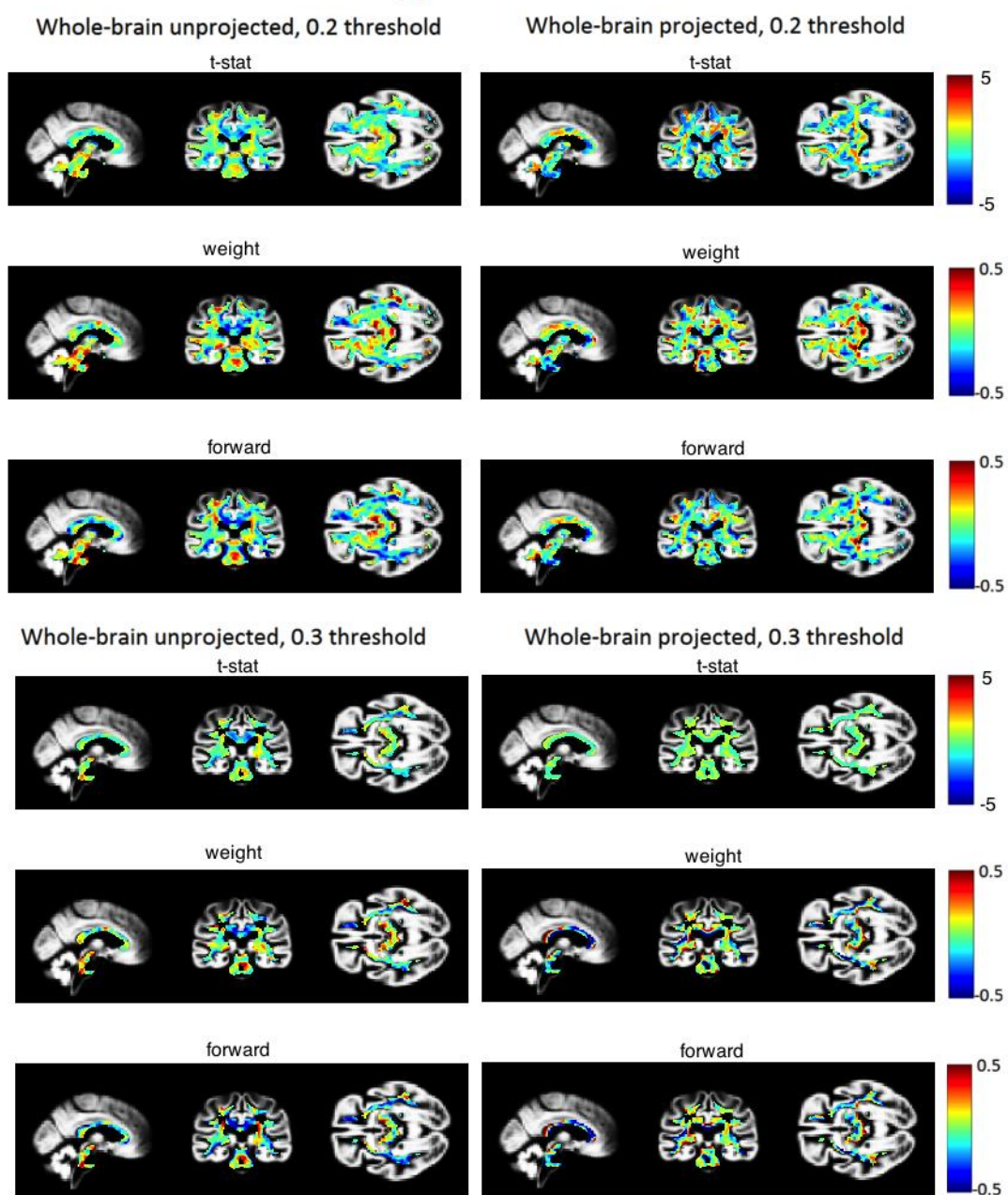


Figure 8-9 Discriminating HR_{PD} subjects from controls. Sets of t-statistic, weight and forward maps from classifiers built using whole-brain, motion corrected FA features varying the two other feature construction factors (projection, minimum FA).

HR_{PD} vs HC

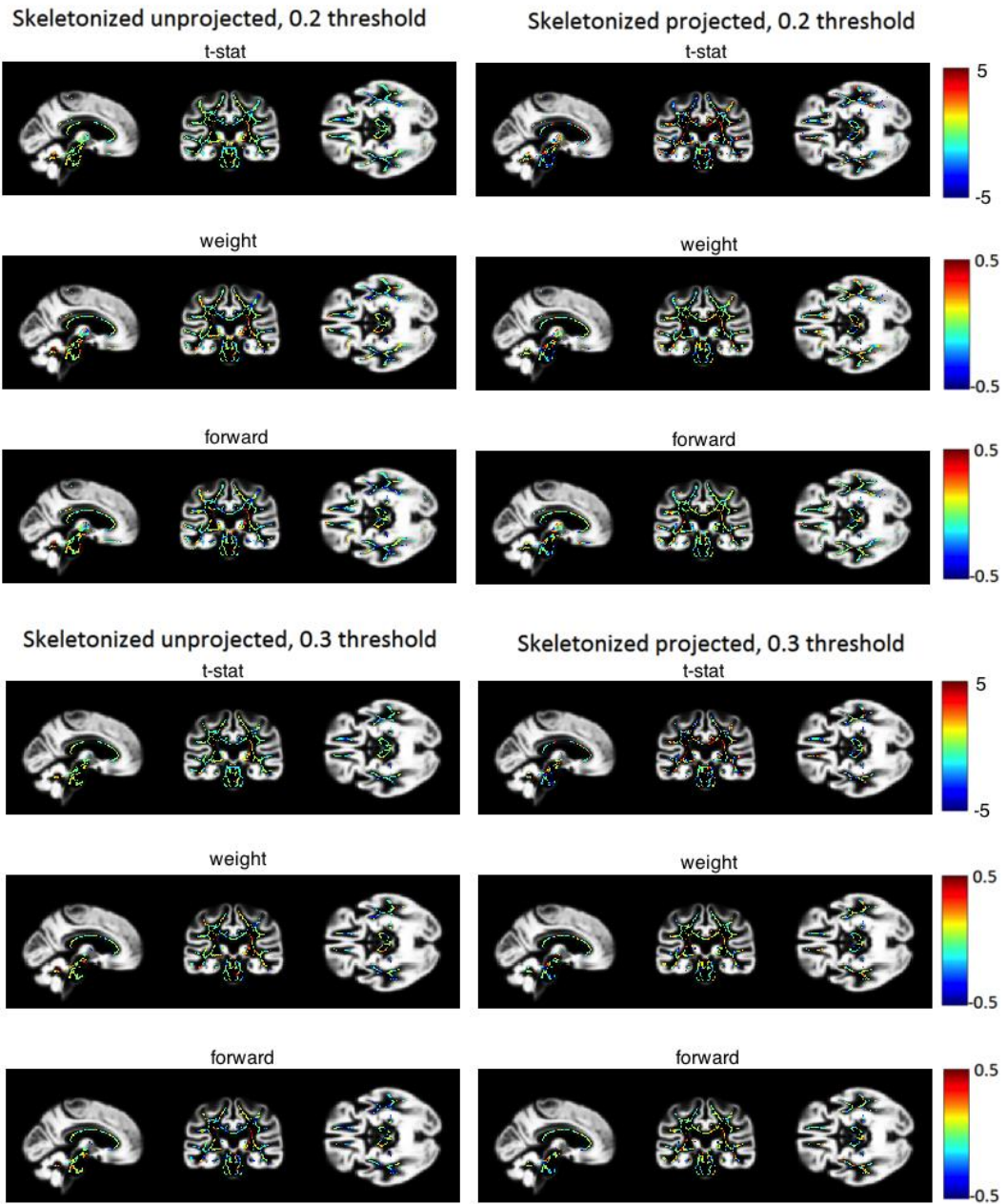


Figure 8-10 Discriminating HR_{PD} subjects from controls. Sets of t-statistic, weight and forward maps from classifiers built using skeletonized, motion corrected FA features varying the two other feature construction factors (projection, minimum FA).

8.4 Discussion

In this chapter, we evaluated the effect of LTC-PCA projection for discriminating PD subjects, those at risk of PD and healthy controls using DTI based features. Unlike chapter 7, there was no significant improvement due to our projection across the three pairwise discriminations. Furthermore, we showed no significant effects due to the choices of FA feature style (whole-brain or skeletonized) or minimum FA threshold (0.2 or 0.3). As a post hoc test we showed a significant improvement in balanced accuracy due to our projection when discriminating PD subjects from controls. Overall, we could not discriminate HR_{PD} subjects from controls nor PD subjects from HR_{PD} subjects.

A previous analysis of the group differences in many of the basic and extended data set biomarkers collected in this study found significant differences in autonomic function between the HR_{PD} group and controls and between the early PD group and controls (Liepelt-Scarfone et al., 2015). In particular, urinary function (not explained by gender differences), a drop in blood pressure (BP) when standing, and the overall number of dysautonomic symptoms significantly differentiated HR_{PD} subjects from controls and early PD subjects from controls. The study also reported no significant group differences between the HR_{PD} and early PD groups across the various biomarkers. The image based analysis we present here is in line with the latter findings. A possible explanation for this, as noted in the description of the study design, is that the clinical profile of the HR_{PD} group overlaps that of the PD group across many measures. It is important to note however that overall these two groups are distinctly different across all diagnostic criteria.

When discriminating PD subjects from controls we described a pattern of FA changes within the brainstem along with some increased FA in cortical white matter. That this pattern was not observed when discriminating HR_{PD} subjects from controls may mean one of at least three different things: (i) if the HR_{PD} group does indeed represent prodromal PD then the changes in FA occur after the prodromal stage; (ii) the changes in FA at the prodromal stage are subtle and may need larger sample sizes to be detected or (iii) the HR_{PD} subjects do not represent prodromal PD and are either healthy or a heterogeneous mixture of various types of early neurodegeneration. The autonomic differences between HR_{PD} subjects and controls,

which have been associated with prodromal PD, provide some evidence against the latter possibility. Further to the second point, we discarded 86 out of 151 images due to artefacts and signal dropout, which may have limited our ability to build a robust longitudinal subspace (see Chapter 7) that could have detected early stage pathology. A great deal of further work is needed to understand the progression and localization of changes to white matter structure within Parkinson's disease.

Chapter 9 Conclusions

9.1 Summary

The goal of this PhD project was to develop methods to better discriminate early stage neurodegeneration by making use of longitudinal information. The preliminary analysis we presented in Chapter 5, which was done early in the course of the project, showed that applying current pattern recognition methods to cross-sectional and longitudinal features derived from existing approaches have difficulty in such discriminations (see Tables 5.4 and 5.5). We noted that there were several potential options for trying to improve on this analysis, such as, for instance, finding novel whole-brain or disease-specific region of interest (ROI) based features. As reviewed in Chapter 2, there is already a great deal of ongoing research in this area, with a variety of different types of features based on a variety of imaging modalities. Instead, we focussed on capitalizing on longitudinal information, viewing it as one of the most promising ways to capture neurodegenerative disease related effects that are marked by changes over time.

Throughout this thesis, we principally made use of data from two ongoing clinical studies, both having a longitudinal study design, with longitudinal neuroimaging data available in each (see Chapter 3) in addition to a publicly available longitudinal dataset. Longitudinal studies are a natural choice for understanding early neurodegeneration, which is marked by subtle within-subject changes in brain anatomy and cognitive function. Our literature review was mostly in line with our preliminary findings, however, noting that longitudinal features typically fare no better at discriminating neurodegeneration than cross-sectional features. Several studies have noted that this may be because longitudinal features provide complementary information that is best used in combination with, rather than in place of, cross-sectional information (McEvoy et al., 2011; Gray et al., 2012). We also discovered that there are few pattern recognition based methods that are specifically designed for longitudinal data. We set out to develop such a method.

The approach we developed and presented in Chapter 6 has some appealing properties: (i) it is a feature construction method and as such it does not depend on the image modality, image preprocessing or subsequent pattern recognition

algorithm it is used with; (ii) it can handle both balanced and unbalanced longitudinal designs as well as an arbitrary number of longitudinal measurements per subject; (iii) it allows both diagnostic and prognostic predictions and (iv) it allows for the transfer of longitudinal information from one set of subjects to make predictions in another set of subjects that may only have cross-sectional information.

Chapter 7 revisited the problem of discriminating MCI subjects from controls using Jacobian determinant features extracted from structural MRI, as in Chapter 5. This time we used our proposed method to create longitudinally projected features that capitalised on longitudinal information and thereby significantly outperformed purely cross-sectional features in terms of balanced accuracy on the HNRS dataset (see Table 7.1 and Section 7.3.1). We sought to validate this finding using the OASIS dataset, a publicly available “early dementia” dataset. Unlike the HNRS dataset that has a balanced two-sample per subject design (see Section 6.4), this dataset has an unbalanced design with between two and five measurements per subject. To compare to the MCI discrimination, we mimicked a two sample balanced design and showed a similar, though smaller, improvement in balanced accuracy due to the projection (see Table 7.3 and Section 7.3.3). We were also able to test the effect of long and short follow-up times between scans as well as the use of all available scans to better estimate the slope coefficient matrix that is needed to form our projection. We found that longer follow-up times between two scans had higher accuracy than shorter follow-up times and that, unsurprisingly, using all available scans per subject resulted in higher balanced accuracy than using any two scans (see Table 7.4 and Section 7.3.4). Chapter 7 also used the HNRS dataset to test the longitudinal information transferring ability of our method. While our method produced accuracy improvements over unprojected (i.e cross-sectional) features of up to 15% these did not achieve statistical significance due to the small sample size (Table 7.2. and Section 7.3.2).

The small sample sizes in the five discriminations performed in Chapter 7 limited our ability to detect differences between feature construction methods within any individual discrimination (with the exception of Table 7.1). However, we showed a significant improvement in both balanced accuracy and ROC AUC across all discriminations that can be attributed to our projection (see Section 7.3.7). Furthermore longitudinally projected features performed significantly better than

cross-sectionally projected features, showing that the predictive performance improvements of our method are not simply due to dimensionality reduction.

Chapter 8 applied our method to pairwise discriminations of the PMPP dataset's three Parkinson's disease (PD) groups (early PD subjects, those at high risk of developing PD and healthy controls). Instead of structural MRI though we used the available diffusion tensor imaging (DTI) data, extracting fractional anisotropy (FA) based features. In addition to testing whether our longitudinal projection significantly improves over cross-sectional features, we were also interested in the effect of two choices that can be made when forming FA based features: feature style (whole-brain versus skeletonized) and minimum FA threshold (0.2 or 0.3). Due, again, to the small number of samples per discrimination, we analysed the effect of these three factors (projection, feature style and minimum threshold) across all discriminations, finding the significant no effect of our projection on balanced. Neither of the other two factors were significant in any of our tests. Additionally, we performed a post hoc test on the effect of motion correction on balanced accuracies, finding no significant effect. In general, subject motion caused by resting tremor is an important consideration when discriminating PD subjects. As resting tremor may become more pronounced over time as the disease progresses, both tremor and change in tremor over time are important when studying PD longitudinally. In the analysis we presented, however, images seemed to have been more affected by direction-dependent artefacts caused by gradient-switching rather than disease-associated tremor.

Finally, we performed a post hoc test that showed a significant improvement due to longitudinal projection when discriminating PD subjects from controls. Further work is needed to confirm this finding and to explore whether a more robust longitudinal subspace would aid in discriminating HR_{PD} subjects from controls. Overall, this chapter demonstrated our method's flexibility by applying it to a different imaging modality, image preprocessing and disease discrimination problem than we had previously.

The latter application chapters demonstrated most of the positive aspects of our method. However we were not able to demonstrate its ability to make prognostic predictions by projecting baseline information onto the longitudinal subspace to

predict a follow-up diagnosis. Both studies lacked the necessary data: very few subjects had a change in diagnostic label between time-points (e.g. a healthy control converting to MCI between baseline and follow-up), possibly due to the studies' durations or sample sizes.

As a consequence of our use of a linear projection in this thesis (though this is not strictly necessary), we retained the ability to visualize a classifier's decision function (i.e. weight maps) and corresponding group differences (i.e. forward maps) as voxel intensities in brain images (see Sections 6.10 and 4.9). We formed these maps from the SVM classifiers we trained in Chapters 7 and 8 and, furthermore, make a qualitative comparison between maps from classifiers that used projected features to those that used purely cross-sectional features. We see in Figure 7.4, which corresponds to the MCI discrimination of Table 7.1, that the projection-based forward map had greater negative group differences in parietal and precuneous areas, the frontal pole and, to a lesser extent, the temporal lobe than the cross-sectional forward map. This pattern of group differences, indicating tissue contraction in disease subjects relative to controls in these areas, is very similar to the pattern of cortical thinning seen in Figure 4 of Frisoni et al., (2010) that is based on a difference between baseline and one-year follow-up images of MCI subjects. In other words, the group differences that are being enhanced by our longitudinal projection are consistent with the literature on longitudinal changes in MCI subjects.

Figure 7.5 shows the patterns of group differences for the discrimination of "very mild dementia" that corresponds to Table 7.3. Both the projection-based and cross-sectional forward maps had very similar patterns of group differences with strong positive group differences in the lateral ventricles and Sylvian fissure, indicating a regional shrinkage of tissue in disease subjects relative to controls. This pattern of longitudinal changes in Alzheimer's disease subjects, marked by the expansion of the lateral ventricles and contraction of cortical grey matter, is also consistent with the literature (Barnes et al., 2008; Nestor et al., 2008; Anderson et al., 2012; Ziegler et al., 2015).

Comparing the two figures, we see the marked expansion of the ventricles in the forward maps of Figure 7.5 is altogether absent from Figure 7.4 and, in turn, the parietal lobe contraction in Figure 7.4 is absent from Figure 7.5. MCI is associated

with a clinical dementia rating (CDR) of 0.5, which was used to form the disease class in the discrimination of Table 7.3, although it is known there is considerable heterogeneity in rates of progression to greater dementia severity within subjects with score of CDR 0.5 (Morris et al., 2001). The very different patterns we see suggest a different, perhaps earlier, stage of pathology in the MCI disease group of the HNRS study, which is explicitly composed of non-demented subjects, compared to the “very mild dementia” group of the OASIS study, which is based purely on clinical assessment, is not intended to diagnose MCI and may contain subjects in the early stages of dementia.

We can similarly compare the forward maps associated with the pairwise discriminations of early PD, high risk, and control subjects in Chapter 8. As we noted in Section 8.3.4, the three types of maps from the early PD versus controls discrimination (Figures 8.7 and 8.8, projected features) show similar patterns of FA changes in the brainstem (both increases and decreases), along with FA increase in cortical white matter. Similar patterns were not observed when discriminating HR_{PD} subjects from controls; we suggested several potential explanations (see discussion in Section 8.4).

9.2 Novel contribution of this thesis

The primary novel contribution of this thesis is a method for constructing features for pattern recognition analysis of neuroimaging data. We showed the value of this method in several contexts and across multiple datasets. The components of our method are not novel: longitudinal trajectories are often modelled as polynomial functions of time and principal component analysis (PCA) is frequently used in neuroimaging to reduce dimensionality or to find a subspace that is assumed to generalize to unseen data. Instead, the novelty lies in the feature construction procedure as a whole and the hypothesis that it represents. That is, that cross-sectional features lying in a longitudinal subspace are better suited to discriminating disease marked by change over time (such as neurodegenerative diseases) than simply using the cross-sectional features themselves. The explicit testing of this hypothesis in Chapter 7 and the finding that this projection results in a statistically significant improvement in discrimination accuracy is a novel and potentially

interesting finding. The application of our method to several different neuroimaging modalities, several disease discrimination problems and a variety of longitudinal study designs, all demonstrated in Chapters 7 and 8, is, of course, also novel.

Finally, in our preliminary analysis (see Chapter 5) we showed that a multi-class classifier can, with an accuracy that is significantly above chance, simultaneously discriminate subjects in one of three different stages of PD using EEG based features. This is a novel and promising finding that warrants further study.

9.3 Future work

Several components of our feature construction procedure have natural extensions that may improve the quality (e.g. the predictive performance) of the resulting features. In particular, the (within-subject) model we build of the longitudinal trajectory of each feature dimension is currently done independently of all other dimensions and all other subjects (see Section 6.2). Both assumptions are unrealistic, ignoring the significant similarity of trajectories between neighbouring voxels as well as the similarity of the trajectory of a particular voxel across subjects. We noted in the discussion of Chapter 7 that multi-task learning is a particularly promising means of modelling inter-subject correlations in each feature dimension's trajectory. For example, a Bayesian linear regression can be used to simultaneously estimate a dimension's trajectory coefficients across all subjects using a covariance structure that is parameterized to allow for inter-subject coupling. The optimal values of these (hyper) parameters can be found by maximizing the marginal likelihood of the available training data, avoiding the need for cross-validation.

Such models could be used to address the one of the primary problems encountered in this thesis: that of missing longitudinal follow-up information that severely limited the sample sizes used to estimate our longitudinal subspace. In particular, in Section 5.2.1 we that of the 252 total subjects in the HNRS dataset, only 54 subjects had both baseline and follow-up images (of which 47 were either MCI or healthy and thus included in our discrimination). Thus we excluded many of the 324 images from baseline or follow-up time-point that met our quality control

procedure when forming the longitudinal subspaces in the discriminations of Table 7.1 and 7.2. An inter-subject coupled model could allow for the estimation of trajectory coefficients for subjects having only one time-point's information, by estimating a common trajectory for each voxel using all subjects' information. The trajectory coefficients for subjects with multiple time-points would be estimated via a (parameterized) mixture of common and subject-specific information. In this way we could make use of cross-sectional information to better estimate longitudinal trajectories when faced with incomplete longitudinal data.

Another seemingly natural extension would be to use a more sophisticated dimensionality reduction method than PCA. Kernel PCA (Scholkopf et al., 1999) can be used to find a non-linear latent space that better captures most of the variability in the coefficient estimates across subjects. The potential disadvantage of kernel PCA compared to linear PCA is that non-linear kernel functions often introduce some additional parameters that are usually tuned via a costly cross-validation procedure. Additionally, the interpretation of the resulting features may no longer be straightforward if the mapping into the implicit feature space that the kernel function induces cannot be inverted, as is the case with the popular Gaussian kernel. Independent component analysis (ICA) is another well-known method that finds a set of statistically independent directions in feature space (Comon, 1994). It has been shown to be superior to PCA at determining the spatial and temporal extent of task-related activation in functional MRI (fMRI) studies (McKeown et al., 1998). Vicente et al., (2007) point out, however, that ICA does not do any better or worse than PCA when used with a rotationally invariant classifier, e.g. with a linear kernel SVM (Haasdonk and Burkhardt, 2007). With this in mind, the idea of replacing principal components, which are relatively easy to interpret and compute, needs to be carefully considered.

A clearer direction of future work involves improving our method's ability to transfer longitudinal information from one set of subjects to another (Sections 6.9 and 7.3.2). One approach would be to form the longitudinal projection using a much larger longitudinal training set, using the ADNI dataset for instance. Further to this, demographic confounds can also be removed from this projection by regressing out these variables at the coefficient matrix level (i.e. modelling the coefficient matrix as a linear combination of demographic factors and then removing these factors from

the coefficient matrix). If enough data are available, it may be preferable to form such a projection using only cognitively normal individuals, to minimize the heterogeneity in trajectories between diseased and healthy individuals. Exploring the effect of disease severity on anatomical trajectories is, however, an important topic of research (Ziegler et al., 2015b).

As discussed in Chapter 7, there is also the possibility of forming projections using higher order coefficients. With enough data or better models (e.g. those that capitalize on between-subject similarities), we can, for example, use second order trajectory models and create a projection using the second order coefficient matrix. Rather than (or in addition to) using eigenslopes, such eigencurvature based projections could be important as the hippocampus, for example, has been shown to exhibit non-linear trajectories in those over 50 years old (Fjell et al., 2013).

As a final point, best practices in the application of pattern recognition and machine learning methods to the analysis of neuroimaging data are still being established (see Section 2.6). In addition to ongoing work on classifier interpretation (e.g. t-stat, weight and forward maps), methods of assessing classifier generalizability continue to evolve. For example, we used leave one out cross-validation (LOO-CV) throughout this thesis, as it is simple and widely used in neuroimaging, where sample sizes have historically been small. Recent research, discussed in Section 4.4.3, showed that repeated random splitting may better estimate generalizability in some cases.

To conclude this thesis, we would like to point out that we have presented one simple approach to better using longitudinal information, which is itself just one particular type of problem structure. It is our hope that, if nothing else, this thesis inspires others to pursue methods that find simple and effective means of exploiting problem structure.

Appendices

Appendix A. Computation of principal components

The computation of the principal components we describe here is based on Section 12.1.4 of Bishop, (2007). It is well known that finding the principal components of a matrix is equivalent to finding the eigenvectors of its covariance matrix. We define the necessary $D \times D$ dimensional covariance matrix $\mathbf{S}_D = \frac{1}{n} \hat{\mathbf{X}}^T \hat{\mathbf{X}}$, where $\hat{\mathbf{X}}$ is the (row) mean-centred version of $\mathbf{X}$, the $n \times D$ matrix whose principal components we seek. The equation describing the i^{th} eigenvalue and eigenvector of $\mathbf{S}_D$ is:

$$\mathbf{S}_D \mathbf{u}_i = \lambda_i \mathbf{u}_i . \quad (\text{A.1})$$

where $\mathbf{u}_i$ is the i^{th} eigenvector of $\mathbf{S}_D$ and λ_i is the corresponding eigenvalue. Computing all of the D eigenvectors of the matrix $\mathbf{S}_D$ is infeasible as it is of size $D \times D$. However, as the original $\mathbf{X}$ is an $n \times D$ matrix, the rank of $\mathbf{S}_D$ will be at most $n - 1$. We therefore follow the development in Section 12.1.4 of Bishop, (2007) to transform the eigenvalues and eigenvectors of the $n \times n$ matrix $\mathbf{S}_n = \frac{1}{n} \hat{\mathbf{X}} \hat{\mathbf{X}}^T$ into those of $\mathbf{S}_D$. The final form of the i^{th} eigenvector, appropriately unit normalized, becomes

$$\mathbf{u}_i = \frac{1}{(n\lambda_i)^{1/2}} \hat{\mathbf{X}}^T \mathbf{v}_i \quad (\text{A.2})$$

where $\mathbf{v}_i$ is i^{th} eigenvector of $\mathbf{S}_n$ and λ_i is the corresponding eigenvalue.

The whole PCA procedure therefore involves computing the $n - 1$ non-zero eigenvalues and eigenvectors of $\mathbf{S}_n$ via a standard eigenvalue solver and then applying equation (A.2) k times to find the eigenvalues and eigenvectors of the desired $\mathbf{S}_D$, where $k < n$. In an array language such as MATLAB we can use the following equation to speed up computation:

$$\mathbf{U}_k = \hat{\mathbf{X}}^T \mathbf{V}_k \mathbf{D}_k \quad (\text{A.3})$$

where we have collected the k largest eigenvectors in the $n \times k$ matrix $\mathbf{V}_k = [\mathbf{v}_1, \dots, \mathbf{v}_k]$ and we form the $k \times k$ diagonal matrix $\mathbf{D}_k$ with elements $\frac{1}{(n\lambda_1)^{1/2}}, \dots, \frac{1}{(n\lambda_k)^{1/2}}$ along the leading diagonal. The $D \times k$ matrix $\mathbf{U}_k$ then contains the desired k principal components of the matrix $\mathbf{X}$ as columns.

Appendix B. Effect of SVM C-parameter variation

In the classifications we present in Chapter 5 (Tables 5.4 and 5.5), Chapter 7 (Tables 7.1-7.5) and Chapter 8 (Tables 8.2 and 8.3) we fixed the value of support vector classifier's C parameter at $C = 1$ throughout. This choice was made based on our experience using SVC with very high-dimensional feature vectors derived from MRI. In Figure A.1 we tested this choice by assessing classifier performance at values of C between 10^{-5} and 10^5 with logarithmic spacing. We considered the classification problem presented in Table 7.1: the Balanced Within-Set Prediction problem of discriminating MCI subjects from healthy controls using the HNRS dataset, with two time-points per subject.

Figure A.1 depicts the classifier balanced accuracies at these values of C for the three types of features we compared: follow-up time-point information, longitudinal (Ashburner and Ridgway, 2013b), and proposed LM-PCA projected follow-up time-point information. For all three types of features we observe stability in classifier performance between $C = 0.1$ and $C = 10^5$, with the cross-sectional follow-up and longitudinal features' performance remaining unchanged and the projected follow-up features' varying between 70% and 75%. The chosen value of $C = 1$, depicted by the dashed line, falls within this stable region. We observed similar stability of classifier performance at $C = 1$ in the other classifications we performed.

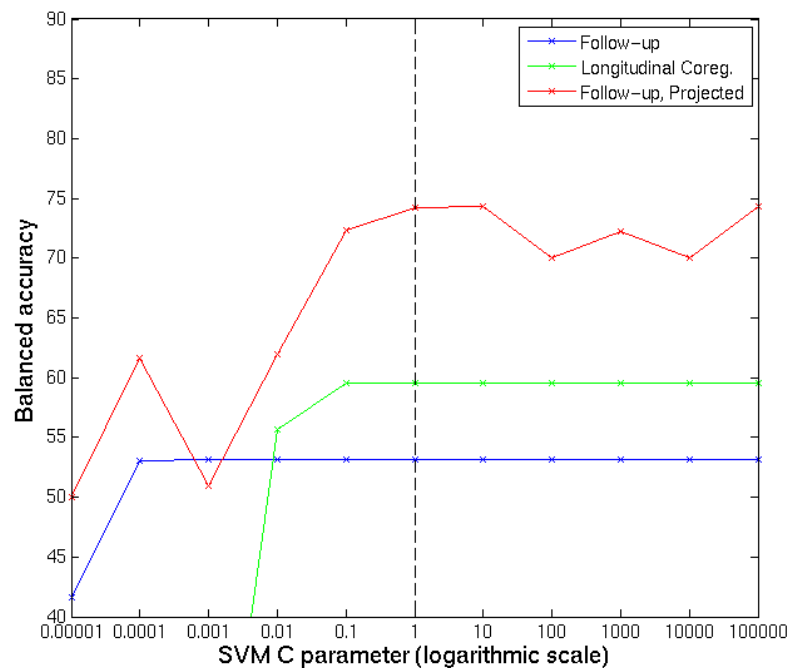


Figure 0-1 Effect of varying SVM C parameter on balanced accuracy of classifying MCI subjects versus Healthy Controls (HC) in HNRS dataset (*Balanced Within-Set Prediction* problem considered in Table 7.1). The $C = 1$ point shown as dashed line corresponds to the fixed value used in all results presented.

Appendix C. Extended dataset of the PMPP study

Imaging: 123I-FP-CIT-SPECT (single positron emission computed tomography), Transcranial Sonography (TCS), Magnetic Resonance Imaging (MRI): detailed in table below

Electrophysiology: Electroencephalography (EEG), Ambulatory polysomnography

Motor assessments: 3D movement analysis (Vicon 612 motion capture system), Quantitative ambulatory motor assessment (DynaPort Hybrid, McRoberts), Hoehn and Yahr Stage

Clinical and demographic data: Sensory function, Family history of PD, Edinburgh Handedness Inventory

Autonomic function: Unified Multiple System Atrophy Rating Scale (UMSARS) Part I (items 9–12), UMSARS Part III

Depression: Lifetime prevalence, Major depression

Neuropsychological test battery: Leistungsprüfsystem (translation from German: Performance test system) (LPS) 50+ (short version), Tower of London, Trail-Making Test A and B, Farb-Wort-Interferenz Test (i.e. Stroop Test), California Verbal Learning Test, Wechsler Memory Scale – Revised (WMS-R) (Logical Memory

part), WMS-R (Digit Span forward & backward part), Wechsler Adult Intelligence Scale (HAWIE) (Block Design part), Testbatterie zur Aufmerksamkeitsprüfung (translation from German: Test battery for attention testing) (TAP) (Alertness), TAP (Divided Attention)

Table A-1 Acquisition details for available MRI sequences, PMPP dataset

Sequence Name	TR (ms)	TE (ms)	TI (ms)	Flip Angle	Matrix	FOV (cm)	Slice Thickness (mm)	# Slices
T1-weighted (MPRAGE)	2300	3.03	1100	8°	256 × 224	25.6 × 22.4	1 mm	176
Gradient Echo	2000	TE1 = 3.37 ΔTE = 5.09 10 echoes		60°	192 × 192	22.9 × 22.9	3 mm	30
Turbo Spin Echo (TSE)	4830	15, 87, 160, 3 echoes			256 × 251	23.0 × 23.0	2 mm, 0.5 mm gap	10
Diffusion Tensor Imaging (DTI)	7500	79		90°	128 × 116	25.6 × 25.6	2 mm	61
3 b = 0 s mm ⁻² , 18 directions with b = 800 s mm ⁻²								
Perfusion	5000	14		90°	64 × 64	23.0 × 23.0	5 mm, 2 mm gap	7
Magnetic Resonance Spectra (MRS)		30					7	

Appendix D. Quality of FA maps registration and thresholding

To quantify the quality of the registration of FA maps in Chapter 8 (See Section 8.2.4) we computed the ‘mean overlap’ (i.e. Dice coefficient) between each registered FA image (the sources, 63 images total) and the white matter portion of the DARTEL based template (the target, a single image) that was created from the corresponding T1 weighted images. The mean overlap (MO) is defined as the size (in number of voxels here) of the intersection between two regions (a source region S and target region T) divided by the mean of the sizes of the target and source

regions, i.e. $MO = 2 \frac{|S \cap T|}{|S| + |T|}$. This is a simple and intuitive metric that has previously been used to compare registration accuracy (Klein et al., 2009).

We computed the MO between each FA map, thresholded at a minimum value of 0.2, and the white matter tissue class image of the template, thresholded at 0.01, as this value resulted in a comparable number of voxels in the target (278,897) and across the source images (median 303,255). Figure A.2 shows the MO across the three diagnostic groups as well as the MO within each group across the available time-points (0 – baseline, 1 – one year follow-up, 2 – two year follow-up). Using one-way ANOVAs, we tested for group differences in MO across and within groups, finding no difference across groups ($F_{2,60} = 0.64$, $p = 0.64$) and no differences within any group across time (PD: $F_{2,15} = 2.34$, $p = 0.13$, HRPD: $F_{2,32} = 0.34$, $p = 0.72$, Healthy: $F_{1,8} = 1.54$, $p = 0.25$).

To qualitatively assess the registration we inspected the FA maps with the highest, median and lowest MO value. Figure A.3 displays these, showing generally good alignment with the template.

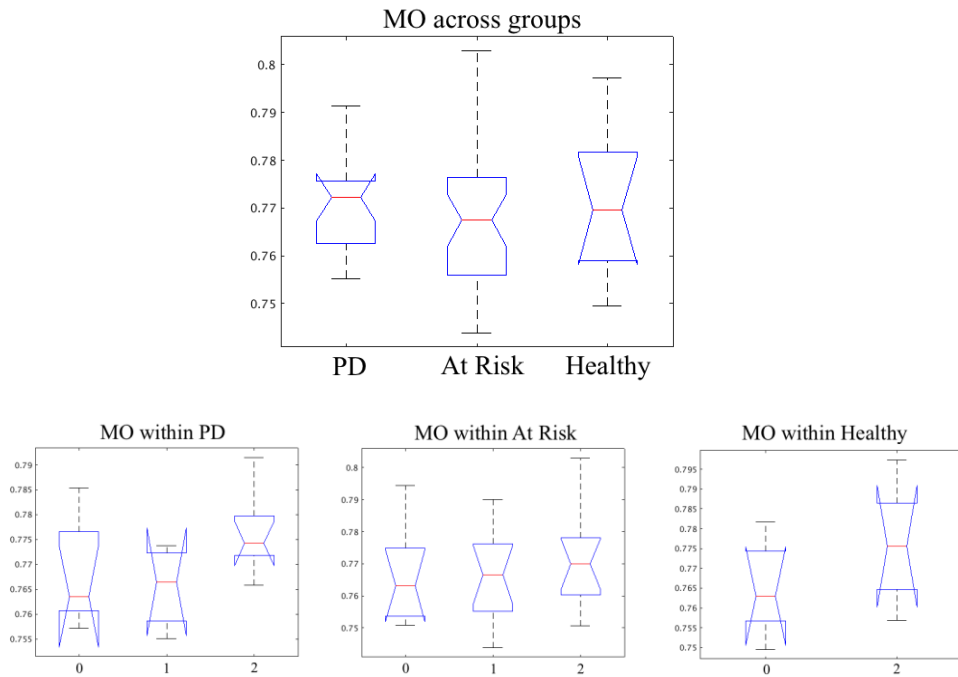


Figure 0-2 Mean overlaps (MO) between each source (FA map) and target (WM template), across the diagnostic groups and across time-points (0, 1, 2) within each group.

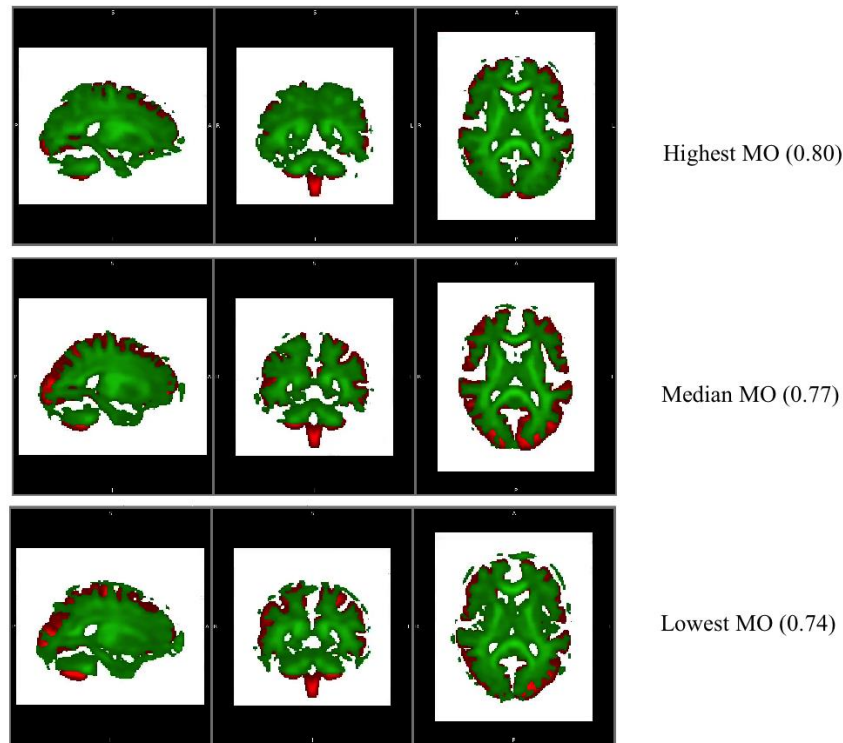


Figure 0-3 Visualization of alignment between the three FA maps (shown in green, top image) with highest, median and lowest mean overlap (MO) with template (shown in red, bottom image).

We performed a similar analysis to compare the number of discarded features across groups due to the minimum FA thresholds we used (0.2 and 0.3). Figure A.4 displays boxplots of these differences, which were not significant for either the 0.2 ($F_{2,41} = 2.16$, $p = 0.13$) or 0.3 threshold ($F_{2,41} = 2.4$, $p = 0.10$). Similarly, there were no significant differences (not displayed) across time within any of the groups for both the 0.2 (PD: $F_{2,9} = 0.01$, $p = 0.98$, HR_{PD}: $F_{2,21} = 1.32$, $p = 0.29$, Healthy: $F_{1,6} = 0.45$, $p = 0.53$) and 0.3 threshold (PD: $F_{2,9} = 0.02$, $p = 0.98$, HR_{PD}: $F_{2,21} = 0.65$, $p = 0.53$, Healthy: $F_{1,6} = 0.05$, $p = 0.83$).

To qualitatively assess the effect of the thresholding across groups, we also visualized the median image of each group after separately thresholding within each group. Figure A.5 shows the overlay of these images for both thresholds. In both cases the healthy control group's image contains slightly more features, while the at risk and PD groups are very similar.

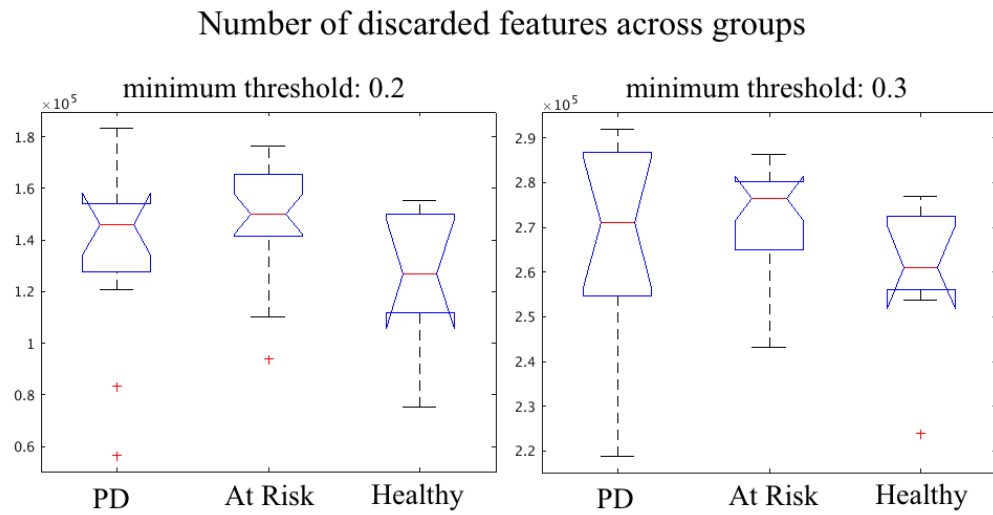
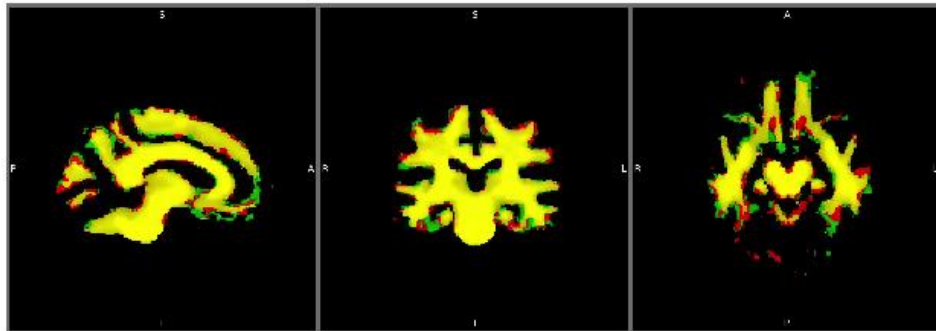
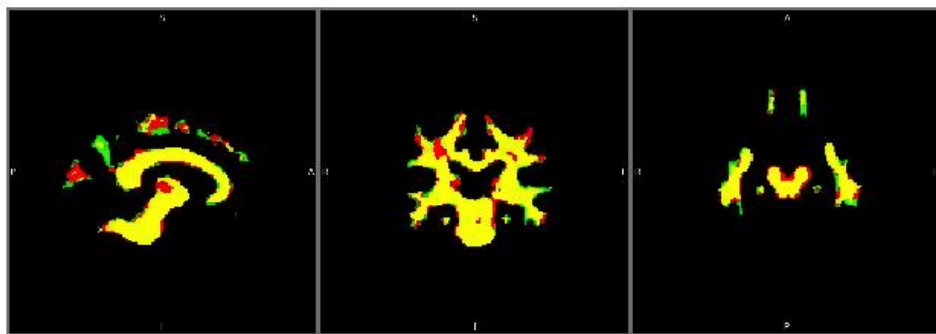


Figure 0-4 Number of discarded features across the diagnostic groups for the 0.2 and 0.3 minimum FA thresholds used in Chapter 8 analysis.

Median image for each group after thresholding



Minimum threshold: 0.2



Minimum threshold: 0.3

Figure 0-5 Overlays of median image of each group (green, bottom image: healthy control group, yellow, middle image: HR_{PD} group, red, top image: PD group). Each group has been separately thresholded at 0.2 (top) or 0.3 (bottom) minimum FA value.

Appendix E. OASIS dataset demographic and clinical information

There is a significant group difference in gender (Fisher's exact test: $p = 0.007$) but no other significant demographic differences (age: $F_{1,47} = 0.2$, $p = 0.65$, years between scans: $F_{1,47} = 0.7$, $p = 0.40$, education years: $F_{1,47} = 3.2$, $p = 0.08$). There is a significant difference in MMSE ($F_{1,47} = 21.9$, $p < 0.001$) but no significant difference in its change over time ($F_{1,47} = 1.4$, $p = 0.24$).

Table A-2 Demographic and clinical information for MCI versus controls discrimination, OASIS dataset

OASIS MRI (Follow-up + long. change)	Healthy (n = 25)	MCI (n = 24)
Male	4 (16.0%)	13 (54.2%)
Female	21 (84.0%)	11 (45.8%)
Age	77.6 ± 9.3	78.7 ± 7.3
Years Between Scans	2.0 ± 0.4	1.9 ± 0.6
Education Years	14.6 ± 2.7	13.1 ± 3.0
MMSE	28.7 ± 1.3	25.1 ± 3.6
MMSE change	-0.5 ± 1.4	-1.1 ± 2.3

Appendix F. Method implementation

An implementation of our method, along with demo data and demo analysis, is available at <https://github.com/LeonAksman/lpr>.

References

- Adeli, E., Shi, F., An, L., Wee, C.-Y., Wu, G., Wang, T., Shen, D., 2016. Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. *NeuroImage* 141, 206–219. doi:10.1016/j.neuroimage.2016.05.054
- Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* 7, 270–279. doi:10.1016/j.jalz.2011.03.008
- Anderson, V.M., Schott, J.M., Bartlett, J.W., Leung, K.K., Miller, D.H., Fox, N.C., 2012. Gray matter atrophy rate as a marker of disease progression in AD. *Neurobiol. Aging* 33, 1194–1202. doi:10.1016/j.neurobiolaging.2010.11.001
- Andersson, J.L.R., Jenkinson, M., Smith, S., 2007b. Non-linear registration aka Spatial normalisation. *FMRIB Tech. Rep. TR07JA2* Wwwfmribboxacukanalysistechrep.
- Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2016. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. doi:10.1016/j.neuroimage.2016.02.079
- Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. *NeuroImage* 38, 95–113. doi:10.1016/j.neuroimage.2007.07.007
- Ashburner, J., Friston, K.J., 2001. Why Voxel-Based Morphometry Should Be Used. *NeuroImage* 14, 1238–1243. doi:10.1006/nimg.2001.0961
- Ashburner, J., Friston, K.J., 2000. Voxel-Based Morphometry—The Methods. *NeuroImage* 11, 805–821. doi:10.1006/nimg.2000.0582
- Ashburner, J., Ridgway, G.R., 2013a. Symmetric diffeomorphic modeling of longitudinal structural MRI. *Brain Imaging Methods* 6, 197. doi:10.3389/fnins.2012.00197
- Assaf, Y., Pasternak, O., 2007. Diffusion Tensor Imaging (DTI)-based White Matter Mapping in Brain Research: A Review. *J. Mol. Neurosci.* 34, 51–61. doi:10.1007/s12031-007-0029-0
- Bach, F.R., Lanckriet, G.R.G., Jordan, M.I., 2004. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm, in: *Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04*. ACM, New York, NY, USA, p. 6–. doi:10.1145/1015330.1015424
- Barnes, J., Scahill, R.I., Frost, C., Schott, J.M., Rossor, M.N., Fox, N.C., 2008. Increased hippocampal atrophy rates in AD over 6 months using serial MR imaging. *Neurobiol. Aging* 29, 1199–1203. doi:10.1016/j.neurobiolaging.2007.02.011
- Bartlett, P., Shawe-taylor, J., 1998. Generalization Performance of Support Vector Machines and Other Pattern Classifiers.
- Basser, P.J., Pierpaoli, C., 1996. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson. B* 111, 209–219.
- Benamer, T.S., Patterson, J., Grosset, D.G., Booij, J., de Bruin, K., van Royen, E., Speelman, J.D., Horstink, M.H., Sips, H.J., Dierckx, R.A., Versijpt, J., Decoo, D., Van Der Linden, C., Hadley, D.M., Doder, M., Lees, A.J., Costa, D.C., Gacinovic, S., Oertel, W.H., Pogarell, O., Hoeffken, H., Joseph, K., Tatsch, K., Schwarz, J., Ries, V., 2000. Accurate differentiation of parkinsonism and essential tremor using visual assessment of [123I]-FP-CIT SPECT imaging: the [123I]-FP-CIT study group. *Mov. Disord. Off. J. Mov. Disord. Soc.* 15, 503–510.
- Berg, D., Roggendorf, W., Schröder, U., Klein, R., Tatschner, T., Benz, P., Tucha, O., Preier, M., Lange, K.W., Reiners, K., Gerlach, M., Becker, G., 2002. Echogenicity of the substantia nigra: association with increased iron content and marker for susceptibility to nigrostriatal injury. *Arch. Neurol.* 59, 999–1005.

- Berg, D., Seppi, K., Behnke, S., Liepelt, I., Schweitzer, K., Stockner, H., Wollenweber, F., Gaenslen, A., Mahlknecht, P., Spiegel, J., Godau, J., Huber, H., Srulijes, K., Kiechl, S., Bentele, M., Gasperi, A., Schubert, T., Hiry, T., Probst, M., Schneider, V., Klenk, J., Sawires, M., Willeit, J., Maetzler, W., Fassbender, K., Gasser, T., Poewe, W., 2011. Enlarged Substantia Nigra Hyperechogenicity and Risk for Parkinson Disease: A 37-Month 3-Center Study of 1847 Older Persons. *Arch. Neurol.* 68, 932–937. doi:10.1001/archneurol.2011.141
- Berg, D., Siefker, C., Becker, G., 2001. Echogenicity of the substantia nigra in Parkinson's disease and its relation to clinical findings. *J. Neurol.* 248, 684–689.
- Bishop, C.M., 2007. *Pattern Recognition and Machine Learning*. Springer, New York.
- Blass, J.P., 2002. Alzheimer's disease and Alzheimer's dementia: distinct but overlapping entities. *Neurobiol. Aging* 23, 1077–1084. doi:10.1016/S0197-4580(02)00036-2
- Braak, H., Tredici, K.D., Rüb, U., de Vos, R.A.I., Jansen Steur, E.N.H., Braak, E., 2003. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging* 24, 197–211. doi:10.1016/S0197-4580(02)00065-9
- Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The Balanced Accuracy and Its Posterior Distribution, in: 2010 20th International Conference on Pattern Recognition (ICPR). Presented at the 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3121–3124. doi:10.1109/ICPR.2010.764
- Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M.E., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A.M., Dolph, C.V., Iftekharruddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage* 111, 562–579. doi:10.1016/j.neuroimage.2015.01.048
- Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* 3, 186–191. doi:10.1016/j.jalz.2007.04.381
- Brun, A., Englund, E., 1986. A white matter disorder in dementia of the Alzheimer type: A pathoanatomical study. *Ann. Neurol.* 19, 253–262. doi:10.1002/ana.410190306
- Buckner, R.L., Snyder, A.Z., Shannon, B.J., LaRossa, G., Sachs, R., Fotenos, A.F., Sheline, Y.I., Klunk, W.E., Mathis, C.A., Morris, J.C., Mintun, M.A., 2005. Molecular, Structural, and Functional Characterization of Alzheimer's Disease: Evidence for a Relationship between Default Activity, Amyloid, and Memory. *J. Neurosci.* 25, 7709–7717. doi:10.1523/JNEUROSCI.2177-05.2005
- Busse, A., Bischof, J., Riedel-Heller, S.G., Angermeyer, M.C., 2003. Subclassifications for mild cognitive impairment: prevalence and predictive validity. *Psychol. Med.* 33, 1029–1038.
- Cattell, L., Platsch, G., Pfeiffer, R., Declerck, J., Schnabel, J.A., Hutton, C., 2016. Classification of amyloid status using machine learning with histograms of oriented 3D gradients. *NeuroImage Clin.* doi:10.1016/j.nicl.2016.05.004
- Cavedo, E., Lista, S., Khachaturian, Z.S., Aisen, P.S., Amouyel, P., Herholz, K., Jack, C., Sperling, R., Cummings, J., Blennow, K., O'Bryant, S., Frisoni, G., Khachaturian, A., Kivipelto, M., Klunk, W., Broich, K., Andrieu, S., Schotten, M.T. de, Mangin, J.-F., Lammertsma, A., Johnson, K., Teipel, S.J., Drzezga, A., Bokde, A.L.W., Colliot, O., Bakardjian, H., Zetterberg, H., Dubois, B., Vellas, B., Schneider, L., Hampel, H., 2014. The road ahead to cure Alzheimer's disease: development of biological markers and neuroimaging methods for prevention trials across all stages and target populations. *J. Prev. Alzheimer's Dis.* 1, 181–202.

- Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., Cercignani, M., 2015. Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *NeuroImage* 112, 232–243. doi:10.1016/j.neuroimage.2015.02.037
- Chang, C.-C., Lin, C.-J., 2011a. LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol* 2, 27:1–27:27. doi:10.1145/1961189.1961199
- Chang, Y.-L., Bondi, M.W., McEvoy, L.K., Fennema-Notestine, C., Salmon, D.P., Galasko, D., Hagler, D.J., Dale, A.M., 2011. Global clinical dementia rating of 0.5 in MCI masks variability related to level of function. *Neurology* 76, 652–659. doi:10.1212/WNL.0b013e31820ce6a5
- Chaudhuri, K.R., Bhidayasiri, R., van Laar, T., 2016. Unmet needs in Parkinson's disease: New horizons in a changing landscape. *Parkinsonism Relat. Disord.*, More than a Century of Apomorphine: The New Faces of an Old Therapy in Parkinson's Disease 33, Supplement 1, S2–S8. doi:10.1016/j.parkreldis.2016.11.018
- Chen, L., Hagenah, J., Mertins, A., 2012. Feature Analysis for Parkinson's Disease Detection Based on Transcranial Sonography Image, in: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 272–279. doi:10.1007/978-3-642-33454-2_34
- Chen, S., DuBois Bowman, F., 2011a. A novel support vector classifier for longitudinal high-dimensional data and its application to neuroimaging data. *Stat. Anal. Data Min.* 4, 604–611. doi:10.1002/sam.10141
- Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A., Nobili, F., 2011. Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. *NeuroImage* 58, 469–480. doi:10.1016/j.neuroimage.2011.05.083
- Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., 2012. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage* 60, 59–70. doi:10.1016/j.neuroimage.2011.11.066
- Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehericy, S., Benali, H., Garnero, L., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus* 19, 579–587. doi:10.1002/hipo.20626
- Comon, P., 1994. Independent Component Analysis, a New Concept? *Signal Process* 36, 287–314. doi:10.1016/0165-1684(94)90029-9
- Costafreda, S.G., Dinov, I.D., Tu, Z., Shi, Y., Liu, C.-Y., Kloszewska, I., Mecocci, P., Soininen, H., Tsolaki, M., Vellas, B., Wahlund, L.-O., Spenger, C., Toga, A.W., Lovestone, S., Simmons, A., 2011. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *NeuroImage* 56, 212–219. doi:10.1016/j.neuroimage.2011.01.050
- Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19, 261–270. doi:10.1016/S1053-8119(03)00049-1
- Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage* 56, 766–781. doi:10.1016/j.neuroimage.2010.06.013
- Darcourt, J., Booi, J., Tatsch, K., Varrone, A., Borgh, T.V., Kapucu, Ö.L., Nägren, K., Nobili, F., Walker, Z., Laere, K.V., 2009. EANM procedure guidelines for brain neurotransmission SPECT using 123I-labelled dopamine transporter ligands, version 2. *Eur. J. Nucl. Med. Mol. Imaging* 37, 443–450. doi:10.1007/s00259-009-1267-x

- Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol. Aging* 32, 2322.e19-2322.e27. doi:10.1016/j.neurobiolaging.2010.05.023
- Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M., 2008. Detection of Prodromal Alzheimer's Disease via Pattern Classification of MRI. *Neurobiol. Aging* 29, 514–523. doi:10.1016/j.neurobiolaging.2006.11.010
- Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage* 41, 1220–1227. doi:10.1016/j.neuroimage.2008.03.050
- de Lau, L.M., Breteler, M.M., 2006. Epidemiology of Parkinson's disease. *Lancet Neurol.* 5, 525–535. doi:10.1016/S1474-4422(06)70471-9
- de Weerd, A.W., Perquin, W.V.M., Jonkman, E.J., 1990. Role of the EEG in the Prediction of Dementia in Parkinson's Disease. *Dement. Geriatr. Cogn. Disord.* 1, 115–118. doi:10.1159/000107129
- deFigueiredo, R.J., Shankle, W.R., Maccato, A., Dick, M.B., Mundkur, P., Mena, I., Cotman, C.W., 1995. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. *Proc. Natl. Acad. Sci. U. S. A.* 92, 5530–5534.
- Desikan, R., Cabral, H., Hess, C., Dillon, W., Glastonbury, C., Weiner, M., Schmansky, N., Greve, D., Salat, D., Buckner, Fischl, B., Initiative, A.D.N., 2009. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain J. Neurol.* 132. doi:10.1093/brain/awp123
- Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980. doi:10.1016/j.neuroimage.2006.01.021
- Dlugaj, M., Weimar, C., Gerwig, M., Wege, N., Verde, P.E., Dragano, N., Moebus, S., Jöckel, K.-H., Erbel, R., Siegrist, J., 2010. Prevalence of mild cognitive impairment and its subtypes in the Heinz Nixdorf Recall study cohort. *Alzheimers Dement. J. Alzheimers Assoc.* 6, S474. doi:10.1016/j.jalz.2010.05.1578
- Doyle, O.M., Westman, E., Marquand, A.F., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Williams, S.C.R., Simmons, A., 2014. Predicting Progression of Alzheimer's Disease Using Ordinal Regression. *PLoS ONE* 9, e105542. doi:10.1371/journal.pone.0105542
- Duchesnay, E., Cachia, A., Roche, A., Riviere, D., Cointepas, Y., Papadopoulos-Orfanos, D., Zilbovicius, M., Martinot, J.L., Regis, J., Mangin, J.F., 2007. Classification Based on Cortical Folding Patterns. *IEEE Trans. Med. Imaging* 26, 553–565. doi:10.1109/TMI.2007.892501
- Duchesne, S., Bocti, C., De Sousa, K., Frisoni, G.B., Chertkow, H., Collins, D.L., 2010. Amnestic MCI future clinical status prediction using baseline MRI features. *Neurobiol. Aging* 31, 1606–1617. doi:10.1016/j.neurobiolaging.2008.09.003
- Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L., 2008. MRI-Based Automated Computer Classification of Probable AD Versus Normal Controls. *IEEE Trans. Med. Imaging* 27, 509–520. doi:10.1109/TMI.2007.908685
- Duchesne, S., Rolland, Y., Vérin, M., 2009. Automated computer differential classification in Parkinsonian Syndromes via pattern analysis on MRI. *Acad. Radiol.* 16, 61–70. doi:10.1016/j.acra.2008.05.024
- Dyrba, M., Grothe, M., Kirste, T., Teipel, S.J., 2015. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* 36, 2118–2131. doi:10.1002/hbm.22759
- Eickhoff, S., Nichols, T.E., Van Horn, J.D., Turner, J.A., 2016. Sharing the wealth: Neuroimaging data repositories. *NeuroImage, Sharing the wealth: Brain Imaging*

- Repositories in 2015 124, Part B, 1065–1068. doi:10.1016/j.neuroimage.2015.10.079
- Fahn, S., Oakes, D., Shoulson, I., Kieburtz, K., Rudolph, A., Lang, A., Olanow, C.W., Tanner, C., Marek, K., Parkinson Study Group, 2004. Levodopa and the progression of Parkinson's disease. *N. Engl. J. Med.* 351, 2498–2508. doi:10.1056/NEJMoA033447
- Fan, Y., Rao, H., Giannetta, J., Hurt, H., Wang, J., Davatzikos, C., Shen, D., 2006. Diagnosis of brain abnormality using both structural and functional MR images. *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. Suppl.* 6585–6588. doi:10.1109/IEMBS.2006.260894
- Fan, Y., Rao, H., Hurt, H., Giannetta, J., Korczykowski, M., Shera, D., Avants, B.B., Gee, J.C., Wang, J., Shen, D., 2007. Multivariate examination of brain abnormality using both structural and functional MRI. *NeuroImage* 36, 1189–1199. doi:10.1016/j.neuroimage.2007.04.009
- Fearnley, J.M., Lees, A.J., 1991. Ageing and Parkinson's Disease: Substantia Nigra Regional Selectivity. *Brain* 114, 2283–2301. doi:10.1093/brain/114.5.2283
- Fernández-Seara, M.A., Mengual, E., Vidorreta, M., Aznárez-Sanado, M., Loayza, F.R., Villagra, F., Irigoyen, J., Pastor, M.A., 2012. Cortical hypoperfusion in Parkinson's disease assessed using arterial spin labeled perfusion MRI. *NeuroImage* 59, 2743–2750. doi:10.1016/j.neuroimage.2011.10.033
- Ferrarini, L., Frisoni, G.B., Pievani, M., Reiber, J.H.C., Ganzola, R., Milles, J., 2009. Morphological hippocampal markers for automated detection of Alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. *J. Alzheimers Dis. JAD* 17, 643–659. doi:10.3233/JAD-2009-1082
- Figueroa, J.J., Singer, W., Parsaik, A., Benarroch, E.E., Ahlskog, J.E., Fealey, R.D., Parisi, J.E., Sandroni, P., Mandrekar, J., Iodice, V., Low, P.A., Bower, J.H., 2014. Multiple System Atrophy: Prognostic Indicators of Survival. *Mov. Disord. Off. J. Mov. Disord. Soc.* 29, 1151–1157. doi:10.1002/mds.25927
- Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M., Initiative, for the A.D.N., 2010. CSF Biomarkers in Prediction of Cerebral and Clinical Change in Mild Cognitive Impairment and Alzheimer's Disease. *J. Neurosci.* 30, 2088–2101. doi:10.1523/JNEUROSCI.3785-09.2010
- Fjell, A.M., Westlye, L.T., Grydeland, H., Amlie, I., Espeseth, T., Reinvang, I., Raz, N., Dale, A.M., Walhovd, K.B., 2014. Accelerating Cortical Thinning: Unique to Dementia or Universal in Aging? *Cereb. Cortex N. Y. NY* 24, 919–934. doi:10.1093/cercor/bhs379
- Fjell, A.M., Westlye, L.T., Grydeland, H., Amlie, I., Espeseth, T., Reinvang, I., Raz, N., Holland, D., Dale, A.M., Walhovd, K.B., 2013. Critical ages in the life course of the adult brain: nonlinear subcortical aging. *Neurobiol. Aging* 34, 2239–2247. doi:10.1016/j.neurobiolaging.2013.04.006
- Focke, N.K., Helms, G., Scheewe, S., Pantel, P.M., Bachmann, C.G., Dechent, P., Ebentheuer, J., Mohr, A., Paulus, W., Trenkwalder, C., 2011. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic parkinson syndrome and healthy controls. *Hum. Brain Mapp.* 32, 1905–1915. doi:10.1002/hbm.21161
- Fox, N.C., Schott, J.M., 2004. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *The Lancet* 363, 392–394. doi:10.1016/S0140-6736(04)15441-X
- Freeborough, P.A., Fox, N.C., 1997a. The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Trans. Med. Imaging* 16, 623–629. doi:10.1109/42.640753
- Freeborough, P.A., Woods, R.P., Fox, N.C., 1996. Accurate registration of serial 3D MR brain images and its application to visualizing change in neurodegenerative disorders. *J. Comput. Assist. Tomogr.* 20, 1012–1022.

- Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi:10.1038/nrneurol.2009.215
- Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi:10.1002/hbm.460020402
- Gaenslen, A., Swid, I., Liepelt-Scarfone, I., Godau, J., Berg, D., 2011. The Patients' Perception of Prodromal Symptoms Before the Initial Diagnosis of Parkinson's Disease. *Mov. Disord.* 26, 653–658. doi:10.1002/mds.23499
- Gaenslen, A., Unmuth, B., Godau, J., Liepelt, I., Di Santo, A., Schweitzer, K.J., Gasser, T., Machulla, H.-J., Reimold, M., Marek, K., Berg, D., 2008. The specificity and sensitivity of transcranial ultrasound in the differential diagnosis of Parkinson's disease: a prospective blinded study. *Lancet Neurol.* 7, 417–424. doi:10.1016/S1474-4422(08)70067-X
- Gallichan, D., Scholz, J., Bartsch, A., Behrens, T.E., Robson, M.D., Miller, K.L., 2010. Addressing a systematic vibration artifact in diffusion-weighted MRI. *Hum. Brain Mapp.* 31, 193–202. doi:10.1002/hbm.20856
- Gelman, A., Stern, H., 2006. The difference between “significant” and “not significant” is not itself statistically significant.
- Gerardin, E., Chételat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.-S., Niethammer, M., Dubois, B., Lehericy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage* 47, 1476–1486. doi:10.1016/j.neuroimage.2009.05.036
- Gray, K.R., Wolz, R., Heckemann, R.A., Aljabar, P., Hammers, A., Rueckert, D., 2012. Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease. *NeuroImage* 60, 221–229. doi:10.1016/j.neuroimage.2011.12.071
- Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci.* 100, 253–258. doi:10.1073/pnas.0135058100
- Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V., 2004. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proc. Natl. Acad. Sci. U. S. A.* 101, 4637–4642. doi:10.1073/pnas.0308627101
- Haasdonk, B., Burkhardt, H., 2007. Invariant kernel functions for pattern analysis and machine learning. *Mach. Learn.* 68, 35–61. doi:10.1007/s10994-007-5009-7
- Haense, C., Herholz, K., Jagust, W.J., Heiss, W.D., 2009. Performance of FDG PET for detection of Alzheimer's disease in two independent multicentre samples (NEST-DD and ADNI). *Dement. Geriatr. Cogn. Disord.* 28, 259–266. doi:10.1159/000241879
- Hagenah, J.M., Hedrich, K., Becker, B., Pramstaller, P.P., Seidel, G., Klein, C., 2006. Distinguishing early-onset PD from dopa-responsive dystonia with transcranial sonography. *Neurology* 66, 1951–1952. doi:10.1212/01.wnl.0000219806.03849.1a
- Han, C.-X., Wang, J., Yi, G.-S., Che, Y.-Q., 2013. Investigation of EEG abnormalities in the early stage of Parkinson's disease. *Cogn. Neurodyn.* 7, 351–359. doi:10.1007/s11571-013-9247-z
- Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., Pollmann, S., 2009. PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics* 7, 37–53. doi:10.1007/s12021-008-9041-y
- Hastie, T., Tibshirani, R., Friedman, J., 2011a. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, 2nd ed. 2009. Corr. 7th printing 2013 edition. ed. Springer, New York, NY.
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014a. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87, 96–110. doi:10.1016/j.neuroimage.2013.10.067

- Hawkes, C.H., 2008. The prodromal phase of sporadic Parkinson's disease: Does it exist and if so how long is it? *Mov. Disord.* 23, 1799–1807. doi:10.1002/mds.22242
- Hawkins, D.M., 2004. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* 44, 1–12. doi:10.1021/ci0342472
- Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, 20 YEARS OF fMRI 20 YEARS OF fMRI 62, 852–855. doi:10.1016/j.neuroimage.2012.03.016
- Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* 293, 2425–2430. doi:10.1126/science.1063736
- Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* 8, 686–691. doi:10.1038/nn1445
- Herholz, K., Boecker, H., Nemeth, I., Dunn, G., 2013. FDG PET in dementia multicenter studies and clinical trials. *Clin. Transl. Imaging* 1, 261–270. doi:10.1007/s40336-013-0018-y
- Herholz, K., Salmon, E., Perani, D., Baron, J.-C., Holthoff, V., Frölich, L., Schönknecht, P., Ito, K., Mielke, R., Kalbe, E., Zündorf, G., Delbeuck, X., Pelati, O., Anchisi, D., Fazio, F., Kerrouche, N., Desgranges, B., Eustache, F., Beuthien-Baumann, B., Menzel, C., Schröder, J., Kato, T., Arahata, Y., Henze, M., Heiss, W.-D., 2002. Discrimination between Alzheimer Dementia and Controls by Automated Analysis of Multicenter FDG PET. *NeuroImage* 17, 302–316. doi:10.1006/nimg.2002.1208
- Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., Alzheimer's Disease Neuroimaging Initiative, 2009. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *NeuroImage* 48, 138–149. doi:10.1016/j.neuroimage.2009.05.056
- Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage* 55, 574–589. doi:10.1016/j.neuroimage.2010.10.081
- Holland, D., Dale, A.M., 2011. Nonlinear registration of longitudinal images and measurement of change in regions of interest. *Med. Image Anal.* 15, 489–497. doi:10.1016/j.media.2011.02.005
- Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24, 417–441. doi:10.1037/h0071325
- Hua, X., Lee, S., Hibar, D.P., Yanovsky, I., Leow, A.D., Toga, A.W., Jack Jr., C.R., Bernstein, M.A., Reiman, E.M., Harvey, D.J., Kornak, J., Schuff, N., Alexander, G.E., Weiner, M.W., Thompson, P.M., 2010. Mapping Alzheimer's disease progression in 1309 MRI scans: Power estimates for different inter-scan intervals. *NeuroImage* 51, 63–75. doi:10.1016/j.neuroimage.2010.01.104
- Hua, X., Lee, S., Yanovsky, I., Leow, A.D., Chou, Y.-Y., Ho, A.J., Gutman, B., Toga, A.W., Jack Jr, C.R., Bernstein, M.A., Reiman, E.M., Harvey, D.J., Kornak, J., Schuff, N., Alexander, G.E., Weiner, M.W., Thompson, P.M., 2009. Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with tensor-based morphometry: An ADNI study of 515 subjects. *NeuroImage* 48, 668–681. doi:10.1016/j.neuroimage.2009.07.011
- Hua, X., Leow, A.D., Lee, S., Klunder, A.D., Toga, A.W., Lepore, N., Chou, Y.-Y., Brun, C., Chiang, M.-C., Barysheva, M., Jack Jr., C.R., Bernstein, M.A., Britson, P.J., Ward, C.P., Whitwell, J.L., Borowski, B., Fleisher, A.S., Fox, N.C., Boyes, R.G., Barnes, J., Harvey, D., Kornak, J., Schuff, N., Boreta, L., Alexander, G.E., Weiner, M.W., Thompson, P.M., Initiative, the A.D.N., 2008. 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *NeuroImage* 41, 19–34. doi:10.1016/j.neuroimage.2008.02.010
- Hughes, A.J., Daniel, S.E., Lees, A.J., 1993. The clinical features of Parkinson's disease in 100 histologically proven cases. *Adv. Neurol.* 60, 595–599.

- Illan, I.A., Gorz, J.M., Ramirez, J., Segovia, F., Jimenez-Hoyuela, J.M., Ortega Lozano, S.J., 2012. Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging. *Med. Phys.* 39, 5971–5980. doi:10.1118/1.4742055
- Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* 7, 257–262. doi:10.1016/j.jalz.2011.03.004
- Jack, C.R., Barnes, J., Bernstein, M.A., Borowski, B.J., Brewer, J., Clegg, S., Dale, A.M., Carmichael, O., Ching, C., DeCarli, C., Desikan, R.S., Fennema-Notestine, C., Fjell, A.M., Fletcher, E., Fox, N.C., Gunter, J., Gutman, B.A., Holland, D., Hua, X., Insel, P., Kantarci, K., Killiany, R.J., Krueger, G., Leung, K.K., Mackin, S., Maillard, P., Malone, I.B., Mattsson, N., McEvoy, L., Modat, M., Mueller, S., Nosheny, R., Ourselin, S., Schuff, N., Senjem, M.L., Simonson, A., Thompson, P.M., Rettmann, D., Vemuri, P., Walhovd, K., Zhao, Y., Zuk, S., Weiner, M., 2015. Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. *Alzheimers Dement. J. Alzheimers Assoc.* 11, 740–756. doi:10.1016/j.jalz.2015.05.002
- Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 9, 119–128. doi:10.1016/S1474-4422(09)70299-6
- Jack, C.R., Shiung, M.M., Gunter, J.L., O'Brien, P.C., Weigand, S.D., Knopman, D.S., Boeve, B.F., Ivnik, R.J., Smith, G.E., Cha, R.H., Tangalos, E.G., Petersen, R.C., 2004. Comparison of Different MRI Brain Atrophy Rate Measures with Clinical Disease Progression in AD. *Neurology* 62, 591–600.
- Jockel, K.-H., 1986. Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests. *Ann. Stat.* 14, 336–347. doi:10.1214/aos/1176349860
- Joliffe, I.T., 2002. *Principal Component Analysis*, Springer Series in Statistics. Springer-Verlag, New York.
- Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi:10.1038/nn1444
- Kerrouche, N., Herholz, K., Mielke, R., Holthoff, V., Baron, J.-C., 2006. 18FDG PET in Vascular Dementia: Differentiation from Alzheimer's Disease Using Voxel-Based Multivariate Analysis. *J. Cereb. Blood Flow Metab.* 26, 1213–1221. doi:10.1038/sj.jcbfm.9600296
- Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., 2015. Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin. Neurophysiol.* 126, 2132–2141. doi:10.1016/j.clinph.2015.02.060
- Kippenhan, J.S., Barker, W.W., Pascal, S., Nagel, J., Duara, R., 1992. Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects. *J. Nucl. Med. Off. Publ. Soc. Nucl. Med.* 33, 1459–1467.
- Kippenhan, J.S., Nagel, J.H., 1990. Diagnosis And Modelling Of Alzheimer's Disease Through Neural Network Analyses Of Pet Studies, in: , *Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1990. Presented at the , *Proceedings of the Twelfth Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1990, pp. 1449–1450. doi:10.1109/IEMBS.1990.691833
- Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage* 46, 786–802. doi:10.1016/j.neuroimage.2008.12.037

- Klöppel, S., Abdulkadir, A., Jack Jr., C.R., Koutsouleris, N., Mourão-Miranda, J., Vemuri, P., 2012a. Diagnostic neuroimaging across diseases. *NeuroImage* 61, 457–463. doi:10.1016/j.neuroimage.2011.11.002
- Klöppel, S., Chu, C., Tan, G.C., Draganski, B., Johnson, H., Paulsen, J.S., Kienzl, W., Tabrizi, S.J., Ashburner, J., Frackowiak, R.S.J., PREDICT-HD Investigators of the Huntington Study Group, 2009. Automatic detection of preclinical neurodegeneration: presymptomatic Huntington disease. *Neurology* 72, 426–431. doi:10.1212/01.wnl.0000341768.28646.b6
- Klöppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008a. Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain* 131, 2969–2974. doi:10.1093/brain/awn239
- Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., 2008b. Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689. doi:10.1093/brain/awm319
- Klunk, W.E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D.P., Bergström, M., Savitcheva, I., Huang, G.-F., Estrada, S., Ausén, B., Debnath, M.L., Barletta, J., Price, J.C., Sandell, J., Lopresti, B.J., Wall, A., Koivisto, P., Antoni, G., Mathis, C.A., Långström, B., 2004. Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B. *Ann. Neurol.* 55, 306–319. doi:10.1002/ana.20009
- Knopman, D.S., DeKosky, S.T., Cummings, J.L., Chui, H., Corey-Bloom, J., Relkin, N., Small, G.W., Miller, B., Stevens, J.C., 2001. Practice parameter: Diagnosis of dementia (an evidence-based review) Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology* 56, 1143–1153. doi:10.1212/WNL.56.9.1143
- Koutsouleris, N., Meisenzahl, E.M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., Gaser, C., 2009. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch. Gen. Psychiatry* 66, 700–712. doi:10.1001/archgenpsychiatry.2009.62
- Kraha, A., Turner, H., Nimon, K., Zientek, L.R., Henson, R.K., 2012. Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity. *Front. Psychol.* 3. doi:10.3389/fpsyg.2012.00044
- Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J., 2005. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 957–968. doi:10.1109/TPAMI.2005.127
- Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I., 2004. Learning the Kernel Matrix with Semidefinite Programming. *J Mach Learn Res* 5, 27–72.
- Landau, S.M., Harvey, D., Madison, C.M., Reiman, E.M., Foster, N.L., Aisen, P.S., Petersen, R.C., Shaw, L.M., Trojanowski, J.Q., Jack, C.R., Weiner, M.W., Jagust, W.J., 2010. Comparing predictors of conversion and decline in mild cognitive impairment(Podcast)(e-Pub ahead of print). *Neurology* 75, 230–238. doi:10.1212/WNL.0b013e3181e8e8b8
- Lang, A.E., 2011. A critical appraisal of the premotor symptoms of Parkinson's disease: Potential usefulness in early diagnosis and design of neuroprotective trials. *Mov. Disord.* 26, 775–783. doi:10.1002/mds.23609
- Langbaum, J.B.S., Chen, K., Lee, W., Reschke, C., Bandy, D., Fleisher, A.S., Alexander, G.E., Foster, N.L., Weiner, M.W., Koeppe, R.A., Jagust, W.J., Reiman, E.M., 2009. CATEGORICAL AND CORRELATIONAL ANALYSES OF BASELINE FLUORODEOXYGLUCOSE POSITRON EMISSION TOMOGRAPHY IMAGES FROM THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (ADNI). *NeuroImage* 45, 1107–1116. doi:10.1016/j.neuroimage.2008.12.072

- Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S.M., Davatzikos, C., 2004. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *NeuroImage* 21, 46–57. doi:10.1016/j.neuroimage.2003.09.027
- Lautrup, B., Hansen, L.K., Law, I., Svarer, C., Strother, S.C., 1994. Massive Weight Sharing: A Cure For Extremely Ill-Posed Problems, in: *Supercomputing in Brain Research: From Tomography to Neural Networks*. World Scientific Pub. Corp. World Scientific, pp. 137–148.
- Leemans, A., Jeurissen, B., Sijbers, J., Jones, D.K., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data, in: *17th Annual Meeting of Intl Soc Mag Reson Med*. p. 3537.
- Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. *NeuroImage* 56, 387–399. doi:10.1016/j.neuroimage.2010.11.004
- Lerch, J.P., Pruessner, J., Zijdenbos, A.P., Collins, D.L., Teipel, S.J., Hampel, H., Evans, A.C., 2008. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiol. Aging* 29, 23–30. doi:10.1016/j.neurobiolaging.2006.09.013
- Lerche, S., Hobert, M., Brockmann, K., Wurster, I., Gaenslen, A., Hasmann, S., Eschweiler, G.W., Maetzler, W., Berg, D., 2014. Mild Parkinsonian Signs in the Elderly – Is There an Association with PD? Crosssectional Findings in 992 Individuals. *PLOS ONE* 9, e92878. doi:10.1371/journal.pone.0092878
- Leung, K.K., Ridgway, G.R., Ourselin, S., Fox, N.C., 2012. Consistent multi-time-point brain atrophy estimation from the boundary shift integral. *NeuroImage* 59, 3995–4005. doi:10.1016/j.neuroimage.2011.10.068
- Libon, D.J., Xie, S.X., Eppig, J., Wicas, G., Lamar, M., Lippa, C., Bettcher, B.M., Price, C.C., Giovannetti, T., Swenson, R., Wambach, D.M., 2010. The heterogeneity of mild cognitive impairment: A neuropsychological analysis. *J. Int. Neuropsychol. Soc.* 16, 84–93. doi:10.1017/S1355617709990993
- Liepert-Scarfone, I., Brändle, B., Yilmaz, R., Gauss, K., Schaeffer, E., Timmers, M., Wurster, I., Brockmann, K., Maetzler, W., Van Nueten, L., Streffer, J.R., Berg, D., 2017. Progression of prodromal motor and non-motor symptoms in the premotor phase study – 2-year follow-up data. *Eur. J. Neurol.* n/a-n/a. doi:10.1111/ene.13397
- Liepert-Scarfone, I., Gauss, K., Maetzler, W., Müller, K., Bormann, C., Fruhmann Berger, M., Timmers, M., Streffer, J., Berg, D., 2013a. Evaluation of progression markers in the premotor phase of Parkinson's disease: the progression markers in the premotor phase study. *Neuroepidemiology* 41, 174–182. doi:10.1159/000353560
- Liepert-Scarfone, I., Pilotto, A., Müller, K., Bormann, C., Gauss, K., Wurster, I., Streffer, J., Berg, D., 2015. Autonomic dysfunction in subjects at high risk for Parkinson's disease. *J. Neurol.* 262, 2643–2652. doi:10.1007/s00415-015-7888-z
- Litvan, I., Mangone, C.A., McKee, A., Verny, M., Parsa, A., Jellinger, K., D'Olhaberriague, L., Chaudhuri, K.R., Pearce, R.K., 1996. Natural history of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome) and clinical predictors of survival: a clinicopathological study. *J. Neurol. Neurosurg. Psychiatry* 60, 615–620.
- Lobo, A., Launer, L.J., Fratiglioni, L., Andersen, K., Di Carlo, A., Breteler, M.M., Copeland, J.R., Dartigues, J.F., Jagger, C., Martinez-Lage, J., Soininen, H., Hofman, A., 2000. Prevalence of dementia and major subtypes in Europe: A collaborative study of population-based cohorts. *Neurologic Diseases in the Elderly Research Group*. *Neurology* 54, S4-9.
- Long, D., Wang, J., Xuan, M., Gu, Q., Xu, X., Kong, D., Zhang, M., 2012. Automatic Classification of Early Parkinson's Disease with Multi-Modal MR Imaging. *PLOS ONE* 7. doi:10.1371/journal.pone.0047714
- Lutkenhoff, E.S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J.D., Owen, A.M., Monti, M.M., 2014. Optimized Brain Extraction for Pathological Brains (optiBET). *PLOS ONE* 9, e115551. doi:10.1371/journal.pone.0115551

- Magnin, B., Mesrob, L., Kinkingnéhun, S., Péligrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehericy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51, 73–83. doi:10.1007/s00234-008-0463-x
- Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2010. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. *J. Cogn. Neurosci.* 22, 2677–2684. doi:10.1162/jocn.2009.21407
- Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., Poewe, W., Mollenhauer, B., Klinik, P.-E., Sherer, T., Frasier, M., Meunier, C., Rudolph, A., Casaceli, C., Seibyl, J., Mendick, S., Schuff, N., Zhang, Y., Toga, A., Crawford, K., Ansbach, A., De Blasio, P., Piovela, M., Trojanowski, J., Shaw, L., Singleton, A., Hawkins, K., Eberling, J., Brooks, D., Russell, D., Leary, L., Factor, S., Sommerfeld, B., Hogarth, P., Pighetti, E., Williams, K., Standaert, D., Guthrie, S., Hauser, R., Delgado, H., Jankovic, J., Hunter, C., Stern, M., Tran, B., Leverenz, J., Baca, M., Frank, S., Thomas, C.-A., Richard, I., Deeley, C., Rees, L., Sprenger, F., Lang, E., Shill, H., Obradov, S., Fernandez, H., Winters, A., Berg, D., Gauss, K., Galasko, D., Fontaine, D., Mari, Z., Gerstenhaber, M., Brooks, D., Malloy, S., Barone, P., Longo, K., Comery, T., Ravina, B., Grachev, I., Gallagher, K., Collins, M., Widnell, K.L., Ostrowizki, S., Fontoura, P., Ho, T., Luthman, J., Brug, M. van der, Reith, A.D., Taylor, P., 2011. The Parkinson Progression Marker Initiative (PPMI). *Prog. Neurobiol., Biological Markers for Neurodegenerative Diseases* 95, 629–635. doi:10.1016/j.pneurobio.2011.09.005
- Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., Mourão-Miranda, J., 2010. Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage* 49, 2178–2189. doi:10.1016/j.neuroimage.2009.10.072
- Marquand, A.F., Brammer, M., Williams, S.C.R., Doyle, O.M., 2014. Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage* 92, 298–311. doi:10.1016/j.neuroimage.2014.02.008
- Marquand, A.F., Filippone, M., Ashburner, J., Girolami, M., Mourao-Miranda, J., Barker, G.J., Williams, S.C.R., Leigh, P.N., Blain, C.R.V., 2013. Automated, High Accuracy Classification of Parkinsonian Disorders: A Pattern Recognition Approach. *PLoS ONE* 8, e69237. doi:10.1371/journal.pone.0069237
- Massano, J., Bhatia, K.P., 2012. Clinical Approach to Parkinson's Disease: Features, Diagnosis, and Principles of Management. *Cold Spring Harb. Perspect. Med.* 2. doi:10.1101/cshperspect.a008870
- McDonald, C.R., McEvoy, L.K., Gharapetian, L., Fennema-Notestine, C., Hagler, D.J., Holland, D., Koyama, A., Brewer, J.B., Dale, A.M., 2009. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. *Neurology* 73, 457–465. doi:10.1212/WNL.0b013e3181b16431
- McEvoy, L.K., Fennema-Notestine, C., Roddey, J.C., Hagler, D.J., Holland, D., Karow, D.S., Pung, C.J., Brewer, J.B., Dale, A.M., Alzheimer's Disease Neuroimaging Initiative, 2009. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology* 251, 195–205. doi:10.1148/radiol.2511080924
- McEvoy, L.K., Holland, D., Hagler, D.J., Fennema-Notestine, C., Brewer, J.B., Dale, A.M., 2011. Mild Cognitive Impairment: Baseline and Longitudinal Structural MR Imaging Measures Improve Predictive Prognosis. *Radiology* 259, 834–843. doi:10.1148/radiol.11101975
- McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. *Hum. Brain Mapp.* 6, 160–188.

- McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M., 1984. Clinical diagnosis of Alzheimer's disease Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 34, 939–939. doi:10.1212/WNL.34.7.939
- McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* 7, 263–269. doi:10.1016/j.jalz.2011.03.005
- Meek, P.D., McKeithan, K., Schumock, G.T., 1998. Economic considerations in Alzheimer's disease. *Pharmacotherapy* 18, 68-73; discussion 79-82.
- Migliaccio, R., Agosta, F., Possin, K.L., Rabinovici, G.D., Miller, B.L., Gorno-Tempini, M.L., 2012. White matter atrophy in Alzheimer's disease variants. *Alzheimers Dement. J. Alzheimers Assoc.* 8, S78–S87.e2. doi:10.1016/j.jalz.2012.04.010
- Miller, D.B., O'Callaghan, J.P., 2015. Biomarkers of Parkinson's disease: Present and future. *Metabolism*. 64, S40–S46. doi:10.1016/j.metabol.2014.10.030
- Minoshima, S., Frey, K.A., Koeppe, R.A., Foster, N.L., Kuhl, D.E., 1995. A Diagnostic Approach in Alzheimer's Disease Using Three-Dimensional Stereotactic Surface Projections of Fluorine-18-FDG PET. *J. Nucl. Med.* 36, 1238–1248.
- Minoshima, S., Giordani, B., Berent, S., Frey, K.A., Foster, N.L., Kuhl, D.E., 1997. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Ann. Neurol.* 42, 85–94. doi:10.1002/ana.410420114
- Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage* 44, 1415–1422. doi:10.1016/j.neuroimage.2008.10.031
- Mönninghoff, C., Dlugaj, M., Kraff, O., Geisel, M.H., Jöckel, K.-H., Erbel, R., Weimar, C., Wanke, I., 2015. Are Transversal MR Images Sufficient to Distinguish Persons with Mild Cognitive Impairment From Healthy Controls? *Acad. Radiol.* 22, 1172–1180. doi:10.1016/j.acra.2015.04.008
- Mørch, N., Hansen, L.K., Strother, S.C., Svarer, C., Rottenberg, D.A., Lautrup, B., Savoy, R., Paulson, O.B., 1997. Nonlinear Versus Linear Models in Functional Neuroimaging: Learning Curves and Generalization Crossover, in: *Proceedings of the 15th International Conference on Information Processing in Medical Imaging, IPMI '97*. Springer-Verlag, London, UK, UK, pp. 259–270.
- Morris, J.C., Storandt, M., Miller, J.P., McKeel, D.W., Price, J.L., Rubin, E.H., Berg, L., 2001. Mild cognitive impairment represents early-stage Alzheimer disease. *Arch. Neurol.* 58, 397–405.
- Mosconi, L., 2005. Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. *Eur. J. Nucl. Med. Mol. Imaging* 32, 486–510. doi:10.1007/s00259-005-1762-7
- Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., de Leon, M.J., 2010. Pre-Clinical Detection of Alzheimer's Disease Using FDG-PET, with or without Amyloid Imaging. *J. Alzheimers Dis. JAD* 20, 843–854. doi:10.3233/JAD-2010-091504
- Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *NeuroImage, Special Section: Social Cognitive Neuroscience* 28, 980–995. doi:10.1016/j.neuroimage.2005.06.070
- Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 12, 229–244. doi:10.1007/s12021-013-9204-3
- Natarajan, N., Dhillon, I.S., Ravikumar, P., Tewari, A., 2013. Learning with Noisy Labels, in: *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13*. Curran Associates Inc., USA, pp. 1196–1204.

- Nayate, A.P., Dubroff, J.G., Schmitt, J.E., Nasrallah, I., Kishore, R., Mankoff, D., Pryma, D.A., 2015. Use of Standardized Uptake Value Ratios Decreases Interreader Variability of [18F] Florbetapir PET Brain Scan Interpretation. *Am. J. Neuroradiol.* 36, 1237–1244. doi:10.3174/ajnr.A4281
- Nestor, S.M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J.L., Fogarty, J., Bartha, R., Initiative, the A.D.N., 2008a. Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's disease neuroimaging initiative database. *Brain* 131, 2443–2454. doi:10.1093/brain/awn146
- Nho, K., Shen, L., Kim, S., Risacher, S.L., West, J.D., Foroud, T., Jack, C.R., Weiner, M.W., Saykin, A.J., 2010. Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging. *AMIA. Annu. Symp. Proc.* 2010, 542–546.
- Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25.
- Nir, T.M., Villalon-Reina, J.E., Prasad, G., Jahanshad, N., Joshi, S.H., Toga, A.W., Bernstein, M.A., Jack Jr., C.R., Weiner, M.W., Thompson, P.M., 2015. Diffusion weighted imaging-based maximum density path analysis and classification of Alzheimer's disease. *Neurobiol. Aging, Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD)* 36, Supplement 1, S132–S140. doi:10.1016/j.neurobiolaging.2014.05.037
- Noirhomme, Q., Lesenfans, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. *NeuroImage Clin.* 4, 687–694. doi:10.1016/j.nicl.2014.04.004
- O'Brien, J.T., Herholz, K., 2015. Amyloid imaging for dementia in clinical practice. *BMC Med.* 13, 163. doi:10.1186/s12916-015-0404-6
- Olanow, C.W., Obeso, J.A., 2012. The significance of defining preclinical or prodromal Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* 27, 666–669. doi:10.1002/mds.25019
- Oliveira, F.P.M., Castelo-Branco, M., 2015. Computer-aided diagnosis of Parkinson's disease based on [123 I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. *J. Neural Eng.* 12, 026008. doi:10.1088/1741-2560/12/2/026008
- Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012a. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neurosci. Biobehav. Rev.* 36, 1140–1152. doi:10.1016/j.neubiorev.2012.01.004
- Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 559–572.
- Péran, P., Cherubini, A., Assogna, F., Piras, F., Quattrocchi, C., Peppe, A., Celsis, P., Rascol, O., Démonet, J.-F., Stefani, A., Pierantozzi, M., Pontieri, F.E., Caltagirone, C., Spalletta, G., Sabatini, U., 2010. Magnetic resonance imaging markers of Parkinson's disease nigrostriatal signature. *Brain J. Neurol.* 133, 3423–3433. doi:10.1093/brain/awq212
- Petersen, R.C., 2004a. Mild cognitive impairment as a diagnostic entity. *J. Intern. Med.* 256, 183–194. doi:10.1111/j.1365-2796.2004.01388.x
- Pezard, L., Jech, R., Růžicka, E., 2001. Investigation of non-linear properties of multichannel EEG in the early stages of Parkinson's disease. *Clin. Neurophysiol.* 112, 38–45. doi:10.1016/S1388-2457(00)00512-5
- Plant, C., Teipel, S.J., Oswald, A., Böhm, C., Meindl, T., Mourao-Miranda, J., Bokde, A.W., Hampel, H., Ewers, M., 2010. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage* 50, 162–174. doi:10.1016/j.neuroimage.2009.11.046
- Platt, J., 1998a. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Microsoft Res.*

- Prashanth, R., Dutta Roy, S., Mandal, P.K., Ghosh, S., 2016a. High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning. *Int. J. Med. Inf.* 90, 13–21. doi:10.1016/j.ijmedinf.2016.03.001
- Prashanth, R., Roy, S.D., Mandal, P.K., Ghosh, S., 2016b. High Accuracy Classification of Parkinson's Disease through Shape Analysis and Surface Fitting in 123I-Ioflupane SPECT Imaging. *IEEE J. Biomed. Health Inform.* PP, 1–1. doi:10.1109/JBHI.2016.2547901
- Przuntek, H., Müller, T., Riederer, P., 2004. Diagnostic staging of Parkinson's disease: conceptual aspects. *J. Neural Transm. Vienna Austria* 1996 111, 201–216. doi:10.1007/s00702-003-0102-y
- Pyatigorskaya, N., Gallea, C., Garcia-Lorenzo, D., Vidailhet, M., Lehericy, S., 2014. A review of the use of magnetic resonance imaging in Parkinson's disease. *Ther. Adv. Neurol. Disord.* 7, 206–220. doi:10.1177/1756285613511507
- Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., Initiative, T.A.D.N., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain* 132, 2036–2047. doi:10.1093/brain/awp105
- Rae, C.L., Correia, M.M., Altena, E., Hughes, L.E., Barker, R.A., Rowe, J.B., 2012. White matter pathology in Parkinson's disease: The effect of imaging protocol differences and relevance to executive function. *NeuroImage* 62, 1675–1684. doi:10.1016/j.neuroimage.2012.06.012
- Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L., 2001. A default mode of brain function. *Proc. Natl. Acad. Sci.* 98, 676–682. doi:10.1073/pnas.98.2.676
- Rasmussen, C.E., Williams, C.K.I., 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass.
- Raz, N., Ghisletta, P., Rodrigue, K.M., Kennedy, K.M., Lindenberger, U., 2010. Trajectories of brain aging in middle-aged and older adults: Regional and individual differences. *NeuroImage* 51, 501–511. doi:10.1016/j.neuroimage.2010.03.020
- Raz, N., Lindenberger, U., 2011a. Only Time will Tell: Cross-sectional Studies Offer no Solution to the Age-Brain-Cognition Triangle—Comment on. *Psychol. Bull.* 137, 790–795. doi:10.1037/a0024503
- Risacher, S.L., Saykin, A.J., 2013. Neuroimaging and Other Biomarkers for Alzheimer's Disease: The Changing Landscape of Early Detection. *Annu. Rev. Clin. Psychol.* 9, 621–648. doi:10.1146/annurev-clinpsy-050212-185535
- Risacher, S.L., Shen, L., West, J.D., Kim, S., McDonald, B.C., Beckett, L.A., Harvey, D.J., Jack Jr., C.R., Weiner, M.W., Saykin, A.J., 2010. Longitudinal MRI atrophy biomarkers: Relationship to conversion in the ADNI cohort. *Neurobiol. Aging, Alzheimer's Disease Neuroimaging Initiative (ADNI) Studies* 31, 1401–1418. doi:10.1016/j.neurobiolaging.2010.04.029
- Ross, G.W., Petrovitch, H., Abbott, R.D., Nelson, J., Markesbery, W., Davis, D., Hardman, J., Launer, L., Masaki, K., Tanner, C.M., White, L.R., 2004. Parkinsonian signs and substantia nigra neuron density in decedents elders without PD. *Ann. Neurol.* 56, 532–539. doi:10.1002/ana.20226
- Ryali, S., Supekar, K., Abrams, D.A., Menon, V., 2010. Sparse logistic regression for whole brain classification of fMRI data. *NeuroImage* 51, 752–764. doi:10.1016/j.neuroimage.2010.02.040
- Sabuncu, M.R., Konukoglu, E., 2014. Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study. *Neuroinformatics* 1–16. doi:10.1007/s12021-014-9238-1
- Salamanca, L., Vlassis, N., Diederich, N., Bernard, F., Skupin, A., 2015. Improved Parkinson's Disease Classification from Diffusion MRI Data by Fisher Vector Descriptors, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Lecture*

- Notes in Computer Science. Springer International Publishing, pp. 119–126. doi:10.1007/978-3-319-24571-3_15
- Salas-Gonzalez, D., Górriz, J.M., Ramírez, J., Illán, I.A., López, M., Segovia, F., Chaves, R., Padilla, P., Puntonet, C.G., 2010. Feature selection using factor analysis for Alzheimer's diagnosis using 18F-FDG PET images. *Med. Phys.* 37, 6084–6095. doi:10.1118/1.3488894
- Salvatore, C., Cerasa, A., Castiglioni, I., Gallivanone, F., Augimeri, A., Lopez, M., Arabia, G., Morelli, M., Gilardi, M.C., Quattrone, A., 2014. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. *J. Neurosci. Methods* 222, 230–237. doi:10.1016/j.jneumeth.2013.11.016
- Schmermund, A., Möhlenkamp, S., Stang, A., Grönemeyer, D., Seibel, R., Hirche, H., Mann, K., Siffert, W., Lauterbach, K., Siegrist, J., Jöckel, K.-H., Erbel, R., 2002. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: Rationale and design of the Heinz Nixdorf RECALL Study. *Am. Heart J.* 144, 212–218. doi:10.1067/mhj.2002.123579
- Schneider, S.A., Edwards, M.J., Mir, P., Cordivari, C., Hooker, J., Dickson, J., Quinn, N., Bhatia, K.P., 2007. Patients with adult-onset dystonic tremor resembling parkinsonian tremor have scans without evidence of dopaminergic deficit (SWEDDs). *Mov. Disord.* 22, 2210–2215. doi:10.1002/mds.21685
- Schnitzler, A., Gross, J., 2005. Normal and pathological oscillatory communication in the brain. *Nat. Rev. Neurosci.* 6, 285–296. doi:10.1038/nrn1650
- Scholkopf, B., Smola, A., Müller, K.-R., 1999. Kernel principal component analysis, in: *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING*. MIT Press, pp. 327–352.
- Schölkopf, B., Smola, A.J., 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.
- Schrouff, J., Rosa, M.J., Rondina, J.M., Marquand, A.F., Chu, C., Ashburner, J., Phillips, C., Richiardi, J., Mourão-Miranda, J., 2013. PRoNTo: Pattern Recognition for Neuroimaging Toolbox. *Neuroinformatics* 11, 319–337. doi:10.1007/s12021-013-9178-1
- Schuster, C., Elamin, M., Hardiman, O., Bede, P., 2015. Presymptomatic and longitudinal neuroimaging in neurodegeneration—from snapshots to motion picture: a systematic review. *J Neurol Neurosurg Psychiatry* jnnp-2014-309888. doi:10.1136/jnnp-2014-309888
- Schwarz, S.T., Abaei, M., Gontu, V., Morgan, P.S., Bajaj, N., Auer, D.P., 2013. Diffusion tensor imaging of nigral degeneration in Parkinson's disease: A region-of-interest and voxel-based study at 3 T and systematic review with meta-analysis. *NeuroImage Clin.* 3, 481–488. doi:10.1016/j.nicl.2013.10.006
- Serrano-Pozo, A., Frosch, M.P., Masliah, E., Hyman, B.T., 2011. Neuropathological Alterations in Alzheimer Disease. *Cold Spring Harb. Perspect. Med.* 1. doi:10.1101/cshperspect.a006189
- Shawe-Taylor, J., Cristianini, N., 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK ; New York.
- Silverman, D.H., Small, G.W., Chang, C.Y., Lu, C.S., Kung De Aburto, M.A., Chen, W., Czernin, J., Rapoport, S.I., Pietrini, P., Alexander, G.E., Schapiro, M.B., Jagust, W.J., Hoffman, J.M., Welsh-Bohmer, K.A., Alavi, A., Clark, C.M., Salmon, E., de Leon, M.J., Mielke, R., Cummings, J.L., Kowell, A.P., Gambhir, S.S., Hoh, C.K., Phelps, M.E., 2001. Positron emission tomography in evaluation of dementia: Regional brain metabolism and long-term outcome. *JAMA* 286, 2120–2127.
- Sjöbeck, M., Haglund, M., Englund, E., 2005. Decreasing myelin density reflected increasing white matter pathology in Alzheimer's disease—a neuropathological study. *Int. J. Geriatr. Psychiatry* 20, 919–926. doi:10.1002/gps.1384
- Smith, S.M., De Stefano, N., Jenkinson, M., Matthews, P.M., 2001. Normalized accurate measurement of longitudinal brain change. *J. Comput. Assist. Tomogr.* 25, 466–475.

- Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., Behrens, T.E.J., 2006. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31, 1487–1505. doi:10.1016/j.neuroimage.2006.02.024
- Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004a. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage, Mathematics in Brain Imaging* 23, Supplement 1, S208–S219. doi:10.1016/j.neuroimage.2004.07.051
- Smith, S.M., Rao, A., De Stefano, N., Jenkinson, M., Schott, J.M., Matthews, P.M., Fox, N.C., 2007. Longitudinal and cross-sectional analysis of atrophy in Alzheimer's disease: Cross-validation of BSI, SIENA and SIENAX. *NeuroImage* 36, 1200–1206. doi:10.1016/j.neuroimage.2007.04.035
- Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P.M., Federico, A., De Stefano, N., 2002a. Accurate, Robust, and Automated Longitudinal and Cross-Sectional Brain Change Analysis. *NeuroImage* 17, 479–489. doi:10.1006/nimg.2002.1040
- Soikkeli, R., Partanen, J., Soininen, H., Pääkkönen, A., Riekkinen, P., 1991. Slowing of EEG in Parkinson's disease. *Electroencephalogr. Clin. Neurophysiol.* 79, 159–165.
- Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H., 2011. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement. J. Alzheimers Assoc.* 7, 280–292. doi:10.1016/j.jalz.2011.03.003
- Stang, A., Moebus, S., Dragano, N., Beck, E.M., Möhlenkamp, S., Schmermund, A., Siegrist, J., Erbel, R., Jöckel, K.H., Heinz Nixdorf Recall Study Investigation Group, 2005. Baseline recruitment and analyses of nonresponse of the Heinz Nixdorf Recall Study: identifiability of phone numbers as the major determinant of response. *Eur. J. Epidemiol.* 20, 489–496.
- Stebbins, G.T., Murphy, C.M., 2009. Diffusion tensor imaging in Alzheimer's disease and mild cognitive impairment. *Behav. Neurol.* 21, 39–49. doi:10.3233/BEN-2009-0234
- Strimbu, K., Tavel, J.A., 2010. What are Biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466. doi:10.1097/COH.0b013e32833ed177
- Studholme, C., Cardenas, V., Blumenfeld, R., Schuff, N., Rosen, H.J., Miller, B., Weiner, M., 2004. Deformation tensor morphometry of semantic dementia with quantitative validation. *NeuroImage* 21, 1387–1398. doi:10.1016/j.neuroimage.2003.12.009
- Tanaka, H., Koenig, T., Pascual-Marqui, R.D., Hirata, K., Kochi, K., Lehmann, D., 2000. Event-related potential and EEG measures in Parkinson's disease without and with dementia. *Dement. Geriatr. Cogn. Disord.* 11, 39–45. doi:17212
- Teipel, S.J., Bokde, A.L.W., Born, C., Meindl, T., Reiser, M., Möller, H.-J., Hampel, H., 2007a. Morphological substrate of face matching in healthy ageing and mild cognitive impairment: a combined MRI-fMRI study. *Brain J. Neurol.* 130, 1745–1758. doi:10.1093/brain/awm117
- Teipel, S.J., Born, C., Ewers, M., Bokde, A.L.W., Reiser, M.F., Möller, H.-J., Hampel, H., 2007b. Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *NeuroImage* 38, 13–24. doi:10.1016/j.neuroimage.2007.07.008
- Turk, M., Pentland, A., 1991. Eigenfaces for Recognition. *J Cogn. Neurosci.* 3, 71–86. doi:10.1162/jocn.1991.3.1.71
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated Anatomical Labeling of Activations in

- SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage* 15, 273–289. doi:10.1006/nimg.2001.0978
- Vaillancourt, D.E., Spraker, M.B., Prodoehl, J., Abraham, I., Corcos, D.M., Zhou, X.J., Comella, C.L., Little, D.M., 2009. High-resolution diffusion tensor imaging in the substantia nigra of de novo Parkinson disease. *Neurology* 72, 1378–1384. doi:10.1212/01.wnl.0000340982.01727.6e
- Vandenberghe, R., Nelissen, N., Salmon, E., Ivanoiu, A., Hasselbalch, S., Andersen, A., Korner, A., Minthon, L., Brooks, D.J., Van Laere, K., Dupont, P., 2013. Binary classification of 18F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI. *NeuroImage* 64, 517–525. doi:10.1016/j.neuroimage.2012.09.015
- Vapnik, 1998. *Statistical Learning Theory*. New York: Wiley Interscience.
- Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage, Individual Subject Prediction* 145, 166–179. doi:10.1016/j.neuroimage.2016.10.038
- Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack, C.R., 2008a. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage* 39, 1186–1197. doi:10.1016/j.neuroimage.2007.09.073
- Vemuri, P., Jack, C.R., 2010. Role of structural MRI in Alzheimer's disease. *Alzheimers Res. Ther.* 2, 23. doi:10.1186/alzrt47
- Vemuri, P., Whitwell, J.L., Kantarci, K., Josephs, K.A., Parisi, J.E., Shiung, M.S., Knopman, D.S., Boeve, B.F., Petersen, R.C., Dickson, D.W., Jack, C.R., 2008b. Antemortem MRI based STructural Abnormality iNDex (STAND)-scores correlate with postmortem Braak neurofibrillary tangle stage. *NeuroImage* 42, 559–567. doi:10.1016/j.neuroimage.2008.05.012
- Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack, C.R., 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects. *Neurology* 73, 294–301. doi:10.1212/WNL.0b013e3181af79fb
- Vicente, M.A., Hoyer, P.O., Hyvarinen, A., 2007. Equivalence of Some Common Linear Feature Extraction Techniques for Appearance-Based Object Recognition Tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 896–900. doi:10.1109/TPAMI.2007.1074
- Weinberger, D.R., Goldberg, T.E., 2014. RDoCs redux. *World Psychiatry* 13, 36–38. doi:10.1002/wps.20096
- Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H., Fratiglioni, L., Frisoni, G.B., Gauthier, S., Georges, J., Graff, C., Iqbal, K., Jessen, F., Johansson, G., Jönsson, L., Kivipelto, M., Knapp, M., Mangialasche, F., Melis, R., Nordberg, A., Rikkert, M.O., Qiu, C., Sakmar, T.P., Scheltens, P., Schneider, L.S., Sperling, R., Tjernberg, L.O., Waldemar, G., Wimo, A., Zetterberg, H., 2016. Defeating Alzheimer's disease and other dementias: a priority for European science and society. *Lancet Neurol.* 15, 455–532. doi:10.1016/S1474-4422(16)00062-4
- Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., Nordberg, A., Bäckman, L., Albert, M., Almkvist, O., Arai, H., Basun, H., Blennow, K., De Leon, M., DeCarli, C., Erkinjuntti, T., Giacobini, E., Graff, C., Hardy, J., Jack, C., Jorm, A., Ritchie, K., Van Duijn, C., Visser, P., Petersen, R. c., 2004. Mild cognitive impairment – beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *J. Intern. Med.* 256, 240–246. doi:10.1111/j.1365-2796.2004.01380.x
- Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: A review of pattern recognition

- for neuroimaging-based psychiatric diagnostics. *Neurosci. Biobehav. Rev.* 57, 328–349. doi:10.1016/j.neubiorev.2015.08.001
- Wolz, R., Aljabar, P., Hajnal, J.V., Rueckert, D., 2010. Manifold Learning for Biomarker Discovery in MR Imaging, in: Wang, F., Yan, P., Suzuki, K., Shen, D. (Eds.), *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 116–123.
- Wu, T., Hallett, M., 2013. The cerebellum in Parkinson's disease. *Brain* 136, 696–709. doi:10.1093/brain/aws360
- Wu, X., Li, J., Ayutyanont, N., Protas, H., Jagust, W., Fleisher, A., Reiman, E., Yao, L., Chen, K., 2013. The Receiver Operational Characteristic for Binary Classification with Multiple Indices and Its Application to the Neuroimaging Study of Alzheimer's Disease. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 173–180. doi:10.1109/TCBB.2012.141
- Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 2564–2577. doi:10.1093/brain/awu176
- Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage Clin.* 2, 735–745. doi:10.1016/j.nicl.2013.05.004
- Yu, P., Dean, R.A., Hall, S.D., Qi, Y., Sethuraman, G., Willis, B.A., Siemers, E.R., Martenyi, F., Tauscher, J.T., Schwarz, A.J., 2012. Enriching amnesic mild cognitive impairment populations for clinical trials: optimal combination of biomarkers to predict conversion to dementia. *J. Alzheimers Dis. JAD* 32, 373–385. doi:10.3233/JAD-2012-120832
- Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* 59, 895–907. doi:10.1016/j.neuroimage.2011.09.069
- Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* 55, 856–867. doi:10.1016/j.neuroimage.2011.01.008
- Zhou, J., Greicius, M.D., Gennatas, E.D., Growdon, M.E., Jang, J.Y., Rabinovici, G.D., Kramer, J.H., Weiner, M., Miller, B.L., Seeley, W.W., 2010. Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. *Brain* 133, 1352–1367. doi:10.1093/brain/awq075
- Ziegler, G., Penny, W.D., Ridgway, G.R., Ourselin, S., Friston, K.J., 2015a. Estimating anatomical trajectories with Bayesian mixed-effects modeling. *NeuroImage* 121, 51–68. doi:10.1016/j.neuroimage.2015.06.094