Multi-Modality Inference Methods for Neuroimaging with Applications to Alzheimer's Disease Research

By

Christopher D. Hinrichs

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2012

Date of final oral examination: 12/12/12

The dissertation is approved by the following members of the Final Oral Committee:

Vikas Singh, Assistant Professor, Biostatistics & Medical Informatics
Sterling C. Johnson, Professor, Medicine
Moo K. Chung, Associate Professor, Biostatistics & Medical Informatics
Grace Wahba, Professor, Statistics
Charles R. Dyer, Professor, Computer Sciences

For Joey-
Your loss was everyone's loss.

Acknowledgments

My being here, having produced this thesis, is a very, very improbable event. An event of such significance does not happen by itself – it took a concerted, perhaps even conspiratorial, effort on the part of a lot of people to put me here. I merely followed their lead, trying to keep up as best I could.

The road to where I am now began in late 2002 with Steven Bandyk. Steve had been my boss at UIC while I worked in the microrepair office, or "the cave" as it was better known. Just as I was finishing my bachelors at UIC, Steve left for greener pastures at U *of* C (*i.e.*, the University of Chicago – the place people *think* you went to when you tell them you're from UIC). As I wasn't having much luck finding a job, he picked me out of a lineup of a hundred or so student workers to join him as a full-time employee in the Physical Sciences Division Desktop Support office – PSDDS. I doubt if it's even theoretically possible for there to be a more enjoyable place to work, and I have many happy memories from that time – not to mention a few unlikely stories as well. But beyond that, being there often meant having a few odd moments to chat with a lot of *very* talented physicists, chemists, string theorists, grad students and scientists of all stripes. Inevitably, this fundamentally altered my perception of what it really means to be in science. The stage was set, and from that point, it only took a nudge to really get the ball rolling.

That nudge came in the form of one Philip Eaton, who, among other things, is well known for demonstrating the synthesis of the most powerful chemical explosive known to humanity. (Yes, that sentence is actually true.) A gracious and soft-spoken man, professor Eaton was the one who told me that I had no business fixing e-mail problems and installing printers for a living, but that instead I should be doing something more with my life. So, in 2004 I enrolled in an evening Masters program at Chicago, with recommendations from Steve, professor Eaton, and one of his colleagues, (another well known chemist whom he prevailed upon to help me out). A year later, I decided to take a "real" computer science course – a doctoral level computability and Kolmogorov complexity theory course with Robert Soare, for which I was thoroughly unprepared, but survived nonetheless. That was a sufficient test case for me to decide that I had some chance of success in a Ph.D. program, and so with professor Soare's help and

guidance, along with several others in the department, I applied to several graduate schools. The outcome should be fairly obvious at this stage.

Of course, taking a single graduate course in computability theory is not sufficient to transform a B-C, (and occasionally even a D!) undergrad into a star graduate student – or even an average one. Unsurprisingly, I struggled quite a lot with the transition during my first few years in Madison. Probably the biggest help in that transition came from Michael Ferris, for whom I was a TA in Linear Programming. Professor Ferris' patient tutoring during our weekly meetings allowed me to make up for the dismal job I had done in that same course a year earlier, and served me *extremely* well once I began my research.

I cannot stress enough that the phrase, "once I began my research", would be a complete counterfactual were it not for my adviser, Vikas Singh. There are too many ways in which I've benefitted from Vikas' mentoring to mention here, but I will nevertheless try to give a sense of a few of them here, in no particular order. Writing: Academic writing is one of the most important skills gained in a doctoral program, and pretty much everything I know on that subject comes from Vikas. Actually, it comprises several subjects because there are a variety of different authorial voices to master, one for each of the different settings in which one has to write. I also include in this category all of the LaTeX tricks I learned from Vikas. Personal: During the four and a half years I've worked with Vikas, I've gone through a number of difficult situations, some of my own making, some not. In all cases, Vikas was a reliable friend and counselor, and he never failed to come through for me. More than that, Vikas has always been a strong advocate for me, even when I myself did not believe I could succeed. I only hope that if I ever have students of my own in the future that I can be as supportive for them as Vikas has been for me. Research: Probably the single thing I will miss the most about working with Vikas is the fluency of communication we've developed. I can explain to Vikas in 10 minutes what it would take 30 minutes to explain to anyone else, and he always sees right away the strengths and weaknesses of each new idea. (Well, almost always.) This single thing, more than any other, makes me want to pursue a career in academia. Support: Vikas hired me in the first place, and helped me substantially in applying for the CIBM fellowship that supported me for three years. It may seem obvious now, but this is really the "zero$^{\text{th}}$ order term".

Naturally, Vikas was not the only mentor I benefitted from. As soon as I started

working with Vikas, I also started working with Sterling Johnson's group in the Geriatrics Research, Educational and Clinical Center (GRECC), situated in the William S. Middleton Memorial Veterans Hospital. Everything I know about Alzheimer's Disease, neurobiology or any related topic I learned either directly from Sterling himself, or from others in his lab. Just to give an idea of how far I've come, when I first presented results at one of Sterling's lab meetings, I showed a sideways coronal slice of a brain image, thinking it was an axial slice. At the time I didn't even know what those terms meant. Those early steps were like being a baby learning to walk all over again. Sterling too has been a strong advocate for me, for which I am grateful.

Rounding out the rest of my committee, Moo Chung contributed a lot of ideas and suggestions right from the beginning, and was an all-around friend and mentor. When my first paper was published in NeuroImage, Moo told me, "One day this paper will have like a hundred citations. You'll see." So far it's up over 40, and counting... Moo also hosted me at Seoul National University for a month, and the work I did there led to a substantial portion of this thesis. My first conversation with Grace Wahba was at a poster session, and went like this: (Grace) "Hi. I'm interested in your research." (Me) "Oh, well I'm interested in *your* research." (Grace again) "Well I asked you first." Since then I've benefitted enormously from Grace's deep knowledge and sharp insights – and from her pointing out the surprising connections between her own research and mine. Though I haven't worked with him as closely as the others, Chuck Dyer went through the draft of this thesis with a *very* fine-toothed comb and made a lot of helpful suggestions. It undoubtedly reads a lot better now, and I am grateful for the assistance.

Other co-authors I've worked with directly include, (in roughly chronological order,) Guofan Xu, Lopamudra Mukherjee, Nagesh Adluru, Jiming Peng, Kamiya Motwani, Deepti Pachauri, Maritza Dowling, and Vamsi Ithapu. One "silent co-author" which has nevertheless been absolutely vital to my career is the Alzheimer's Disease Neuroimaging Initiative, or ADNI. Without the ADNI, I would not have had access to nearly the wealth of data that I did, (though Sterling's group is like an ADNI all by themselves). Other sources of support include the Computational Informatics in Biology and Medicine (CIBM) training program (NLM 5T15LM007359), whose administrators Louise Pape and Karen Nafzger really looked out for me; several major NIH grants, (NIH R21-AG034315) and (NIH R01-AG021155), and others. Ozioma Okonkwo has also been a substantive role model, with whom I hope to continue

collaborating. It's been my privilege also to share an office with Wonhwa Kim, Jia Xu and Maxwell Collins. Finally, Sangkyun Lee, Gary Pack, Jerry Zhu, Ben Recht and Steve Wright all gave helpful discussions along the way.

Outside of the academic sphere, a number of people helped keep me grounded throughout my time here in Madison. My sister Angie and her wife Danielle LeMay helped me keep my sanity, and made their home the default venue for Christmas each year. My Dad and stepmom Jody Glittenberg were especially supportive as I struggled to get through my quals. Brandon Smith, Maxwell Collins, and a host of others have been co-conspirators in home brewing for almost three years now. (Three years!) One time, thanks to a tip from David Malec I ended up having a few beers at the Dane with Aubrey de Grey. I've also had several awesome housemates: Matt Elder, Bill Harris, Evan Driscoll, (also a rock-climbing buddy,) Josh Yanchar and Mikola Lysenko. Nagesh Adluru has simply been a great friend.

Last, but by no means least, I would like to thank Chris Kielch, Alina MacKenzie, Mike & Kelly Lakas, Fredy Peralta and Robert B. Schmickelheimer for being campmates, drinking buddies, travelling companions, and fellow combatants in the never-ending war on *ennui* that is a well-rounded life.

Contents

Contents vi

List of Tables viii

List of Figures ix

NOMENCLATURE xi

Abstract xii

1 Introduction 1
1.1 Motivation and Context 1

2 Background 7
2.1 Alzheimer's Disease 7
2.2 Medical Imaging Modalities 15
2.3 Related Work in the AD Literature 20
2.4 Kernel Methods 24
2.5 Multi-Kernel Methods 28

3 Adaptation of Learning Methods to Neuroimaging Problems: Single Modality Methods 32
3.1 Motivation 33
3.2 Spatially Augmented LP Boosting 33
3.3 **Q***-SVM* 44

4 Adaptation of Learning Methods to Neuroimaging Problems: Multi-modality Methods 50
4.1 Examination of p*-norms* 50
4.2 Robustness to Outliers 52

5 Exploiting Modality-modality Interactions 61
5.1 **Q***-MKL* 61

5.2 The Case for **Q***-MKL* 65
5.3 Experiments 77
5.4 Conclusions 84

6 Machine Learning Approaches to Scientific Investigation of Alzheimer's Disease 88
6.1 Predictive Multi-modality Markers of Neurodegeneration 88
6.2 Discovery of Anomalous Subjects 90
6.3 Clinical Trial Enrichment 101

7 Linear Outcome Measures in Clinical Trials 108
7.1 Outcome Measures, and Related Statistical Concepts 111
7.2 A Motivating Example 113
7.3 Power Calculations 120
7.4 Monte Carlo Evaluations 122
7.5 Simulations Using the AD cohort 125
7.6 Conclusions 127

8 Future Directions and Open Questions 129
8.1 Efficient Large-scale Permutation Testing via Matrix Completion 129
8.2 Ongoing Applications to Planned Clinical Trials 137

9 Conclusion 140
9.1 Contributions 141
9.2 Summary 145

Bibliography 146

List of Tables

3.1 Results of classification experiments on ADNI data 40

4.1 Comparison of different MKL norms in the presence of uninformative kernels 51
4.2 Accuracy results for Robust MKL 58

5.1 **Q**-functions, their arguments, and their uses. 71
5.2 Comparison of **Q**-MKL & MKL 80
5.3 Performance measures for several **Q** functions on UCI datasets 87

6.1 Comparison of relevant biomarkers in group I AD and group II AD . . . 93
6.2 Comparison of relevant biomarkers in group I CN and group II CN . . . 94
6.3 Estimated sample cohort sizes for single modal and multimodal inclusion criteria 106

List of Figures

2.1 Several examples of anatomical variation among ADNI subjects 14

3.1 Spatial relationships between voxels 34
3.2 Weak classifier outputs 37
3.3 Classifier output and ROC curves 41
3.4 GMP brain regions selected 41
3.5 Classifier output and ROC curves 43
3.6 FDG-PET brain regions selected 43
3.7 **Q**-matrices and trained classifiers on an MNIST digits task 47
3.8 Comparison of spatial smoothness of **Q**-SVM and SVM weights 48

4.1 Toy example demonstrating the effect of outliers on kernels 57
4.2 ROC curves for Robust MKL 58
4.3 GMP and FDG-PET classifier weights 60

5.1 Covariance **Q** before taking Laplacian; and several eigen-vectors 81
5.2 Histogram **Q** before taking Laplacian; and several eigen-vectors 83
5.3 Eigen-space alignment **Q** before taking Laplacian; and several eigen-vectors 84
5.4 Training error covariance **Q** before taking Laplacian; and several eigen-vectors 85
5.5 Relevance maps in each modality assigned by **Q**-MKL 86

6.1 MMDMs applied to the MCI population 89
6.2 Weak classifier outputs and percent of weak classifiers giving incorrect outputs 92
6.3 Voxel weights assigned by the MKL classifier for FDG-PET images at baseline, with and without outliers 96
6.4 Voxel weights assigned by the MKL classifier for FDG-PET images at 24 months, with and without outliers 97
6.5 Voxel weights assigned by the MKL classifier for GM density images at baseline, with and without outliers 98

6.6 Voxel weights assigned by the MKL classifier for TBM images, with and without outliers 99
6.7 Sample cohort sizes as a function of number of TBM voxels and MCI subjects 105

7.1 Two cases which illustrate the relative strengths of each method 111
7.2 Sample covariance of the MCI cohort 116
7.3 Eigen-values of the MCI cohort sample covariance matrix 117
7.4 Experimental results with simulated Multivariate Gaussian data, using Σ_{MCI} 117
7.5 Experimental results with simulated Multivariate Gaussian data using $\Sigma = \mathrm{I}$ 118
7.6 p-values from simulated trials using all voxels, mapping stable MCI to converting MCI. 125
7.7 p-values from simulated clinical trials using CSF voxels only, mapping stable MCI to converting MCI. 126
7.8 p-values from simulated trials using all voxels, mapping CN to AD. . . . 127
7.9 p-values from simulated clinical trials using CSF voxels only, mapping CN to AD. 127

8.1 Distribution of the maximum t-statistic reconstructed from a varying sample as a percentage of all voxels. 136

DISCARD THIS PAGE

Nomenclature

LP-Boost	Linear Program Boosting – a type of linear classification algorithm
SVM	Support Vector Machine
Q-SVM	Support Vector Machine augmented by a quadratic regularizer
MKL	Multi-Kernel Learning
Q-MKL	Multi-Kernel Learning model augmented by a quadratic regularizer
AD	Alzheimer's Disease
CN	Cognitively Normal, *i.e.*, control subjects
MCI	Mild Cognitive Impairment
CSF	Cerebro-Spinal Fluid
APOE	Apolipo-Protein E: a gene known to influence Alzheimer's Disease
$A\beta$	Amyloid β protein
τ	τ protein
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
FDG	^{18}F Fluoro-Deoxy Glucose (PET tracer which mimics glucose)
PIB	Pittsburgh Compound B (PET tracer which binds to amyloid protein)

Abstract

An emphasis in ongoing Alzheimer's Disease (AD) research is identifying biomarkers which best predict future cognitive decline, especially at the earliest stages of disease progression. Ultimately, it is hoped that these biomarkers can serve as early markers for diagnosis, but in a more immediate time frame the can be used for selection of subjects into clinical trials. Recent results suggest that the identification of such discriminative biomarkers is possible by adapting machine learning methods for this problem: but studies have primarily used modalities in isolation so far. The sensitivity/specificity offered by these methods is as yet unsatisfactory for more clinically relevant questions, especially: which Mild Cognitive Impairment (MCI) patients will convert to AD? Answering such questions requires new methods that leverage all data sources (e.g., imaging modalities, CSF measures) in conjunction. A significant portion of this thesis is devoted to demonstrating that significant improvements in sensitivity and specificity for discriminating AD, MCI, and healthy controls at the level of individual subjects are possible by making use of multiple modalities (together with longitudinal data) simultaneously. This thesis focuses on showing how data from multiple biomarkers should be optimally aggregated to best predict future cognitive decline, using specially developed variants of Multiple Kernel Learning (MKL); how these models can improve sample size estimates in clinical trials for AD; and perhaps even redefine the central question under consideration in clinical trials by using customized outcomes for evaluating new treatment procedures.

This dissertaion presents new algorithms for a) introducing inductive biases into existing machine learning methods which are designed to fully capture and exploit the structure of the image data; b) combining various imaging modalities into a single multi-kernel predictive model via robust learning loss functions, and quadratic regularizers based on modality-modality interactions; c) using the above frameworks to derive sensitive custom measures of disease progressiong from medical neuroimaging data for use in clinical trials. I present extensive empirical evaluations and theoretical evidence that illustrate how tailor-made machine learning algorithms can transform neuroimaging analysis and clinical trials.

Chapter 1

Introduction

1.1 Motivation and Context

Recent decades have seen a remarkable and accelerating growth in the variety, power, and availability of medical scanning technologies, allowing unprecedented views of the inner workings of living tissues and organs. Starting in the 1970s, X-ray systems have developed into Computed Tomography (CT) systems; Nuclear Magnetic Resonance (NMR) led to the development of Magnetic Resonance Imaging (MRI); MRI further branched into functional MR Imaging (fMRI), Diffusion Tensor Imaging (DTI), Arterial Spin-Labeled perfusion imaging (ASL), and a host of others; Positron Emission Tomography (PET) uses a variety of tracer compounds to illuminate various biological processes, such as blood perfusion, glucose metabolization, dopamine uptake, amyloid protein deposition, etc. In addition to clinical applications, new avenues of scientific inquiry are opened up as well. The study of neurodegenerative disorders – Alzheimer's Disease (AD) in particular – has undergone a renaissance as an array of imaging modalities can now be used to track physiological changes and neurological degeneration at increasingly earlier stages, and to discover new risk factors and eliminate confounding variables. It is hoped that these investigations will yield sufficiently sensitive markers that interventions may be developed and deployed before irreversible degeneration sets in.

Yet, having more data is only one half of the equation. It is equally important that models and algorithms rise to the challenges posed by this wealth of data. When analyzing a large body of 3D scans in search of differentiative disease patterns, the sheer volume of data can itself become a challenge in several ways – first, computation times can become prohibitive, meaning that more efficient models will become more and more important. Second, as the dimensionality of imaging data grows, issues such as multiple comparisons and model complexity make results more difficult to interpret. Third, imaging data has certain inherent structural properties (or, covariance structures)

owing to the proximity of neighboring voxels, yet, statistical models often assume that variables are independent, or ignore dependency structures when they are known. These issues are only compounded by the growing number of modalities, which introduce even more complex dependencies among covariates. Recently, however, neuroimaging-driven AD research has turned a new corner as machine learning methods have been given a more central analysis role. Machine learning methods differ from traditional statistical methods in that they place more focus on generalizing to future, unseen data, in addition to explaining existing samples.

By acquiring scans of subjects suffering from a pathological condition of interest, as well as healthy controls, investigators can examine in detail the effects of pathology by separating individual variation from group-wise variation. If clinical groups are properly controlled then they will systematically differ only on the basis of disease, allowing hypotheses relating to disease processes to be tested. Traditionally this has been done by way of standard univariate statistical models which test whether *e.g.*, *means* vary between groups relative to their *standard deviations*.

Dimensionality problems are often controlled by taking a mean over a particular anatomical Region of Interest (ROI – a pre-defined set of voxels in the brain) which can obscure a large portion of the signal being sought. More recent investigations have moved beyond simple ROI analysis by utilizing Statistical Parametric Mapping (SPM) in which a group-wise statistic is computed separately at every voxel, giving a map of statistical significance levels. This map can then be interpreted in terms of known anatomical and functional regions, but at the cost of incurring serious multiple-comparisons issues. What differentiates the machine learning approach from these existing models is that rather than treating neuroimaging data as a collection of separate variables, *each of which gives rise to a separate model*, machine learning methods treat *only the patients* as being separate and independent, meaning that significance calculations are left as the very last step in evaluating a model, rather than an intermediate step to be corrected later. Moreover, machine learning methods explicitly focus on controlling for high dimensionality and model complexity as a primary point of interest.

In the broadest sense, machine learning is the automated search for models which better approximate objects of interest as larger and larger bodies of training data can be acquired. Such "objects of interest" are usually real-valued functions whose values are

known only for a training sample, but in general they may include pair-wise relations, distance metrics, trees, graphs, or vector-valued functions; classification (in which we want to predict only which of two or more groups, or classes, a new example belongs to), and regression (in which we want to predict an unknown real-valued quantity as a function of several observable variables), are settings of this type, and are of primary interest in this work, though future work will explore applications to more generalized settings.

Using machine learning methods, novel scientific hypotheses can be posed which go beyond traditional tests of separation-of-means statistics: for instance, we may want to test whether certain *predictions* can be made about disease progression at the *individual* level, and assign significance levels based on the quality of those predictions on held-aside data. Or, we may want to look for more subtle variations between groups by examining the classifier function returned by the algorithm, or by examining subpopulations in terms of their tendency to be misclassified (also known as novelty detection) – for which machine learning methods are well suited. Inversely, by posing neuroscience questions in machine learning settings we can identify capabilities in need of development, and refine our notions of what it means to learn from imaging data. Ultimately, insights gained in this process are potentially of significant interest to the machine learning community at large.

The principal concern of this thesis is to describe modelling and algorithmic advances based on machine learning methods which are specifically designed to facilitate understanding of the pathological processes underlying AD (and potentially other neurodegenerative disorders such as Fronto-Temporal Dementia) how they progress over time, and how we can make *patient-specific* prognostications. In addition, various ways of incorporating these developments into the design of clinical trials for novel treatments are proposed. The work described here can be divided into three main sub-themes:

1. Imbuing the learning process with structural biases based on the inherent spatial structure of neuroimaging data;

2. Integrating multiple data sources into a single classification model, with emphasis on multiple kernel methods;

3. Using machine learning methods to drive advances in clinical trial design.

Structural biases for learning

The first theme is to imbue the learning process with systematic and statistically sound learning biases based on the inherent structure of biomedical imaging data. That is, rather than shoe-horn a generic learning algorithm which is agnostic to the special characteristics of medical images, one might leverage knowledge relating to the acquisition, dimensionality, and spatial characteristics of such data to better address the clinical question under study. This allows a significant reduction in the search space of candidate classifiers, without sacrificing classification accuracy. In addition, it facilitates interpretation of the learned disease pattern. Simply put, if the algorithm knows something about the data from which it is tasked with learning a disease or other anatomical pattern, then it can do a better job. Models I have developed to address these issues are described in Chapter 3.

Multi-modality learning

The second major theme is to integrate multiple data sources into a single classification model. When multiple scanning modalities are available, each can offer a unique view or vantage point of brain structure and function. However, naïve methods of combining data sources will lead to statistical over-fitting issues, so it is necessary to develop novel algorithms that combine all the available information without losing anything in the process. I believe that combining kernels, either additively or non-linearly, is the most promising direction to search. Multi-Kernel Learning (MKL) is the most commonly used frame-work for learning linear combinations of kernels, and has been the basis for several novel algorithms. Chapters 4 and 5 will describe this work in detail. Chapter 6 describes results of several investigations into how machine learning methods can contribute to the science of AD by examining the performance of MKL with regard to its ability to combine imaging modalities for better predictions, and how those multi-modality predictions can be used as a screening measure in clinical trials to reduce cohort sizes.

Applications to clinical trial design

In Chapter 7 I describe some of our recent work that uses customized neuroimaging-derived outcome measures to increase sensitivity to treatment-related effects, and show that drastic reductions in the required sample sizes are realizable. This line of research has the potential to significantly increase the statistical sensitivity of clinical trials as well as reduce the number of patients required in order to demonstrate a drug or other treatment's effectiveness at combating AD pathology. The key observation is that the two steps of reducing high-dimensional data to a univariate measure of atrophy, and assigning a level of statistical confidence to observed differences in this univariate measure, can be done jointly, but without incurring statistical penalties.

I have also made preliminary investigations into a novel approach to sample size and power estimation in a common statistical setting that includes both clinical trials and neuroimaging studies. If we view a statistical parametric map (*e.g.*, a set of t-statistics, one at each voxel) as a sample from a low-rank stochastic process, then by decomposing the permutation matrix into a low-rank product with a sparse residual, we can analyze the entire phenomenon in terms of a small set of independent random variables. The aim is to apply these techniques to the independent components of this decomposition, and derive empirical estimates of Type-I and Type-II error rates without inducing multiple-testing issues, which are a scourge in high dimensional statistical parametric mapping studies. The approach I am investigating is to use low-rank random matrix methods to sub-sample a permutation test matrix, and re-create the remaining results. These investigations are presented in Chapter 8.

Organization

The progression of topics in this thesis is as follows: I begin with a discussion of relevant background concepts in Chapter 2. I then discuss learning methods which incorporate imaging-specific domain knowledge and inductive biases in single-modality settings in Chapter 3 before moving to a discussion of ways of combining multiple modalities in Chapter 4. I then consider a broader class of machine learning methods that can identify and exploit relations and interactions between input streams in Chapter 5. In Chapter 6 I examine applications of the machine learning approaches I have developed which may further drive neuroscience questions. In Chapter 7 I propose a novel

clinical trial design based on a machine learning test, discuss ways of analyzing it in terms of its predictive power, and present details of simulated clinical trials which strongly suggest that this method has the potential to boost the sensitivity of neuroimaging-based clinical trials. In Chapter 8 I describe ongoing work and open questions. First, I present preliminary results leading towards a novel methodology for performing experiment-wise Type I error calculations for clinical trials and other neuroimaging studies. Then, I discuss plans to validate the methods proposed in Chapter 7 by including them in a live clinical trial currently in the planning stages. Finally, in Chapter 9 I summarize the contributions of this thesis, and offer some concluding remarks.

Chapter 2

Background

Before discussing the major content of my research, I first present relevant background material to provide context.

2.1 Alzheimer's Disease

Several types of neurodegenerative dementia tend to occur in late life, of which Late Onset Alzheimer's Disease (LOAD)[1] is the most common. There are currently over 5.3 million sufferers of AD in the US, and the impact on the US economy is well over $100 billion annually [Wimo et al., 2006]. There are no known cures, but several treatments are under development which could potentially slow, or defer the effects of this disease. As Alzheimer's-related atrophy can precede loss of cognitive function by many years, and neural death is irreversible, these efforts require test subjects who are experiencing early stage AD in order to meaningfully evaluate the effectiveness of such treatments. The combination of medical imaging technologies and machine learning algorithms offer the potential to serve this need. Clearly, MRI and other scanning modalities are vital for detecting structural atrophy directly, while machine learning algorithms are well suited to extracting meaningful, predictive patterns which can *discriminate* between patients and healthy elderly subjects, even when signs of atrophy are subtle. In this section I will briefly outline what is known about Alzheimer's pathology, and how this knowledge impacts learning strategies aiming to predict its onset and progression.

The human brain can be segregated into regions of Gray Matter (GM), which is largely composed of neuron cell bodies and dendrites (incoming connections) as well as glial cells, which facilitate neural metabolic processes; White Matter (WM), which is composed of bundles of millions of axons, which transmit signals between

[1] The acronym LOAD is used to distinguish the late-onset variant of AD from familial, or early onset AD. Throughout this thesis, the term "AD" will be used to denote LOAD exclusively, as LOAD accounts for over 95% of all cases [alz, 2007].

disparate brain regions; and Cerebro-Spinal Fluid (CSF) which suspends the brain and helps absorb shocks. Gray matter tissues atrophy all through adult life. White matter, however, is known to gradually increase until early middle age, before showing signs of atrophy later. In late life, this process accelerates slightly, leading in some cases to deterioration or loss of cognitive function. Alzheimer's Disease (AD) is a neurodegenerative process in which GM atrophy (and to a lesser extent WM atrophy), is greatly accelerated, ultimately leading to complete debilitation.

While root causes of AD are still a matter of active investigation, some proximal causes are well known. At a histopathological (cellular) level, two distinct pathologies are present in AD: amyloid plaques, and neurofibrillary tangles. Amyloid plaques are masses of amyloid proteins which build up at the interfaces between neurons, and which can interfere with transmissions from axon to dendrite. For reasons which are not yet fully understood, these plaques are associated with a breakdown of subcellular structures, which ultimately leads to deterioration and death of the cell. One possible explanation is that the interference in neuron firing leads to atrophy through underuse, which is a naturally occurring mechanism in neurons. As neurons begin to deteriorate, the microtubules which carry electrical impulses begin to disintegrate, dispersing τ-proteins (microtubule building blocks) into the CSF, which can then be detected through protein assays. As microtubules disintegrate they lose local structural cohesion, and form neurofibrillary tangles which remain even after cell death, and are the second visible marker of AD pathology. If enough cells die in a certain localized region, then cortical GM will thin noticeably, and in some regions can atrophy away completely.

As these cellular degenerative processes proceed through several stages over many years, there are various ways of detecting them. Abnormal levels of Amyloid-β (Aβ) and τ proteins in CSF samples are among the earliest known signs of AD. However, this is complicated by the fact that many cognitively healthy elderly subjects also have abnormal levels of these proteins. As it is unknown how many, and which, among these subjects will progress to AD, it is difficult to make predictions solely on this basis. Several protective factors are known (*e.g.*, education, APOE ε_2 allele) which allow cognitive function in spite of a heavier pathological burden, which can also make such a determination difficult. An intermediate stage sign of AD pathology is depressed glucose metabolism in lateral parietal regions, and to a lesser extent, in para-hippocampal, and entorhinal cortices. Among these the Default Mode Network,

(a set of regions believed to be "active" when subjects are not involved in "tasks" such as those given in functional MRI scans, though they may also be involved in introspection and meta-cognition,) is strongly implicated, though the link between the functional role played by these regions and the biochemical processes leading to cell death is not yet entirely clear. Such reductions in glucose metabolization can precede detectable macro-scale atrophy, though cellular degeneration may be progressing undetected. As increasing numbers of neurons succumb to AD-type degenerative events GM volume can measurably decrease, which can be detected in MR scans. A challenge in detecting such atrophy is that the AD pattern of atrophy is similar, but not identical, to normal geriatric GM degeneration. In its final stages, atrophy overwhelms the *cognitive reserves*, *i.e.*, spare or redundant gray matter in certain vital regions, leading to dementia and debilitation.

A defining characteristic of AD is that such cell-death events are not distributed throughout the brain uniformly at random. Rather, there are distinct stages to the degeneration; these stages are characterized by the anatomical regions in which plaques and tangles, as well as associated GM atrophy, are found, as well as the overall severity of the histopathological burden [Braak and Braak, 1991]. This non-uniformity may be due to varying functional burdens (use / disuse) of different regions, or to differing physiological characteristics, or, to some unknown interaction between them. [2]

In the earliest stage at which cellular pathologies are detectable, atrophy is seen in the hippocampus and other basal nuclei, and in medial temporal regions. In later stages, atrophy spreads to pre-frontal cortical regions. While local atrophy may be very slight – on the order of 1% to 3% loss of hippocampal volume per year in AD subjects, and 0.5% per year in healthy older adults – cumulatively, atrophy is more visible in expanded CSF volumes such as the ventricles (fluid filled structures near the center of the brain) and sulci (clefts, or crevices between wrinkles in cortical surfaces.) By late-stage AD (post-onset of clinical dementia) the degeneration in the hippocampus is often quite striking – in some cases the hippocampus can almost disappear completely,

[2] For instance, Brodmann's areas are a set of roughly 50 (the list has undergone some revisions in the last 100 years,) sharply delineated brain regions where neuronal organization varies visibly; this variation was the original basis for Brodmann's segregation of regions. Some of these variations are known to be tied to functional characteristics of each region, *e.g.*sensory regions have more incoming connections, while motor regions have more outgoing, however not all variations can be fully characterized in this way.

and the ventricles can grow to occupy a significant cross-section of the brain.

It is important to note that while these pathologies are *necessary* for diagnosis as AD, they are not necessarily *sufficient*; many cognitively healthy older adults show such pathologies at autopsy. This means that a learning algorithm must do more than simply *detect* secondary signs of AD pathology, but rather it must *gauge* whether such signs are sufficiently severe that the subject is likely to be cognitively impaired.

Risk factors

Several genetic, health, and lifestyle risk factors are known to influence the probability of developing AD. First, of all is age – very few subjects under the age of 55 ever develop (late onset) AD. Second, various hereditary factors are known to be involved as well. Family history of AD among parents and their siblings is well known to increase the risk of AD [Johnson et al., 2006]. Recently, two major genetic influences have also been identified – the gene coding for apolipoprotein E, APOE is considered to be the primary genetic risk factor for AD, and has three major alleles. Of these ε_2 is somewhat rare, but has a slight protective effect; ε_3 is neutral, while ε_4 confers significant risk. TOMM40 (Translocase of Outer Mitochondrial Membrane) is a repeating sequence on chromosome 19, and is very near to APOE [Roses et al., 2009]. This gene varies by length, ranging from Short (< 20 thymidine bases) to Very Long (> 30), and while shorter forms are generally associated with healthy status, and longer forms with increased risk of AD, the interaction with APOE genotype is somewhat more complicated [Johnson et al., 2011]. Finally, several health and lifestyle factors can also contribute to the risk of developing AD. Vascular health, low body mass index (BMI), and low insulin resistance are all known to be associated with lower risk. Perhaps the most intriguing is that [Querbes et al., 2009, Butler et al., 1996] as well as other studies have shown that subjects with more education as a group have greater levels of atrophy, and in some cases can continue to have normal cognitive functioning in spite of levels of atrophy comparable to that of dementia patients. The exact mechanism for this is unknown.

Mild Cognitive Impairment

A milder form of dementia termed Mild Cognitive Impairment (MCI) has been documented in many cases which, while not meeting the diagnostic requirements of AD, often progresses to full AD. To be diagnosed as MCI, a subject must have a memory complaint, but no other serious cognitive dysfunction. Annually about 13% to 15% of MCI sufferers convert to AD on average, however some remain as stable MCI without ever converting. Thus MCI is usually, but not always, a precursor to AD. Because of this, there is particular interest in discriminating which MCI subjects will progress, and which will not, as any discriminative model which can do so can potentially highlight factors which differentiate AD from normal aging.

Implications for learning

As AD is very well studied, there is a wealth of domain knowledge pertaining to its characteristics, and there are several different means of detecting, and tracking its progress. We can therefore expect that machine learning methods which take full advantage of these resources are more likely to successfully uncover the disease patterns of interest. This domain knowledge can be broadly partitioned into several categories of primary driving influences, which will inform the majority of my research.

- **Imaging characteristics.** 3-dimensional images are needed to capture characteristics of the entire brain volume, which leads to very high-dimensional data. That is, while a 100×100 image might be considered thumbnail-sized, a $100 \times 100 \times 100$ image has 1,000,000 voxels. Thus, a driving factor in our choice of learning algorithms is the need to *control the dimensionality* of the subjects, and the corresponding model which classifies them. A large part of this job is handled by image registration, warping, and segmentation techniques which can separate out the GM regions of interest, discarding the rest. However, this typically leaves on the order of several hundred thousand voxels.

  A second characteristic of interest is that the machine learning algorithm sees each image not as a 3-D volume, but as a very long vector. Whereas at the time of acquisition we know which voxels are neighbors to one another, this information is lost when the images are treated as simple vectors. Since this information is

present in the data, it should be used to refine the space of classifiers to choose from.

- **Specific pattern of atrophy $\rightarrow$ brain regions.** AD is characterized by *specific* patterns of atrophy, owing to the physiological and functional segregation of the brain. Thus, if we can learn exactly where AD-specific atrophy occurs, we can ignore unrelated global atrophy to get a better signal. This is exactly the kind of task for which machine learning algorithms are designed. The implication for neuroscientific inquiry is that rather than using voxel-wise independent statistical maps which measure differences of means, and which are subject to multiple testing issues, we can train a single machine learning model to detect the pattern of interest, which may narrow the search for functional and anatomical correlates of AD.

- **Classification tasks.** In pursuit of questions related to AD, there are several classification tasks of interest:

  - **AD vs. Controls.** The most straight-forward way of studying AD is to try to classify images as being healthy or diseased. The resulting disease pattern represents late-stage AD.
  - **MCI vs. Controls and AD.** As MCI subjects are a more interesting group clinically, it is more interesting to study the effects seen only in MCI, and by implication, in an earlier stage of AD. Conversely we may also be interested in studying how AD progresses from MCI to full Alzheimer's disease. These tasks are more difficult, as the distinguishing signal is weaker.
  - **Stable vs. progressing MCI.** The most interesting, and the most challenging, task of all is to try to predict which MCI subjects will remain stable, and which will progress to AD.
  - **APOE genotype.** As APOE genotype is a major risk factor for AD, it may be interesting to examine its effect by classifying ε_4-allele subjects (higher risk) subjects from ε_2- and ε_3-subjects. By comparing a discriminative map for AD with one for APOE genotype, we may be able to identify residual factors in AD.

- **Inclusion of multiple modalities and biological measures.** As remarked on above, there are several imaging modalities, as well as CSF protein assays, genotype, cognitive measures, family history, and other biological measures which can influence a classifier's outcome. As is already the case with imaging data, dimensionality is a major issue, which is only compounded by the volume of data available. More interestingly, however, is that just as neighboring voxels are known to be interdependent, we can be certain that some entire imaging modalities will be related to one another than differently than they are to biological measures or genotype. Hence, machine learning methods which can capitalize on this information may be expected to perform those which do not.

Alzheimer's Disease Neuroimaging Initiative

Throughout this report, experimental validation of proposed methods has relied on copious subject data, which has been provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AlzheimerâŁ™s disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, M.D., VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research - approximately 200 cognitively normal older individuals to be followed for 3 years, 400 MCI patients

to be followed for 3 years, and 200 early AD patients to be followed for 2 years. Empirical results described in this thesis were performed using ADNI subject data almost exclusively.

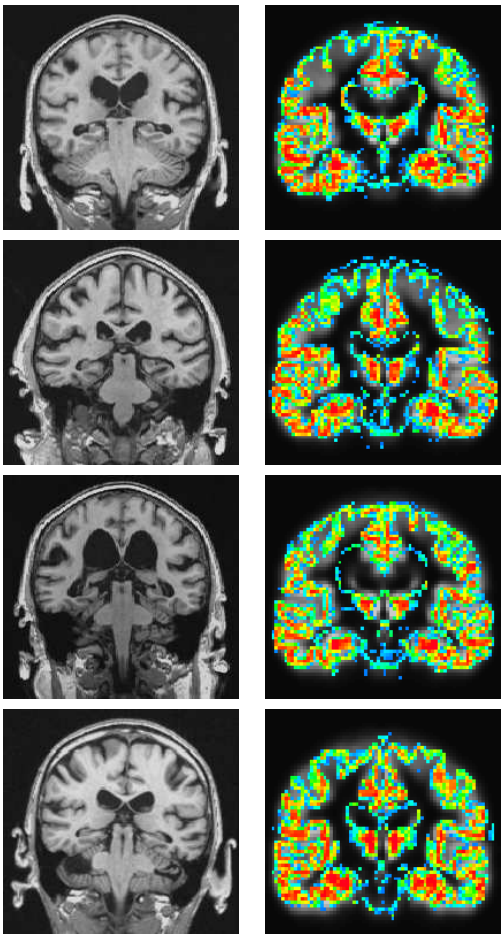


Figure 2.1: Several examples of anatomical variation among ADNI subjects. Left: original MRI scans. Right: Computed GM concentrations.

2.2 Medical Imaging Modalities

A large variety of medical imaging technologies are available which can identify signs of AD at different stages. In this section I will describe the types of relevant medical scanning technologies, and give an overview of the standard registration and normalization techniques. Special attention will be given to MR and FDG-PET imaging modalities, as they are the ones used in validation experiments described throughout, though I will briefly describe several others which can be readily adapted to the proposed learning frameworks.

Magnetic Resonance Imaging

Nuclear Magnetic Resonance is a well-studied physical phenomenon. Briefly, the protons and neutrons of every atom are constantly spinning. This spinning produces magnetic fields, which cancel each other when the axes of rotation are random. When placed in a strong enough magnetic field, however, the axes align themselves to the field. This field is called B_0. If other magnetic fields are applied, then laws of gyroscopic motion and magnetic fields dictate that a proton's axis of rotation will *itself* begin to rotate, or *precess*, (the common analogy is of a top, which continues to spin even as its major axis also wobbles). By applying a particular sinusoidally varying magnetic field, it is possible to encourage atoms to precess at a particular angle. It happens that a sinusoidally varying magnetic field is essentially a Radio Frequency (RF) transmission. Precession in the presence of the B_0 field produces a measurable RF signal, as the atoms shed the same RF energy that perturbed them in the first place. As hydrogen is the most prevalent type of atom in the body, and responds strongly to such RF perturbations, the frequency of the RF signal is chosen specifically to affect hydrogen atoms. Hence, to a first approximation, MR signal can be thought of as a measurement of proton density. For a more detailed treatment of the subject of MR imaging, see [Prince and Links, 2006].

T_1-weighted MRI

Once protons are made to precess by the RF signal, there are two factors which cause that signal to decay – the time it takes for the protons to revert to being aligned with

the main magnetic field, which is called the T_1 time, and the time in which it takes the spins of the various atoms to become *de-phased* with respect to one another, which also attenuates the RF signal, which is called the T_2 time. For reasons having to do with the density of protons, and the de-phasing effects of different substances, GM and WM have different T_1 and T_2 times. By varying the parameters of acquisition we can induce contrasts in measured signal relating to different T_1 and T_2 times in different tissues. As the T_1 times of GM and WM differ more than their T_2 times, a T_1 contrast is better for delineating the boundaries between the two tissue types, allowing a better segmentation and estimation of their relative volumes. Thus, T_1-weighted MRI is synonymous with structural MRI, as it is used for imaging anatomical structures. T_2-weighted MRI can be thought of as measuring differences in fluid concentration, which can be useful for diagnosing strokes, cancers and other types of brain injuries, but is less useful for measuring AD-related atrophy.

Diffusion weighted MRI

If a magnetic field is applied in a certain direction, and canceled a short time later, it is possible to measure the level of *diffusion* of water molecules, because for moving molecules, the two applications will not exactly cancel. Thus, moving water molecules will cause a measurable attenuation of the signal. If this process is repeated in many directions, we can measure the directional distribution of diffusion at each location in the brain. In anatomical terms, diffusion weighted MRI is best adapted for detecting axon bundles present in the WM, because water can diffuse along the length of axons, but not through their myelin coating, (myelin is a lipid which protects the axons, and promotes electrical conductivity). As neurons begin to die, so do their axons. Thus, it is conceivable that early signs of atrophy may be detected in the axon bundles emanating from regions most affected by AD pathology. Another possibility is that the mechanisms causing cell death in AD may also attack the protective myelin sheath.

Functional MRI

Functional MRI (fMRI) is based on the attenuation of MR signal relating to oxygenated blood. MR acquisition parameters may be set to maximize the sensitivity to this attenuation. This type of signal is called Blood Oxygenation Level Depletion (BOLD).

BOLD signal is largely driven by two opposing factors: first, as glucose is metabolized to produce ATP, oxygen is used up, contributing to BOLD signal. Often, the increase in metabolism is related to neural activity. However, brain vasculature is highly sensitive to such depletion, and in as little as two seconds, vessels may dilate so as to allow for more blood flow. This can actually reduce the BOLD signal. Recent studies have shown [Xu et al., 2009] that some of the early-stage functional abnormalities in AD may be detected in resting fMR (*i.e.*, fMR images acquired when the subject is at rest, and not performing any directed cognitive task).

Positron Emission Tomography

Another modality of imaging relies on radioactive tracers. Some radionuclides are known to emit an anti-electron, or positron. When this positron encounters an electron, the two particles annihilate, and produce two gamma rays which are oriented exactly 180° apart from one another. A special detector can then detect these two particles, and reconstruct their point of origin, which can be assembled into a global picture of where the radiotracer is concentrating. This type of imaging is known as Positron Emission Tomography, or PET imaging. An advantage of PET imaging is that a tracer can be tailored to bind to specific biological compounds of interest, giving a very specific picture of where such compounds are concentrated, without any unrelated signal from other biological structures or processes. However, PET suffers from a large impulse-response function, which reduces the effective resolution achievable, because the positron has to travel some distance – sometimes by as much as a centimeter – before annihilating. Various tracers are available, which have useful characteristics for measuring AD-related pathology.

- **Fluoro-DeoxyGlucose (FDG) PET**
  FDG-PET uses a form of glucose in which one O atom is replaced by an ^{18}F tracer, called Fluoro-deoxy glucose, or FDG. FDG-PET indirectly measures the level of glucose concentration in various tissues, and thereby the rate of metabolic and neural activity. This is extraordinarily useful for detecting early drops in activity in the Posterior Cingulate Cortex (PCC) and lateral parietal lobes. As these functional declines can precede atrophy, FDG-PET is an extremely useful marker of AD.

- **Pittsburgh Compound B (PiB) PET**
  Another type of radiotracer is the Pittsburgh Compound B (PiB) which is able to bind to amyloid proteins, thus revealing the location and density of amyloid plaques throughout the brain. However, as PiB is expensive to produce, and has a very short half-life, and because amyloid plaques are also found in cognitively healthy older adults, PiB imaging is somewhat less common. The ADNI does provide some PiB images, but only for a smaller subject cohort.

- **O^{15} Perfusion PET**
  Perfusion PET uses water molecules in which the O atom has been replaced by a ^{15}O radionuclide. Thus, perfusion imaging can show how well water, typically in the form of blood, perfuses the brain, and hence measures vascular health, as well as atrophy.

While ADNI only provides FDG-PET and T_1 (structural) MR scans, (and PIB-PET for a smaller group of subjects) the methods I have developed can be applied to other imaging modalities and degenerative disorders as well, so long as some aspect of pathology is measured. However, before using imaging data in a learning algorithm, we must first extract a feature representation so that standard learning frameworks can be employed.

Registration and normalization

Human subjects vary widely in the size, and shape, of their various anatomical features; several examples are shown in Figure 2.1. However, if we wish to provide a learning algorithm with a set of meaningful features, then point-wise anatomical correspondences must first be established. In other words, we must be able to find and delineate regions such as the hippocampus, entorhinal cortex, and lateral parietal lobules in all subjects in order to compare them. The simplest way to go about this is to rigidly or affinely warp the brain scans together, however, the greater the extent of the atrophy present, the greater the abnormality in the shape of the brain. Further, atrophy in smaller structures such as the hippocampus is sometimes most visible at the edges, meaning that even slight differences at the global level can result in these regions being mapped to completely different locations in different subjects. Experimental

evaluations described in this report which make use of subject brain images rely on two non-linear registration methods, which are part of a standard neuroimaging pre-processing pipeline.

Voxel Based Morphometry

Voxel Based Morphometry (VBM) [Ashburner and Friston, 2000] is a method of non-linearly warping volumetric data to a common template (called a stereotactic space). A non-linear warp is defined by a *deformation field*, which is represented as a piece-wise affine function which maps each voxel in the input image to a location in the target. This deformation field is chosen so as to minimize residual errors (in a sum of squares sense), with regularization which encourages spatial smoothness. The VBM approach assumes that large-scale, global individual variation is captured by the non-linear deformation (which is discarded) while local anatomical variations are reflected in differences from the template itself. The template itself is pre-segmented into GM and WM regions, so warping the volumes also produces a segmentation, or at an even finer-grained view, a relative map of GM concentration, (not to be confused with neuron density, as MRI does not have sufficient resolution to distinguish individual neurons).

Tensor Based Morphometry

An alternative approach is to use the deformation field itself as a measure of relative (*i.e.*, individual) variations. A numerical indication of relative volume is computed from the Jacobian determinant of the affine transform for each voxel. The Jacobian of an affine transformation is a tensor, and hence this method is called Tensor Based Morphometry (TBM). A single voxel in a TBM map represents the amount of growth or shrinkage the corresponding voxel in the source image was required to undergo in order to match the target. A particularly interesting use of TBM is in representing *longitudinal* data. When longitudinal data are available, we may be more interested in the change in volume over time rather than the baseline GM concentration. Thus, the second time point can be matched to the first to reveal the rate of expansion (of CSF for instance) or contraction. Then, if the baseline image has been registered to a

template using VBM, the TBM map can then be registered to the template using the same flow-field, giving a map of voxel expansion or contraction.

Both VBM and TBM can be followed by a Gaussian smoothing step, which is necessary for parametric statistical tests which assume normally distributed variables (convolution results in Central Limit behaviors) but need not be the case for discriminative learning methods. Once images have been registered, GM voxels may be selected relatively easily, to provide the learning algorithm with fixed-length feature vectors. VBM and TBM, as well as related morphometric methods such as HAMMER and RAVENS [Shen and Davatzikos, 2002, Davatzikos et al., 2001] are used extensively in scientific studies of neuroimaging data. A notable alternative is Region of Interest (ROI) analysis, in which a specific brain region is segmented, either manually or automatically, from which summary statistics are extracted. ROI analysis has the advantage that no registration is required (which may lose information if the registration does not exactly match the subject's anatomical features), but it may lose information through aggregation into summary statistics, and by ignoring whole-brain features. I will next discuss recent developments in the neuroscience literature which form the backdrop for my research.

2.3 Related Work in the AD Literature

Having discussed the characteristics of AD and relevant medical imaging technologies, I now turn to a discussion of related works, placing this work in an important context. A recent trend in the neurological sciences community is that a growing body of literature has been published reporting on applications of machine learning tools to the problem of developing markers of AD [Davatzikos et al., 2008b, Klöppel et al., 2008, Vemuri et al., 2008, Duchesne et al., 2008, Arimura et al., 2008, Zhang et al., 2011b]. These efforts have primarily utilized brain images, though some have also used other available biological and demographic indicators. Machine learning algorithms such as Support Vector Machines (SVMs), logistic regression, and ADABoost are among the most often utilized, often with some post-processing.

Studies comparing AD and control subjects

One of the earliest works in this area by Klöppel et al. [Klöppel et al., 2008] used linear SVMs to classify AD subjects from controls. In addition, they were also successful in separating AD cases from other types of dementia (Fronto-Temporal Lobar Degeneration or FTLD) using whole-brain images, highlighting the potential for machine learning applications beyond AD. The authors reported above 90% accuracy on autopsy-confirmed AD patients (vs. controls), and less where post-mortem diagnosis was unavailable. Independently, Vemuri et al. [Vemuri et al., 2008] showed promising evaluations on another dataset, obtaining $88 - 90\%$ classification accuracy (also using linear SVMs). The authors observed that using *all* image voxels as features within their framework was counter-productive, as many of these voxel-wise features had spurious correlations with the disease. To address these difficulties the authors added demographic and Apolipoprotein E genotype (APOE) data to their model as features, and adopted significant pre- (and post-) processing on the images. Feature selection was performed by training a linear SVM, and discarding negative-weight voxels, and then training a second linear SVM on the remaining voxels as the core learning algorithm. Having trained the classifier, the authors down-sampled the data to $22 \times 27 \times 22$ voxels, effectively aggregating many voxels' outputs into a single voxel at lower resolution. Then, they discarded voxels with less than 10% tissue densities in half or more of the images, and finally used an ROI to remove the cerebellum. In order to compensate for SVMs' inability to directly consider spatial relationships between voxels, they pruned the weights from the learned model by only retaining non-zero weights in a spatially contiguous $3 \times 3 \times 3$ neighborhood around top-ranked voxels.

In [Fan et al., 2008a,b, Davatzikos et al., 2008a,b], the authors implemented a classification / pattern recognition technique using structural MR images provided by the Baltimore Longitudinal Study of Aging (BLSA) dataset [Shock et al., 1984]. The proposed methodology was to first segment the images into different tissue types, and then perform a non-linear warp to a common template space using tools developed by their group. Feature selection was performed using standard statistical measures of (clinical) group differences, used to train a linear Support Vector Machine (SVM) [Cortes and Vapnik, 1995]. The reported accuracy was quite encouraging. More recently, the methods in [Fan et al., 2008b, Misra et al., 2008] have been applied to the

ADNI dataset, which utilized a large set of MR and FDG-PET images, giving accuracy measures similar to those reported in [Fan et al., 2008a,b, Davatzikos et al., 2008a,b].

Conversion of MCI to AD

Several recent studies [Schroeter et al., 2009, deToledo Morrell et al., 2004, Dickerson et al., 2001, Hua et al., 2008] have shown that certain markers are significantly associated with *conversion from MCI to AD*. In [deToledo Morrell et al., 2004, Dickerson et al., 2001], the authors show that traced volumes of the hippocampus and entorhinal cortex show significant group-level differences between converting and non-converting MCI subjects. In [Hua et al., 2008] a large number of ADNI subjects were tracked longitudinally using Tensor-Based Morphometry (TBM). The authors compared conversion from MCI to AD over 1 year with atrophy in various regions, but a discussion of the predictive potential was relatively limited (i.e., included p-values of 0.02 between converters and non-converters). Group-level differences such as these are a necessary first step in developing markers of AD, and are suggestive of features which may be useful for *predicting* onset of AD among MCI subjects in particular.

In [Davatzikos et al., 2009], the authors applied statistical learning techniques to both ADNI and BLSA subjects [Shock et al., 1984]. A classifier was trained using ADNI subjects, and applied to MCI and control subjects (in the BLSA cohort) to provide a SPARE-AD disease marker. This procedure could successfully separate MCI and control subjects with high confidence (AUC of 0.885), and it was demonstrated that the MCI group had a larger increase in SPARE-AD scores longitudinally. However, the main focus in [Davatzikos et al., 2009] was *not* on predicting which MCI subjects would progress to AD, but rather on finding a marker for MCI itself. In [Querbes et al., 2009], cortical thickness measures were used on a large set of ADNI subjects to characterize disease progression in AD and MCI subjects. Freely available tools (FreeSurfer) were used to calculate cortical thickness values at points on the surface of each subject's brain (after warping to Montreal Neurological Institute (MNI) template space) and then the thickness measures were agglomerated into 22 Regions of Interest (ROI), which the authors used as features in a logistic regression framework. Using age as another feature, a set of AD and control subjects were used to train a logistic regression classifier for each subject, yielding a Normalized Thickness Index (NTI). It

was found that this NTI was able to give 85% accuracy in separating AD subjects vs. controls, and had 73% accuracy (0.76 AUC) in predicting which MCI subjects would progress to full AD within 3 years.

Clinical trials

One of the motivating goals of the ADNI is to expedite the translation of AD research into clinically practical methodologies. Notably, much of the support for ADNI is from major pharmaceuticals. In [Hua et al., 2009] Hua and colleagues analyzed the statistical power of TBM in the context of measuring annual atrophy rates. More sensitive markers of atrophy rates can be used to increase the statistical power of clinical trials of treatments aimed at reducing such atrophy. Using TBM summary measures as a measure of atrophy, the authors were able to show that a hypothetical trial would need fewer subjects by over an order of magnitude than the alternative, which is to use cognitive measures used in AD diagnosis such as the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog). Summary measures were computed as the mean TBM-derived rate of change in a statistical ROI, which is a set of voxels chosen so as to achieve a desired level of Type I and Type II error. In [Kohannim et al., 2010] an SVM was trained on AD and control subjects using ROI volume measures and FDG-PET summary statistics, as well as APOE genotype, BMI and other demographic markers as features. Predictive scores were then calculated for AD and MCI subjects who had been held aside. By excluding subjects who were less likely to decline, they were able to improve the sensitivity of the TBM-derived outcome measure, yielding even smaller hypothetical cohort sizes. Such an approach is known as "sample enrichment".

Summary

The collective body of work described above demonstrates the potential of, and the level of interest in, applications of machine learning methods as statistical analysis tools in the study of AD. While several of these studies have proposed slight modifications to common learning algorithms by adding special pre- or post-processing steps, the core learning algorithms are left unchanged. My research has shown that even more gains are possible if the entire learning framework is adapted to take into account data

characteristics and available domain knowledge pertaining to the discriminative task being performed. As a significant portion of this domain knowledge can be expressed in terms of *dependencies* between features and modalities, a major focus will be on incorporating known dependency structures in learning frameworks in a way that is targeted specifically towards developing improved markers of AD. As I will describe next, kernel methods, which are based on the mathematical theory of Reproducing Kernel Hilbert Spaces are a convenient and natural way of approaching these issues.

2.4 Kernel Methods

Reproducing Kernel Hilbert Spaces (RKHS) have very well-developed theoretical underpinnings, and there is a long history of application of RKHS theory in various fields, including statistics, physics, mathematics and machine learning. As we are primarily interested in classification tasks, the application of RKHS ideas to Support Vector Machine (SVM) theory is the most relevant. In this section I give a brief coverage of RKHS theory, and its application to SVMs. Other kernel methods such as kernel density estimation and smoothing splines are also presented for comparison.

Reproducing Kernel Hilbert Space theory

A mathematical space is simply a set of points, typically having infinite cardinality. Several special types of spaces are defined: Banach spaces specify a *norm* on each point, $\|x\|$; Hilbert spaces also specify an *inner product* between all pairs of points $\langle x, x' \rangle$. Both types of space must also be closed, *i.e.*, they require that all Cauchy sequences of points in the space converge to a point in the space. The most prevalent example of such spaces is the n-dimensional vector space $\mathbb{R}^n$. Note that any Hilbert space is also a Banach space by setting $\|x\| = \langle x, x \rangle$.

The Aronszajn-Moore theorem establishes the equivalence between RKHSs and positive definite functions, making it a fundamental result in the study of both [Aronszajn, 1950]. A function $\mathcal{R}(x, x')$ is positive definite if for any finite sample of points $\mathcal{X}$, the matrix $R_{ij} = R(X_i x_j), x_i, x_j \in \mathcal{X}$ has all positive eigen-vectors, or equivalently, all of the eigen-functions of $\mathcal{R}$ have positive corresponding eigen-values. According to the Aronszajn-Moore theorem, each RKHS $\mathcal{H}$ must define a positive definite kernel

function $\mathcal{R}$, while any positive definite kernel function must *uniquely* define an RKHS. The "reproducing" part of the name is due to the following property [Wahba, 1990]:

$$\langle \mathcal{R}_s, \mathcal{R}_t \rangle = \langle \mathcal{R}(s, \cdot), \mathcal{R}(t, \cdot) \rangle = \mathcal{R}(s, t)$$

That is, by fixing one parameter and taking $\mathcal{R}$ over the entire set, we can generate a *representer of evaluation*, which must be a point in the RKHS itself. Thus, we can "reproduce" each point in the space through its kernel, and the inner product of any pair of points is the same as the inner product of their representers of evaluation. The importance of this fact cannot be overstated, because it means that if we have access to the kernel function, *then we need not have access to the data points themselves*. Thus, if we want to apply a kernel method to imaging data, we need only define a positive definite kernel function between pairs of subjects. In practical settings, the kernel function $\mathcal{R}$ is represented as a square symmetric kernel matrix K, (also known as a Gram matrix) and expresses some notion of similarity between the examples.

The characteristics of RKHSs drive the philosophy behind most kernel methods. First, every linear function f of the points in the RKHS is itself a point in the RKHS. This means that $f(x)$ is evaluated by taking $\langle f, \mathcal{R}_x \rangle$. A second consequence of this is that every such function f, as a point in the RKHS, has a norm. By choosing a function with the least possible norm, we have a way or restricting the class of functions being searched, which generally results in better models.

A second major result is known as the Representer theorem [Kimeldorf and Wahba, 1971, Schölkopf and Smola, 2002], which states that all functions in $\mathcal{H}$ which optimize some (monotonic) loss function can be expressed in terms of linear combinations of the examples. This can easily be seen by considering that any function of a set of examples decomposes into a function of those examples in the span of their representers, and a function of those examples in their orthogonal complement. The second part must be equal to 0, because it cannot improve the loss function, (but it does contribute to the norm on f), and the first part can be represented as a linear combination of representers because it must be a point in the span of the examples, evaluated in terms of inner products with the representers, $f(x) = \langle f, \mathcal{R}_x \rangle = \sum_i a_i \mathcal{R}(x, x_i)$.

As noted above, an important consideration when using imaging data is that the dimensionality of the data should not overwhelm the sample size. The following two

points allow kernel learning methods to address this issue:

- The effective dimensionality of a sample cannot be more than the number of points, because they must all lie within their own span. That is, the measurable portion of the representers of evaluation is restricted to the span of the data. Thus, this dimensionality can never be more than the number of examples (which is still undesirable – see second point). More importantly, the *rank* of the kernel matrix K is equal to the span of the data points, meaning that *a low-rank kernel matrix can restrict the complexity of a kernel model*.
- Well-known results show [Wahba, 1990, Cortes et al., 2010] that if the eigen-values decay at a certain rate, then even further restrictions on model complexity are possible.

These two points give two avenues for pursuing better generalization – first, we can seek low rank matrices, and second, we can try to engineer kernels to have decaying spectra (eigen-values). Moreover, if linear models are too constraining, or if we know for certain that they are not appropriate, we can look for a non-linear kernel function, and so long as it has an appropriate spectrum, *there is no sacrifice in model complexity*. The only extra parameters we incur are those of the non-linear kernel function, which are typically far fewer than for arbitrary non-linear functions.

Commonly used kernel functions

There are far too many kernel functions used in practice to list them all here. Instead, I will list the ones used in validation experiments described throughout. They are, Linear, Polynomial, and Gaussian kernels. A linear kernel function is simply the inner product of two examples in the original data space; for instance, unmodified SVMs implicitly use a linear kernel. A polynomial kernel function is one in which each entry of K is squared (or cubed etc.). Such kernels allow for polynomial functions, rather than simple linear models such as hyperplanes. Finally, Gaussian, or Radial Basis Function (RBF) kernels are based on the Euclidean distance between examples, by the formula

$$\exp\left(\frac{-\|\mathrm{x_i} - \mathrm{x_j}\|}{2\sigma}\right)$$

where σ is a bandwidth parameter and x_i and x_j may denote examples i and j. Note that this depends only on the *distance* between examples, and is thus radially symmetric, and insensitive to translation. The bandwidth parameter σ has an effect on the rank of the kernel in that if it is chosen too small, then examples are effectively orthogonal to each other, leading to a high rank kernel, while if it is chosen too large, then local structure can be washed out, and the resulting kernel's rank may be too low to aid in classification. A common solution to this issue is to choose σ equal to a small number (*i.e.*, in the range of one to two times the number of features, assuming the features themselves are normalized to roughly fall in the range of zero to one; this way, σ approximates the average distance between examples, which is believed to be the "right" scale in which to represent them. A broader treatment of commonly used kernel functions and kernel engineering methods is given in [Bishop, 2006].

A final note about normalization: when comparing kernels, it is often preferable for the geometric distribution of points in corresponding kernel spaces to be comparable. One way of ensuring this is to first center the examples about the origin, by subtracting row and column means and then adding back in the global mean, because $K_{ij} = \langle x - c, x' - c \rangle = \langle x, x' \rangle - \langle c, x' \rangle - \langle x, c \rangle + \langle c, c \rangle$, where c is the sample mean. Row, column and global means are represented by the last three terms. Having first centered the data, normalizing either the trace, or the maximum element of each matrix controls some measure of the spread of the data points in that space, allowing for a better comparison. Note that these methods only offer a first approximation of a solution to this problem. A full treatment should make a more principled definition of what is being compared, and how to normalize different kernels.

Kernel Support Vector Machines

Support Vector Machines (SVMs) are an extremely widely-used learning model based on the idea that a better classifier will result from maximizing the margin between different classes of objects. Thorough treatments of SVM theory are given in [Vapnik, 2000](ch. 5) and [Schölkopf and Smola, 2002, Cortes and Vapnik, 1995]. A key characteristic of SVMs is that they admit a kernel-based formulation. The SVM dual problem (5.7) is shown for reference:

$$\max_{\alpha} \sum_{i} \alpha_i - \sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{x_i^T x_j}_{\text{kernel}} \tag{2.1}$$
$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C \quad \forall i$$
$$\sum_{i} y_i \alpha_i = 0 \quad \forall i$$

Note that the examples (*i.e.*, AD / MCI cases, or controls) only occur as inner products $\langle x_i, x_j \rangle$. These inner products can be captured in a single $n \times n$ kernel matrix, $\mathcal{K}$. If we apply a non-linear function to $\mathcal{K}$ which preserves positive (semi-)definiteness, corresponding to a non-linear transformation of the data, then the learned classifier is a linear function (i.e., separating hyperplane) in the kernel space $\mathcal{H}$. Such a function typically maps back to a non-linear decision function in the original data space. In the SVM literature, the translation from the original data space to $\mathcal{H}$ is commonly denoted as $\phi(x)$; when the kernel function $\mathcal{R}$ is modified, the kernel space $\mathcal{H}$ and translation function $\phi(x)$ are correspondingly modified. In particular, Gaussian kernel-based SVMs can be thought of as training a Gaussian mixture model as the pattern classifier, and are therefore related to kernel density estimation of class probabilities, except that they lack constraints which would force them to give probabilities as outputs.

2.5 Multi-Kernel Methods

In applied kernel inference settings, it often happens that one has access to more than one kernel. If so, one must either select one kernel, and discard the rest, or, somehow combine the kernels in some way which preserves the information contained among them. One of the simplest, and most commonly used, ways of doing so is simply to add the kernels together in a linear combination. This is equivalent to taking the direct sum of kernel spaces, and makes the assumption that *all available kernel spaces are orthogonal* [Aronszajn, 1950, Wahba, 1990]. To see this, consider $\mathbb{R}^n$, which has an inner product defined as $\langle x, x' \rangle = \sum_i^n x_i x_i'$ where i iterates over any orthonormal basis. Orthogonal subspaces are therefore spanned by disjoint sets of basis elements, and inner products from each subspace therefore sum to give the inner product of the

whole. If two subspaces have any overlap, then their sum will be greater than that of the combined space. In this section I will describe several multi-kernel settings which are relevant to my research.

Multi-Kernel Learning

Multi-kernel learning (MKL) [Lanckriet et al., 2004, Sonnenburg et al., 2006, Rakotomamonjy et al., 2008, Gehler and Nowozin, 2009b, Mukherjee et al., 2010] is a generalization of SVMs to the multiple kernel case. This is achieved by adding a set of optimization variables called *subkernel weights* which are coefficients in a linear combination of kernels. If combining kernel matrices corresponds to concatenating feature spaces, then multiplying a kernel matrix by a constant merely scales the *axes* of that space. Therefore, we may interpret MKL as choosing a *scaling* for each RKHS, such that an SVM trained on their concatenation achieves the greatest margin possible. MKL maximizes the combined classifier's margin by minimizing the sum of the squared ℓ_2-norms of weight vectors $\|\mathbf{w}_m\|_2^2$ in each RKHS (indexed by m). This sum of squared norms represents the squared norm of the classification vector $\mathbf{w}$ in the combined space. In order to preserve convexity of the primal problem, rather than directly scale the kernels, MKL down-weights the norm penalty on $\|\mathbf{w}_m\|_2^2$ by the corresponding sub-kernel weight β_m [Kloft et al., 2011]. Thus, the regularization term $\frac{1}{2}\|\mathbf{w}\|_2^2$ in a one-kernel SVM becomes $\frac{1}{2}\sum_{m=1}^{M}\frac{\|\mathbf{w}_m\|_2^2}{\beta_m}$. Notice that as $\beta_m \to \infty$, the penalty on $\|\mathbf{w}_m\|_2^2$ approaches zero. To rule this out, and to control model complexity, a norm penalty on β ensures that the units in which the margin is measured are meaningful (as long as the base kernels are normalized).

The MKL primal problem is:

$$\begin{aligned} \min_{\mathbf{w},b,\beta\geqslant 0,\xi\geqslant 0} \quad & \frac{1}{2}\sum_{m}^{M}\frac{\|\mathbf{w}_m\|_2^2}{\beta_m} + C\sum_{i}^{n}\xi_i + \|\beta\|_p^2 \\ \text{s.t.} \quad & y_i\left(\sum_{m}^{M}\langle \mathbf{w}_m, \phi_m(\mathbf{x}_i)\rangle_{\mathcal{H}_m} + b\right) \geqslant 1-\xi_i, \end{aligned} \tag{2.2}$$

This problem is convex, and can be solved via block-wise coordinate descent [Rakotomamonjy et al., 2008]. Shogun (`http://shogun-toolbox.org`) is a

software package which implements several variants of MKL [Sonnenburg et al., 2006]. MKL experiments have made use of Shogun. Recently, much attention has been devoted to generalizing the norm on β to include arbitrary p-norms of the form $\|x\|_p = \left(\sum_i x_i^p\right)^{\frac{1}{p}}$ where $p \geqslant 1$ [Kloft et al., 2009, 2011, Orabona et al., 2010]. If $p = 1$, then the result is a well-known sparsifying effect on β, meaning that many kernels will be discarded. In cases where many kernels are known or expected to be uninformative, this is a desirable characteristic, however if this is not the case then we may wish to consider a less drastic approach. By varying p from 1 to 2, we can control the degree of sparsity, with $p = 2$ giving a non-sparse solution. Note that for $p = \infty$, the optimal solution has all of the weights being equal, and corresponds to an unweighted combination of kernels.

Related work

MKL has been a very active area of research, resulting in an array of models, as well as optimization strategies. The development of MKL methods began with [Lanckriet et al., 2004], which showed that the problem of learning the right kernel for an input problem instance could be formulated as a Semi-Definite Program (SDP). Subsequent papers have focused on designing more efficient optimization methods, which have enabled its applications to large-scale problem domains. To this end, the model in [Lanckriet et al., 2004] was shown to be solvable as a Second Order Cone Program [Bach et al., 2004], a Semi-Infinite Linear Program [Sonnenburg et al., 2006], and via gradient descent methods in the dual and primal [Rakotomamonjy et al., 2008, Orabona et al., 2010]. More recently, efforts have focused on generalizing MKL to arbitrary p-norm regularizers where $p > 1$ [Orabona et al., 2010, Kloft et al., 2009, 2011] while maintaining overall efficiency. In [Kloft et al., 2011], the authors briefly mentioned that more general norms may be possible, but this issue was not further examined. A non-linear "hyperkernel" method was proposed [Ong et al., 2005] which implicitly maps the kernels themselves to an implicit RKHS, however this method is computationally very demanding, (it has 4^{th} order interactions among training examples). The authors of [Mukherjee et al., 2010] proposed to first select the sub-kernel weights by minimizing an objective function derived from Normalized Cuts, and subsequently train an SVM on the combined kernel. In [Gehler and Nowozin, 2008, 2009b], a method was

proposed for selecting an optimal finite combination from an infinite parameter space of kernels. This method requires a line-search to be performed in the parameter space of kernels, and the resultant 'combined' kernel was shown to have a margin comparable to any infinite combination in that same space. Contemporary to these results, [Bach, 2008] showed that if a large number of kernels had a desirable shared structure (*e.g.*, followed directed acyclic dependencies), extensions of MKL could still be applied. In [Gönen and Alpaydin, 2008] an MKL model was proposed which varies the sub-kernel mixing weights locally throughout the kernel space, by adapting a set of "gating functions" which control the variation in weights. This method was shown to work well experimentally, and again in a thorough review paper [Gönen and Alpaydın, 2011]. Recently in [Gehler and Nowozin, 2009a], a set of base classifiers were first trained using each kernel and were then boosted to produce a strong multi-class classifier. At this time, MKL methods [Gehler and Nowozin, 2009a, Yang et al., 2009] provide some of the best known accuracy on image categorization datasets such as Caltech101/256 (see `www.robots.ox.ac.uk/~vgg/software/MKL/`).

Chapter 3

Adaptation of Learning Methods to Neuroimaging Problems: Single Modality Methods

---

As imaging data proliferates and study cohorts continue to grow, automation becomes essential – there are many stages in the image analysis pipeline, from acquisition, to modeling, to integration of results. For interesting sample cohort sizes it quickly becomes prohibitive for these tasks to be entirely manual. Moreover, it is desirable for these steps to be independent of any operator-dependent bias so as to facilitate comparison. Fortunately, much progress has been made on all of these fronts, resulting in a growing body of large-scale neuroscience studies. As mentioned in Section 2.3, advances have been made by incorporating machine learning methods into this process.

In order for a machine learning algorithm to do significantly more than memorize the training examples, it must make some assumptions about the structure of future observations. Well known results from learning statistical theory [Mitchell, 1997, Vapnik, 2000] show that inclusion of effective priors (introducing bias) to regularize the classification model is a vital component in any learning framework. Such assumptions are known as *inductive bias*. For example, a common inductive bias is to assume that the function being learned has a certain degree of smoothness, *i.e.*, that examples very near to one another will have the same target value. Another is the *max-margin* assumption that different classes or clusters will be separated by a relatively unpopulated (low density) region of space. In this chapter I will describe two models which address some of the challenges in applying machine learning methods to neuroimaging problems by developing specific inductive biases, expressed as regularization terms which are targeted towards learning from neuroimaging data. Both of these resulting models are based on the observation that the structure of imaging data can be represented as a set of relationships between features, as depicted in Figure 3.1 and these relationships should also hold among corresponding model parameters.

3.1 Motivation

In a neuroimaging setting, the learning task is to utilize training data (where confirmed or highly likely diagnosis of the patients into diseased or healthy *classes* is given) to learn a classifier to be used for disease diagnosis. If the data is in the form of images, the first step is to encode the image as a feature vector. Notice that an image volume of size $100 \times 100 \times 100$ in the training set yields a 10^6-dimensional vectorial representation. However, available image datasets are in general relatively small (with at most several hundred images) due to practical difficulties in volunteer recruitment and associated cost. As a result, the feature space is very sparsely populated, and the classification model may very easily overfit, leading to poor generalization [Mangasarian and Wild, 2004, Mitchell, 1997]. If some information about the data is given ahead of time (*e.g.*, the distribution is Gaussian), we may be able to effectively employ such knowledge by choosing a parametric model based on this assumption. Another common strategy to address the high dimensionality is to explicitly utilize dimensionality reduction tools such as Principal Components Analysis (PCA) [Jolliffe, 2002]. However, PCA estimates the spatial distribution of examples in a high-dimensional space (under linearity and Gaussian assumptions) rather than spatial information in the 3D coordinate system of the images themselves. Moreover, despite the fact that PCA operates in a high dimensional space, it is limited to the span of the training subjects, and so there is no way for PCA, or any other dimensionality reduction method, to discover all of the correlations in the data (which are potentially quadratic in the number of features – which are already quite numerous) simply by examining the distribution of the data only within the span of the available subjects. In the following, I will describe methods of encoding priors relating to the spatial structure of our data directly in the learning model.

3.2 Spatially Augmented LP Boosting

In [Hinrichs et al., 2009] my colleagues and I proposed a new framework for AD classification which makes use of Linear Program (LP) Boosting with novel additional regularization based on spatial "smoothness". The algorithm, called Spatially Augmented LP Boosting (SA-LPB), formalizes the expectation that since the examples

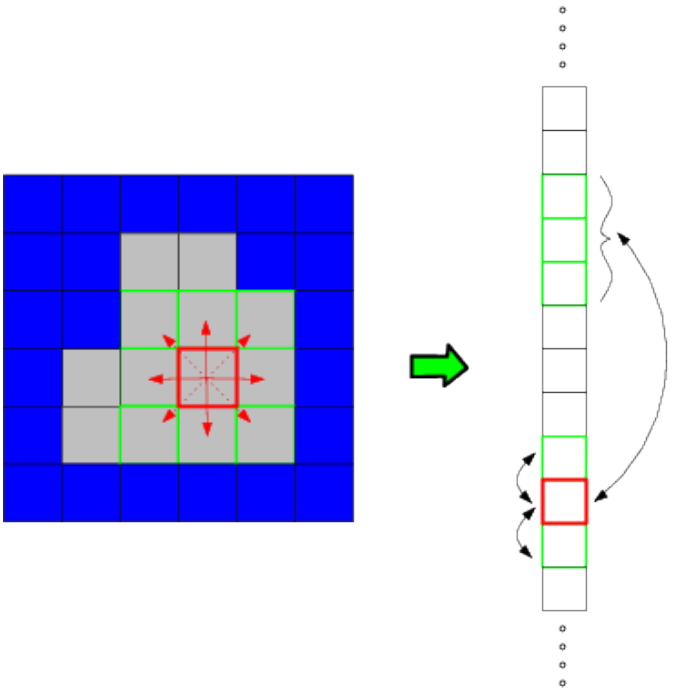


Figure 3.1: Spatial relationships between voxels which may be lost when transforming to a vector representation.

for training the classifier are images, the voxels eventually selected for specifying the decision boundary should constitute spatially contiguous chunks. In other words, a classifier composed of "regions" should be preferred over one composed of isolated voxels. This prior belief turns out to be useful for significantly reducing the space of possible classifiers and leads to substantial benefits in generalization. In this method, the requirement of spatial contiguity (of selected discriminating voxels) is incorporated within the optimization framework directly. Therefore, unlike some of the existing methods, post-processing of the optimized classifier to ensure spatial smoothness is not required. An initial attempt at modeling these interactions was proposed in [Singh et al., 2008] in the context of classifying autism. This model was significantly extended in [Hinrichs et al., 2009]. My contribution was to develop an efficient implementation, propose several modifications for classifying Alzheimer's Disease, and perform extensive evaluations of the algorithm on MR and FDG-PET images from

the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, which are detailed next.

Boosting Approach and Weak Classifiers

The SA-LPB classification method is built from the notion of boosting. Boosting seeks to "boost" the accuracy of weak (or base) classifiers – the general idea is to assign each classifier a weight in a way that will improve their aggregate response [Freund and Schapire, 1995, Mitchell, 1997, Schapire, 1990, Demiriz et al., 2002]. The weak classifiers, when considered individually, may have low predictive power. However, the underlying premise is that if the weak classifiers' errors are uncorrelated, their combination gives a better approximation of the underlying "signal". Linear Programming boosting (LP Boosting) is a boosting approach [Demiriz et al., 2002, Grove and Schuurmans, 1998] where the final classifier is learnt within a linear optimization framework but with a soft margin bias based on the hinge-loss function. The model places a 1-norm penalty on the weights, which also has the effect of reducing many of the weights to zero[1]. The SA-LPB model builds on the LP Boosting model by including additional priors. Weak classifiers in the neuroimaging case correspond to individual voxels (or features), which I describe in detail in this section.

Let us denote the set of images in the training set as $\mathcal{I} = \{\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_n\}$ with known class labels $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$, $y_i \in \{+1, -1\}$. Without loss of generality, AD-positive patients (and controls) are denoted as -1 (and $+1$) respectively, and $\mathcal{I} = \mathrm{I}_{\mathrm{AD}} \cup \mathrm{I}_{\mathrm{CN}}$ where I_{AD} (and I_{CN}) are the image sets of the registered AD (and control) groups. The set of image volumes in $\mathcal{I}$ are spatially normalized to a common template space, as a first step. Therefore, a voxel located at (x, y, z) in one image roughly corresponds to the voxel located at (x, y, z) in other images in $\mathcal{I}$.

The proposed method makes no assumptions on a specific imaging modality. For instance, when utilizing T_1-weighted MR scans, the images are segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), and probability maps of different tissue types are generated using standard techniques [Ashburner and Friston, 2000, Ashburner, 2007]. Either one of these quantities (voxel intensities) may

[1] In linear SVMs, the penalty is on the 2-norm of the weights, which places more emphasis on the *width* of the margin, in a *Euclidean* sense.

then be used to construct weak classifiers. Each weak classifier at a voxel $(\mathrm{x}, \mathrm{y}, \mathrm{z})$ tries to correlate variation at that voxel with the likelihood of AD diagnosis. Since AD is characterized by atrophy in specific brain regions, we should expect some weak classifiers to be more discriminative than others. The SA-LP Boost algorithm seeks to automatically select and boost such classifiers. For notational convenience, in the remainder of this section I will refer to voxels using a single index such as i, rather than $(\mathrm{x}, \mathrm{y}, \mathrm{z})$.

Suppose we have a list of the intensities of voxel i of all images in the training set, $\mathcal{I}$. Many functions of a single variable will give a reasonable classifier, but for boosting purposes we need only construct one having accuracy greater than random chance, and the most straight-forward way of doing so is to use the labels on the training data to determine a best-fit threshold.

The responses of the weak classifiers will populate a matrix, H, of size $\mathrm{m} \times \mathrm{n}$, where m is the number of images and n is the number of classifiers (or voxels). Motivated by the observation that the weak classifiers are not the most reliable predictors, I adopted a "soft" thresholding approach, *i.e.*, the response of the weak classifier assigns a confidence score to the classification for each image rather than explicitly classifying it in either group. I chose a logistic sigmoid function with a variable 'steepness' parameter ρ, and adjust the range to be $[-1, +1]$. To do so, one first chooses a voxel specific threshold, τ_{i}, so that the response is negative (or positive) if less than (or greater than) the threshold. The τ_{i} value is calculated as the midpoint between the voxel intensities' means at voxel i for the I_{AD} and I_{CN} groups. Because a decline in GM concentration or FDG-PET intensity is a sign of atrophy, a clinically consistent assumption here is that the control group mean, $\mu_{\mathrm{CN}}(\mathrm{i})$, is greater than the AD group mean, $\mu_{\mathrm{AD}}(\mathrm{i})$ [Fox and Schott, 2004]. My choice of an adjusted logistic sigmoid curve is based on the fact that its first derivative closely approximates the Gaussian distribution, because the value of the sigmoid (before adjustment) corresponds to the area under the Gaussian density function up to that point. This means that while the weak classifiers do not output actual probabilities, the level of confidence is related to the probability of class membership.

Let H_{ij} be the output of a weak classifier i (a certain voxel or feature) on image j.

$$H_{ij} = \frac{2}{1 + \exp\left(\tau_i - \rho \cdot I_j(i)\right)} - 1$$

where ρ is the "steepness" parameter, $I_j(i)$ is the GMP at voxel i in image $I_j \in I$, and the threshold is given as $\tau_i = (\mu_{CN}(i) - \mu_{AD}(i))/2$. The observed steepness as a function of ρ is shown in Figure 3.2. Having described the weak LP Boosting approach (and particular modifications necessary to adapt it to imaging data), I can now turn to the added improvements to the model itself.

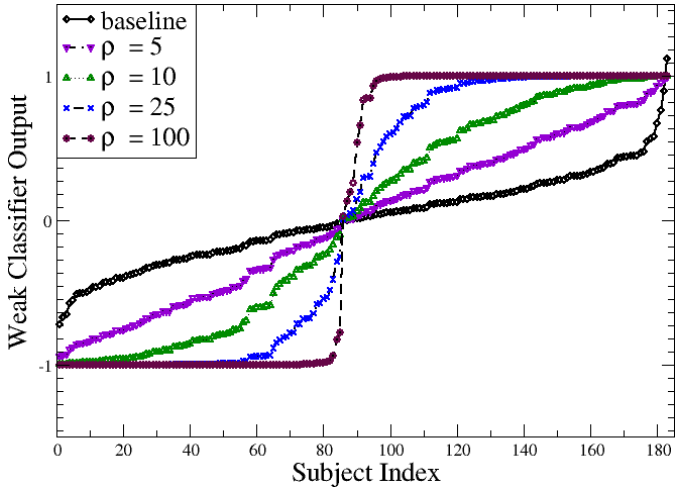


Figure 3.2: Weak classifier outputs as a function of various ρ values.

Spatial Constraints

As discussed at the opening of this chapter, a characteristic of the problem is that the feature vectors are representations of image data. This results in a certain dependency between the feature vector coordinates, and also the weak classifiers, (see Figure 3.1).

This property of the data can be leveraged to introduce a bias (or prior) in the classification which has an advantage of constraining the complexity (expressiveness or degree of freedom) of possible classifiers, encouraging better generalization.

The classifier consists of a set of weights on weak classifier outputs to define

a separating hyperplane. SA-LPB enforces spatial regularity by requiring that the weights assigned to neighboring weak classifiers should be similar. Such a spatial regularizer also has the benefit that it avoids selecting individual, spatially isolated voxels. Rather, it prefers spatially localized 'regions' – a desirable characteristic since isolated voxels are seldom clinically relevant; instead, markers of AD derived from patient imaging must be spatially localized.

Classification model

The SA-LPB optimization model is given as

$$\begin{aligned} \min_{\mathbf{w},\xi_i,t_{jk}} \quad & \mathbf{w}^T\tilde{p} + C\sum_i \xi_i + D\sum_{j\sim k} t_{jk} \\ \text{s.t.} \quad & y_i\mathbf{w}^T H_i + \xi_i \geqslant 1 \quad \forall\, i \\ & w_j - w_k - t_{jk} \leqslant 0 \quad \forall\, j\sim k \\ & w_k - w_j - t_{jk} \leqslant 0 \quad \forall\, j\sim k. \end{aligned} \tag{3.1}$$

As in SVMs, the vector $\mathbf{w}$ defines a classifying hyperplane, chosen to minimize hinge loss, with 1-norm regularization, which has the effect of selecting a *sparse* set of the most discriminative voxels. This allows for an easier clinical interpretation as the output consists of only a few highly discriminative (highly weighted) localized regions, and serves a feature selection purpose [Fung and Mangasarian, 2004] in many applications. The vector, $\tilde{p}$, represents the training set error rate of every weak classifier (the first term in the objective). This results in a *weighted 1-norm* regularization. This way, by adjusting the penalty on each weight w_j relative to its training set error rate, we allow weak classifiers with greater accuracy to be given slightly greater weight. The auxiliary variables, t_{jk}, represent the absolute difference between weights on neighboring voxels j and k (indicated as $i \sim j$). These variables are similarly penalized, which leads the optimizer to choose a separating hyperplane whose weights correspond to a set of spatially coherent voxels. Note that if $t = |w|$ then $t \geqslant w$ and $t \geqslant -w$ must both hold simultaneously. Thus, $t_{jk} = |w_j - w_k|$. The parameter C controls the amount of emphasis placed on *training set accuracy* relative to *model regularization*.

The emphasis on *spatial regularity* is similarly controlled by D. In Model (3.1) above, we observed that in practice $\mathrm{D} > 10\mathrm{C}$ is a reasonable choice to sufficiently enforce the neighborhood constraints.

The linear program in (3.1) can be *optimally* solved efficiently in polynomial time using standard solvers. Once the solution is obtained, the weights $\mathbf{w}$ can be interpreted as the coefficients of a separating hyperplane in the feature space. We use this hyperplane *directly* as our classifier, *and no additional post-processing is required*.

During experimental evaluations I observed that despite the 1-norm penalty, a feature selection step is still necessary. This is mainly for computational reasons, as well as to mitigate the possibility of over-fitting. In these experiments I used a very simple t-test on each voxel (using only training examples) and selected the top 2000–3000 most significant voxels, corresponding to roughly 1% of those available. However, more sophisticated methods can be utilized if desired and will likely further improve the empirical performance of the system. As a result of this feature selection step, I observed that some voxels selected had no neighbors that were also selected, meaning those voxels were not subject to spatial regularization, and the model therefore placed extra weight on them. Pruning such voxels resolved the issue.

Experiments and Results

I validated the SA-LPB algorithm using ADNI data, and present an analysis of its performance characteristics here. Classification experiments were performed using leave-two-out cross-validation. I chose this form of cross-validation because it requires fewer folds, while still behaving similarly to leave-one-out; the size of the training set in each fold is not much different from leave-one-out, but the number of folds is halved. In this section I will first cover results on experiments using T_1-weighted MR images, before moving to evaluations with FDG-PET image data in Section 3.2.

MR image data

In the first set of experiments I used only gray matter probability maps (GMPs, derived from VBM registration). The cross-validated classification accuracy of the model using GMPs was 82%, and the sensitivity (and specificity) was 85% (and 80%). In order to verify that the neighbor constraints are indeed having the desired effect I re-ran the

experiments with $\mathrm{D} = 0$, which effectively reduces the model to standard LP Boosting. In addition to causing a deterioration in accuracy, the number of non-zero voxel weights returned by the algorithm dropped by about a factor of 100, demonstrating the effect that the augmentation has on the algorithm. The results are summarized in Table 3.1, and suggest that the proposed technique works well for the AD classification task using MR image data.

Classifier outputs (*i.e.*, confidence levels) are shown in Figure 3.3. In Figure 3.3(a) we see that the classifier output on AD cases is concentrated between 0 (closest to the classification boundary) and -3 (farthest from the classification boundary), but the model incorrectly classifies some cases (which account for the misclassifications in the accuracy reported in Table 3.1 below). Classification confidence can also be used to generate Receiver Operating Characteristic (ROC) curves, in which the True Positive Rate (TPR) (sensitivity) is plotted as a function of the False Positive Rate (FPR), $(1 -$ specificity). Here, "positive" refers to AD subjects. The points in this plot are generated by setting different thresholds at which the classifier predicts that the subject has AD. That is, the confidence of every subject is used as a threshold, and all subjects with confidence higher than that threshold are classified as AD, and a TPR/FPR point is calculated from this, resulting in the curve shown in Figure 4(b). The area under the curve (AUC) of 0.8789 suggests a good predictive accuracy.

Data set	Accuracy	Sensitivity	Specificity	Area under ROC
GMP	82%	85%	80%	0.8789
FDG-PET	80%	78%	78%	0.8781

Table 3.1: Results of classification experiments on ADNI image data. One set of experiments were conducted with Gray Matter Probabilities (GMP) derived from T_1-weighted MR images as input. The other set of experiments were conducted with FDG-PET images.

An important component of our experiments was to evaluate the relative importance of various brain regions in terms of specifying a good classifier, and whether these regions are consistent with clinically accepted distribution of AD-specific pathology.

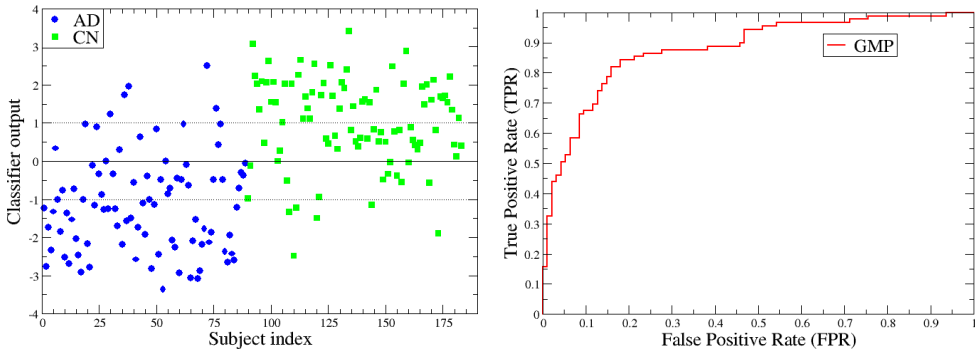


Figure 3.3: (a) Classifier's output for test images on the MR population. (b) ROC curves on the MR population.

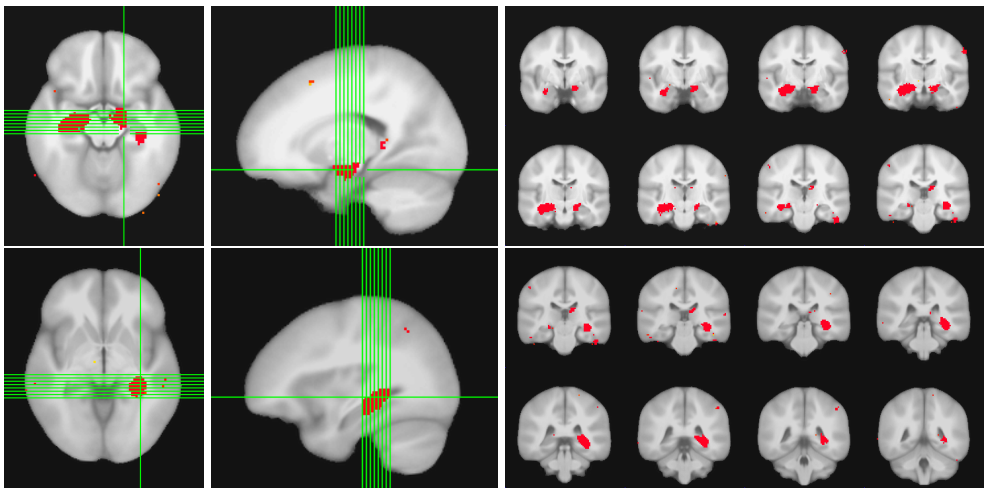


Figure 3.4: Brain regions selected when using GMPs derived from MR scans as input.

Figure 3.4 shows our results for the entire MR population. We can see that the selected voxels (or weak classifiers) are concentrated in the hippocampus and parahippocampal gyri, but that there are also some voxels in the medial temporal lobe bilaterally, and scattered in other regions. We find these results encouraging because the selected regions are all known to be affected in AD patients [Braak and Braak, 1991].

FDG-PET image data

I also applied the SA-LPB algorithm to the FDG-PET scans from the ADNI dataset. In all, there were 149 subjects in the MR population who also had FDG-PET scans. Hereafter I refer to this group as the FDG-PET population. SA-LPB obtained 80% classification accuracy on the FDG-PET population. The specificity was 78% and the sensitivity was 78% while the area under the ROC curve was 0.8781 as shown in Table 3.1. With the spatial constraints removed by setting the D parameter to 0, the number of non-zero weights dropped significantly as it did for GMP data; with the spatial constraints the algorithm typically chose between 150 and 500 non-zero voxels on FDG-PET data. Removing the spatial augmentation did not have a significant effect on accuracy. Because FDG-PET data is highly smooth to begin with, we do not expect as significant a gain in generalization performance by using spatial constraints. Because the level of accuracy was not significantly different, we do not present the results of these experiments.

Figure 3.5(a) shows the SA-LPB's output on the FDG-PET population. Similar to the MR population, most of the AD subjects are concentrated between -1 and -2 (and similarly the control subjects are concentrated between 1 and 2), while some subjects were misclassified. Again, the area under the ROC curve in Figure 3.5(b) is an indication of the accuracy of this method. Note that while most of the examples are clustered around their respective classes, some of the controls show the beginnings of decline, even to the point of resembling AD subjects. It is possible that some of these controls are showing early signs of AD-like symptoms prior to clinical dementia, however, one must also note that a group of AD subjects appears to be healthy (and another group is distinctly in the middle), which suggests that there are either some image processing artifacts, or that there is a small group of AD subjects having abnormal atrophy patterns not detected by SA-LPB. Section 6.2 further elaborates on these issues.

The brain regions selected by SA-LPB in the experiments utilizing FDG-PET scans also showed relevant brain regions. From Figure 3.6 we can see that the posterior cingulate cortex and bilateral parietal lobules are well represented, as well as the left inferior temporal lobe. These regions are known to have well established associations with AD-related neurophysiological changes, and no spurious regions were selected.

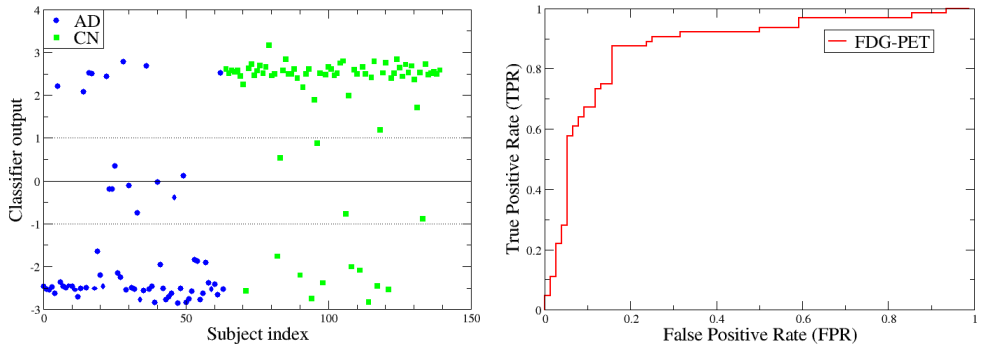


Figure 3.5: (a) Classifier's output for test images on the FDG-PET population. (b) ROC curves on the FDG-PET population.

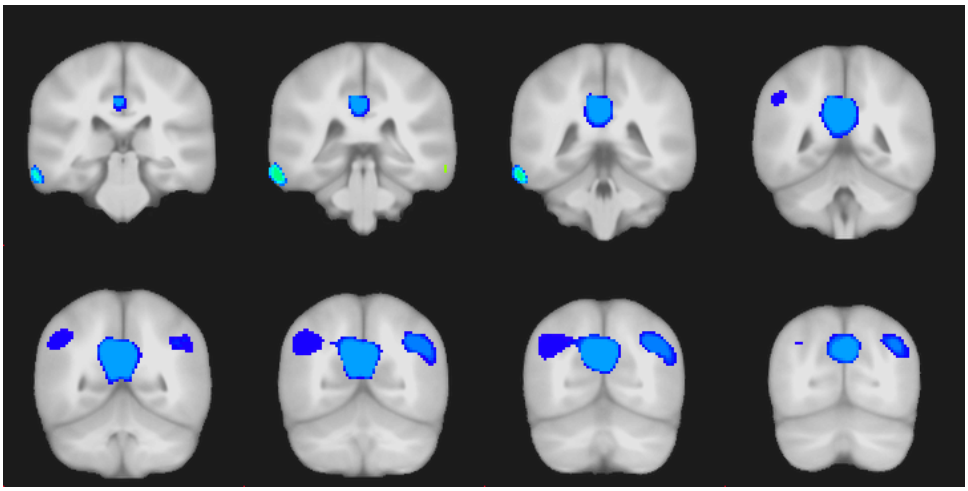


Figure 3.6: Brain regions selected when using FDG-PET scans as input. Lighter colors indicate greater weight.

These results illustrate that the SA-LPB algorithm is able to reliably determine clinically relevant regions in several different scanning modalities.

As a final observation about the SA-LPB model, note that the penalty on dissimilarity between neighboring weights is also a 1-norm penalty, with resulting sparsity behavior. Sparsity in this domain corresponds to many t_{jk} variables being equal to

0, which is the case when neighboring voxels have exactly equal weights. Thus, in many situations, the optimal classifier consists of a set of contiguous sets of voxels, all having uniform weights. While the spatial smoothness-inducing side-constraints do combat overly sparsifying tendencies of the 1-norm regularizer, there is nevertheless a noticeable "shrinking bias" similar to that described in the graph-cuts based segmentation literature [Kolmogorov and Boykov, 2005, Vicente et al., 2008][2]. These issues were a part of the motivation behind **Q**-SVM, described in the next section.

3.3 Q-SVM

SA-LPB improves upon LP-Boosting by incorporating side-constraints which require that neighboring voxels should have similar weights, and that any divergence from this ideal must be due to sampling artifacts, and should be discouraged. The way this belief is encoded, however, is in terms of a 1-norm penalty on differences only between direct neighbors. This encourages a kind of sparsity behavior which is somewhat drastic, leading to "flat" regions having exactly the same value, which may be overly constricting. Further, note that this is a penalty on the surface of transition between classifier regions and non-classifier regions, which means that it is occasionally necessary to carefully tune the C and D parameters in order to get reasonable regions in the output classifier. This also means that two new constraints must be added to the LP for each *pair* of voxels, which can lead to some difficulties with scalability – for more than 10,000 to 20,000 voxels, the associated running time created difficulties.

Perhaps a preferable approach might consider pair-wise feature relationships in terms of a *covariance structure*, which can then be regularized as a quadratic function $\mathbf{w}^\mathsf{T}\mathbf{Q}\mathbf{w}$, where $\mathbf{Q}^{-1}$ encodes the covariance between features. (In order to address scalability, note that this covariance structure can be encoded as a sparse matrix.) As a motivating example, consider handwritten digit recognition. (See experiments below.) When examples correspond to images, features may correspond to pixels – there is a natural correlational structure inherent in the data, as neighboring pixels are expected

[2] We can think of SA-LPB as simultaneously solving a classification task and a segmentation task, in that we would like to examine the optimized classifier in order to extract relevant discriminative "foreground" regions.

to be highly correlated. By encoding this relation in $\mathbf{Q}$, we can bias the classifier to choose a smoothly varying pattern of pixels, which is more likely to correspond to the true separation between digit classes.

This has several advantages. First, the strength of the connection between voxels can be specified as well, which can provide an enhanced range of priors to impose on the model. Second, as a sparse matrix, $\mathbf{Q}$ can encode a larger number of interactions in a way that scales much better than adding constraints to an LP. Third, there is less of a concern about discarding so many voxels as in SA-LPB under the sparsity-inducing 1-norm regularization on $\mathbf{w}$. Instead, the sparsity domain is in the *eigenvectors* of $\mathbf{Q}$ – sparsity in this case is driven more by the decaying eigenvalues of $\mathbf{Q^{-1}}$. That is, the eigenvalues of $\mathbf{Q}$ are just the inverses of those of $\mathbf{Q^{-1}}$, meaning that a few eigenvalues will be penalized much less than the rest.

The $\mathbf{Q}$-SVM model is given as,

$$\min_{\mathbf{w},\xi\geqslant 0,b} \quad \mathbf{w}^{\mathsf{T}}\mathbf{Q}\mathbf{w} + C\sum_{i} \xi_i \tag{3.2}$$

$$\text{s.t. } y_i\left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b\right) \geqslant 1 - \xi_i \tag{3.3}$$

Note that $\mathbf{Q}$-SVM generalizes the regular SVM (by setting $\mathbf{Q} = I$). $\mathbf{Q}$-SVM improves over SA-LPB by eliminating the overly sparsity-inducing behavior of SA-LPB's neighbor constraints; it provides this capability by encoding a non-sparsifying regularizer, not on the *differences* of the weights, but on their *products*. That is, rather than penalizing differences between neighboring voxel weights, $\mathbf{Q}$-SVM can encourage weights to be similar by rewarding a high product between neighboring weights (by placing a negative entry in $\mathbf{Q}_{i,j}$) while still penalizing their individual squared weights (since $\mathbf{Q}_{i,i}$ must be positive and since $\mathbf{Q}$ must be diagonally dominant). For instance, if we were to calculate $\mathbf{Q}$ as a Gaussian kernel between the (x, y, z) coordinates of each voxel, and take a graph Laplacian, then there will be a higher reward for voxels that are close to one another having more similar weight, than for voxels that are far apart, but without the sparsity inducing behaviors of SA-LPB. The comparison is highlighted in experiments with $\mathbf{Q}$-SVM on ADNI data, described below.

The $\mathbf{Q}$-SVM model subsumes several other models that have been proposed in the machine learning literature recently. The model of [Xiang et al., 2009] is similar in concept to (3.2) except that their formulation is an adaptation of ADABoost (for fMRI voxels), while $\mathbf{Q}$-SVM is a direct extension of the SVM model. A recently proposed model for natural language processing [Bergsma et al., 2010] is also a special case of (3.2), in which we set $\mathbf{Q} = \mathbb{I}^{M \times M} - pp^T$, where $p_i = \frac{1}{M},\ \forall i$, (or some other distribution if it can be specified through domain knowledge). One can interpret this model as adding a 1-norm *reward* on top of the 2-norm regularizer, which has the effect of making the weights more uniform. The authors of [Cuingnet et al., 2010] proposed a regularization scheme in which the classifier $\mathbf{w}$ is transformed according to $\mathbf{w} \rightarrow e^{\frac{1}{2}L}\mathbf{w}$, where L is a graph Laplacian (or Laplace-Beltrami operator) based on distances between voxel-wise features, and exponentiation is applied element-wise. We can see that the norm on $\mathbf{w}$ becomes $\mathbf{w}^T(L^TL)\mathbf{w}$ which is a case of $\mathbf{Q}$-SVM. As noted, it is almost always beneficial to regularize a model so as to better reflect the correlations present in the data, especially when this allows us to bring in extra information that goes beyond what we can infer from the data (*i.e.*, out-of-band information).

Classification experiments

As an illustrative example, consider hand-written digit recognition from the MNIST digits dataset [LeCun et al., 1998]. When examples correspond to images, and features correspond to pixels, there is a natural correlational structure inherent in the data, as neighboring pixels are expected to be highly correlated. Ordinarily, the relation between neighboring pixels is lost when converting images to feature vectors – however, by encoding this relation in $\mathbf{Q}$, we can bias the SVM classifier to choose a smoothly varying pattern of pixels, which is more likely to correspond to the true separation between digit classes. Three such $\mathbf{Q}$-matrices are shown in the top row of Figure 3.7. The first is the covariance function between pixels, which can be thought of as an indication for a given pair of pixels of how likely they are to both be included in a given character. This is similar to a spatial smoothness prior, except that it also encodes directionality. The second is a Gaussian kernel between 2D pixel coordinates, and the third is their element-wise product (which is also p.s.d.). I then trained a $\mathbf{Q}$-SVM classifier to distinguish between '3' and '8' digits. The resulting weight

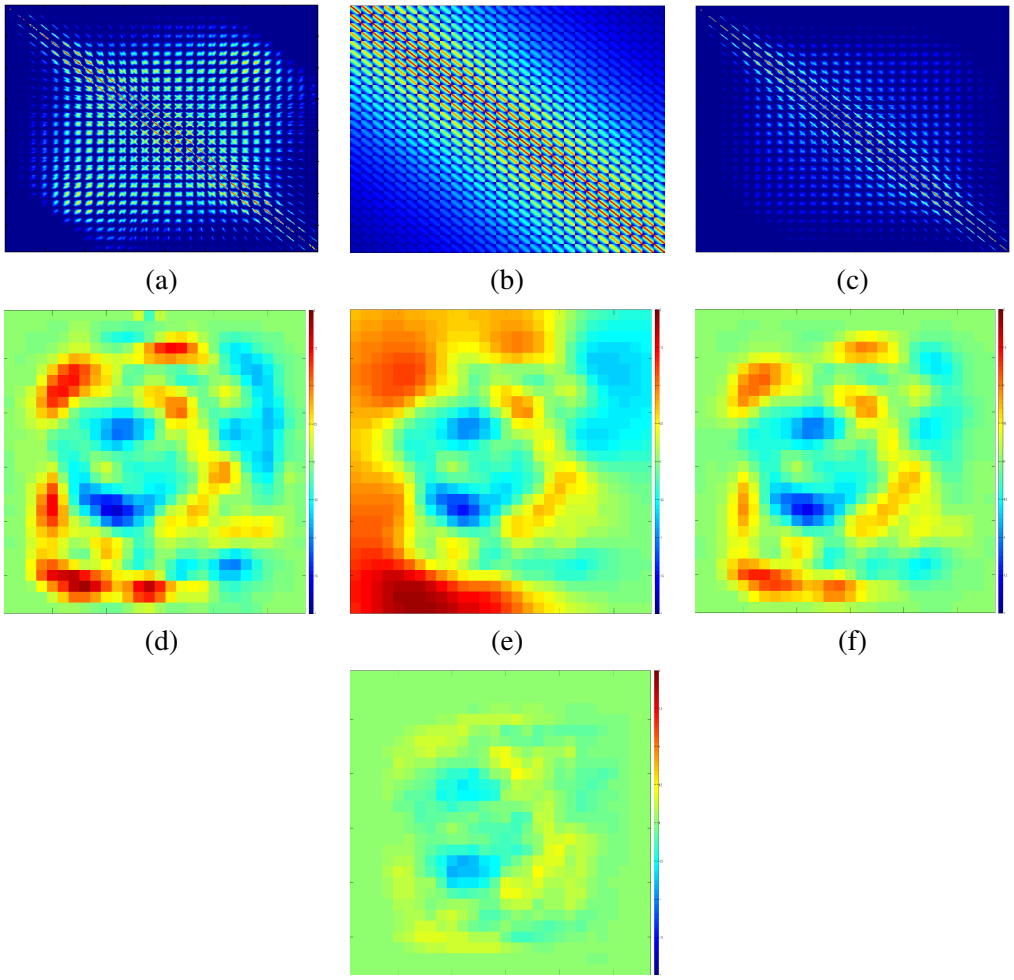
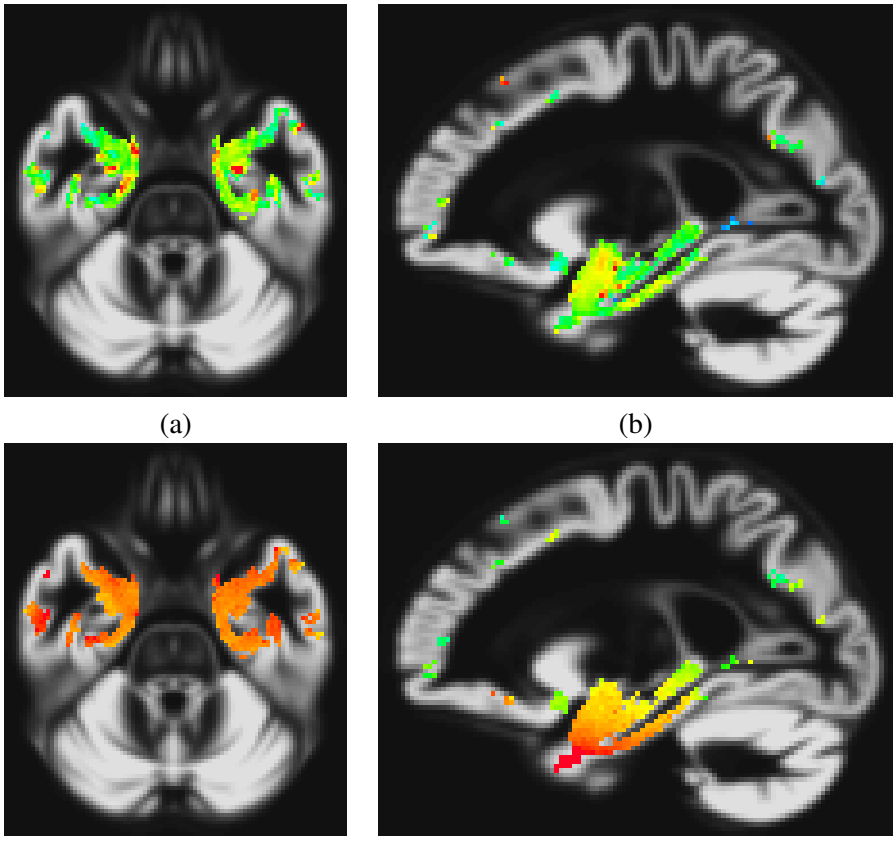


Figure 3.7: Top Row: Three **Q**-matrices for pixel-wise features in the MNIST dataset. (Each is 784×784.) (a): pixel-wise intensity covariances over the entire MNIST data set; (b): Gaussian kernel between pixel coordinates; (c): element-wise product of (a) and (b). Middle Row: (d – f) Pseudo-Inverses of the above **Q**-matrices (a – c) were used as **Q**-SVM regularizers, and the resulting weight vectors are depicted here as 28×28 images. Bottom Row: (g) Weight vector trained by a standard SVM. Color scale in all images goes from $-2 \times 10^{-3} - -2 \times 10^{-3}$. All **Q**-matrices were normalized to have trace equal to the number of rows, to facilitate comparison with the Identity matrix (*i.e.*, standard SVM).



(a) (b)

(c) (d)

Figure 3.8: Comparison of spatial smoothness of the weights chosen by **Q**-SVM and SVM with computed gray matter (GM) density maps. Left (a-b): classifier weights given by a standard SVM; Right (c-d): classifier weights given by **Q**-SVM .

vectors are shown in the second row of Figure 3.7. For comparison, the weight vector chosen by a standard SVM is shown in the far right column. It is clear from the figure that the corresponding regularizers differ, and they lead to different classifiers.

To demonstrate the influence on the learned classifier in a neuroimaging context, I performed classification experiments with the Laplacian of the inverse distance between voxels as a **Q** matrix, and voxel-wise GM density (VBM) as features. Using 10-fold cross-validation with 10 realizations, **Q**-SVM 's accuracy was 0.819, compared to the regular SVM's accuracy of 0.792. These accuracies are significantly different at

the $\alpha = 0.0005$ level under a paired t-test. In Figure 3.8 a comparison is shown of weights trained by a regular SVM (a–b), and those trained by a spatially regularized $\mathbf{Q}$-SVM (c–d). Note the greater spatial smoothness and lower influence given to isolated "pockets". Refinement of ideas relating to $\mathbf{Q}$-SVM ultimately led to the $\mathbf{Q}$-MKL model, discussed in Chapter 5.

Chapter 4

Adaptation of Learning Methods to Neuroimaging Problems: Multi-modality Methods

As mentioned in Section 2.3, several recent papers have demonstrated that discrimination of AD subjects from controls is possible with MR or PET images using machine learning methods such as SVM and boosting. These algorithms learn the classifier using one *type* of image data, yet AD is not well characterized by one imaging modality alone, and analysis is typically performed using several image types, each measuring a different type of structural/functional characteristic. In this chapter I explore AD classification using multiple modalities *simultaneously*. The difficulty with this approach is that any issues with dimensionality are compounded by the number of imaging modalities introduced, meaning that extra means of controlling model complexity are required. To tackle this problem, we utilize and adapt a recently developed class of machine learning tools called Multi-Kernel learning (MKL). Essentially, each imaging modality spawns one (or more) kernels and we simultaneously solve for the kernel weights and a maximum margin classifier. In the immediately following section I will discuss investigations into the utility of MKL for boosting discriminative power from multiple imaging modalities, as well as other sources of information such as genotype, CSF protein assays or demographics, before moving to a discussion of how outlier robustness can improve classification accuracy in MKL.

4.1 Examination of p-norms

When applying MKL to AD classification, one has the option of choosing between several norm regularizers on the subkernel weight vector β. The 1-norm is the sparsest, *i.e.*, it gives solutions in which most of the subkernel weights drop to 0, effectively discarding those kernels, while the 2-norm is non-sparse, meaning that all kernels will be included in the final model, though some may have very small weights. For norms between 1 and 2, the solution is of intermediate sparsity. By choosing a norm, we are

essentially making a guess as to what percentage of the kernels are essentially "noise", and should be discarded. The question then is, are there any "useless" kernels that should be discarded? On the one hand, all kernels are derived from imaging data, and should in theory contain useful information, but on the other hand there is redundancy due to overlap between feature sets and using different kernel functions with the same data. Results are shown in Table 4.1. Briefly, norms greater than 1.5 performed about equally, and were superior to the 1-norm, suggesting that too much sparsity is not always a good thing.

For these experiments, 48 AD subjects and 66 controls were chosen who had both T_1-weighted MR scans and Fluoro-Deoxy-Glucose PET (FDG-PET) scans at two time-points two years apart. Standard diffeomorphic methods (SPM, `www.fil.ion.ucl.ac.uk/spm/`) were used to register scans to a common template and calculate Gray Matter (GM) densities at each voxel in the MR scans. We also used Tensor-Based Morphometry (TBM) to calculate maps of longitudinal voxel-wise expansion or contraction over a two year period. Feature selection was performed separately in each set of images by sorting voxels by t-statistic (calculated using training data) and choosing the highest 2000, 5000, 10000,…,250000 voxels in 8 stages. I used linear, quadratic, and Gaussian kernels: a total of 24 kernels per set, (GM density maps, TBM maps, baseline FDG-PET, FDG-PET at 2-year follow up) for a total of 96 kernels. Note that the same experimental setup (including subjects and kernels) was used in work described in Chapter 5, and in Section 6.3.

MKL norm used	Accuracy	Sensitivity	Specificity	Area under ROC
1.0	0.914	0.867	0.949	0.977
1.25	0.916	0.865	0.954	0.980
1.5	0.921	0.874	0.956	0.982
1.75	0.923	0.872	0.961	0.982
2.0	0.922	0.870	0.959	0.981
SVM	0.882	0.844	0.910	0.970

Table 4.1: Comparison of different MKL norms in the presence of uninformative kernels, and an SVM trained on a concatenation of all features for comparison.

4.2 Robustness to Outliers

In this work, my colleagues and I studied the problem of AD classification using *multi-modal* image data, as above. To make the MKL model robust, I developed strategies to suppress the influence of a small subset of outliers on the classifier, on a *per-kernel* basis, giving a variant on MKL which I call Robust MKL. Though the primary model is not convex, I developed an alternative minimization-based algorithm for Robust MKL. To evaluate Robust MKL's efficacy, I performed *multi-modal* classification experiments on images from the ADNI project, with promising results. In this section I will motivate the robustness modification, discuss the related optimization issues, and present experimental results.

Outlier Ablation

As discussed above, MKL finds an optimal regularized linear combination of kernels concurrently with the optimal classifier. In addition to solving for the combination of kernels, however, it may be desirable to identify and suppress the influence of one or more mislabeled subjects (examples) may have on the classifier. This is important in AD classification because of: (1) Co-morbidity: In some cases, AD is coincident with other neurodegenerative diseases such as Lewy bodies or FTLD, and because subjects may have varying degrees of cognitive reserve due to protective effects such as education [Querbes et al., 2009]; and (2) while the image data may suggest signs of pathology characteristic of AD, these usually *precede* cognitive decline. As a result, a subject may be cognitively normal (and labeled as control) in spite of early-stage AD pathology or because the subject has greater cognitive reserve. To ensure that the algorithm is robust for this problem and other applications, we would like to identify such outliers within the model. Note that this does *not* mean that the classifier will be able to identify such subjects in the future – rather, this robustness will allow the classifier to focus on the common cases, without attempting to correctly classify outlier subjects. This way, we can expect unseen (test-set) outlier subjects will appear even more strikingly as outliers, making it easier to recognize them. In order to do this, one option within the SVM setting is to replace the regular loss function with the "robust" hinge loss function which differs only in that it is capped at 1:

$$\text{robust-hinge}(w, x, y) = \min(1, (1 - y w^T x)_+), \tag{4.1}$$

where $y_i \in \{+1, -1\}$ are the class labels. This means that once an example falls on the wrong side of the classifier there is no additional increase in penalty. To address the non-convexity of Equation (4.1), Xu et al. [2006] replaced the usual hinge loss function with the $\eta-$hinge loss function, which uses a discount variable η_i for each example. That is,

$$\begin{aligned} \eta\text{-hinge}(\mathbf{w}, x, y) &= \eta(1 - y \mathbf{w}^T x)_+ + (1 - \eta) \\ 0 &\leqslant \eta \leqslant 1 \end{aligned} \tag{4.2}$$

where $(\cdot)_+$ truncates negative values to 0. The result in [Xu et al., 2006] shows that η-hinge loss has the same optimum and value as robust-hinge loss. Our proposed model makes use of such a parameter to serve as both an outlier indicator and also to adjust the influence of this example on the classifier in the MKL setting. Robust MKL is formulated as,

$$\begin{aligned} \min_{\eta} \min_{\mathbf{w}, \xi} \quad & \textstyle\sum_k \|\mathbf{w}_k\|^2 + C \sum_i \xi_i - D \sum_{i,k} \eta_{i,k} \\ \text{s.t.} \quad & y_i(\textstyle\sum_k \eta_{i,k} \mathbf{w}_k{}^T \phi_k(\mathbf{x}_i)) \geqslant 1 - \xi_i \ \forall i \\ & 0 \leqslant \eta_{i,k} \leqslant 1 \ \forall i, k \\ & \xi_i \geqslant 0 \ \forall i. \end{aligned} \tag{4.3}$$

Here, $\mathbf{w_k}$ is the set of weights for the kernel k, ξ_i is the slack for example i (similar to SVMs), and $\eta_{i,k}$ is the discount on example i's influence on training classification in kernel k (described in detail below).

The role of η in Model 4.3 can be understood by examining its effect on the loss function: $\eta_{i,k}$ introduces a discount for $\mathbf{x}_i$'s contribution to the classifier *only in kernel* k. This means that an example which is badly characterized in some kernels can *still be used* effectively in other kernels where it is more accurately characterized. In this

way, the proposed model performs *automated* outlier identification *and* suppression in the MKL setting. This is balanced by the *positive reward* for making η as large as possible (objective term $-D\sum_{i,k}\eta_{i,k}$). Note that this term can be equivalently expressed as $+D\|1-\eta\|_1$: a 1-norm penalty on $1-\eta$. The corresponding sparsity behavior ensures that for most (i,k), $\eta_{i,k}=1$, meaning that *the least number of discounts possible are given out through* η. This is important because if we allow the algorithm to discard any example that contributes to the loss function, there can be no guarantee on learning or generalizability. Setting D large enough will ensure that this is the case. Also note that η has cardinality $M\times N$, where M is the number of base kernels, and N is the number of examples – this means that in order to fully escape the loss function, the algorithm would incur a penalty of $M\times N$, where the hinge loss would only be on the order of N times the average margin violation.

Alternative Minimization

While (4.3) accurately expresses our problem, efficiently optimizing the objective function is difficult because it is non-convex. To address this problem, we "relax" this formulation by performing a block-wise coordinate descent, treating the discount coefficients η fixed at each iteration, and solving the corresponding SVM problem. The value is iteratively updated according to the following expression:

$$\eta_{i,k} \;=\; \frac{\left(y_i(\alpha\circ\mathbf{y})^T K_k(\mathbf{x}_i,\cdot)\right)_-}{\left|\sum_j\left(y_j(\alpha\circ\mathbf{y})K_k(\mathbf{x}_j,\cdot)\right)_-\right|}+1 \tag{4.4}$$

The numerator represents the degree of loss incurred by examples that are actually *misclassified* in kernel k, over and above the hinge-loss, and the denominator represents a normalization over all examples within a single kernel. This is necessary because different kernels have different error variances, which must be accounted for (since we are combining kernels). Adding 1 converts the fraction of incorrectness to a fraction of remaining certainty that kernel k correctly characterizes example i. Having calculated η via Equation (4.4), we construct the kernel matrix K such that $K(\mathbf{x}_i,\mathbf{x}_j)=\sum_k\eta_{j,k}K_k(\mathbf{x}_i,\mathbf{x}_j)$, and train an SVM on this kernel. This process repeats until convergence.

Note that when manipulating kernel matrices it is essential that positive semi-definiteness be preserved, or the alternative formulation will not be any more solvable than the original Model (4.3). Fortunately, setting $\eta_{i,k}$ is equivalent to scaling down an entire row-column pair (recall that kernel matrices are symmetric) which may not preserve positive *definiteness* (as it can lead to 0 elements on the main diagonal) but it does preserve positive *semi*-definiteness.

Experimental Results

As in other works, I evaluated Robust MKL's performance on image scans from the ADNI dataset. In these experiments, I used MR and FDG-PET scans of 159 AD patients (77 AD, 82 controls) from this dataset. These experiments focused on two main questions: does outlier ablation have a noticeable effect on the learned kernel, and, does it result in appreciable gains in predictive accuracy? To address the first question, I analyzed the variation in the kernel matrices as a response to outlier identification and suppression. Second, I evaluated the efficacy of the Robust MKL framework (with outlier detection) as a classification system, with respect to its accuracy using ROC curves.

Evaluation of Outlier Detection

In this section I evaluate the usefulness of outlier detection in the classification setting. Recall that an ideal input to any maximum margin classifier is a dataset where each class is separated from the other by a large margin. The relationship between distance metrics and kernel functions can be understood via the following identity:

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_j)^2 =& \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle \\ =& \langle \mathbf{x}_i, \mathbf{x}_i \rangle - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle \\ =& K_{ii} + K_{jj} - 2K_{ij} \end{aligned}$$

(Note that the diagonal dominance property of positive semi-definite matrices is necessary for a real, *i.e.*, non-imaginary, distance metric.) The term $(\alpha \circ \mathbf{y})^T K (\alpha \circ \mathbf{y})$ in the SVM dual problem represents a minimization of off-diagonal kernel entries

corresponding to support vectors of the same class, and a maximization of off-diagonal entries for support vectors of opposing classes. While counter intuitive at first, this can be understood by recalling that the SVM algorithm chooses the examples that are *most difficult to classify* to be the support vectors. Following this line of reasoning then, if there are some outlier subjects that fall deep within the opposite class, they *will* be picked up as outliers, however, *they may overwhelm the more important support vectors which are the ones on the actual margin*. Robust MKL seeks to compensate for this behavior, and place more weight on examples nearer to the margin. This effect can be seen by visually inspecting the resulting kernel matrices, giving a qualitative evaluation of Robust MKL's performance.

If we order the subjects so that they are grouped by class, then the kernel matrix can be divided into four contiguous regions: similarities between AD subjects(A), between control subjects(C), and between the AD and control subjects (B and B^T.)

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

Each support vector will appear as a "band" which is relatively stronger in B and B^T than in A or C. Thus, outliers will appear as a few extremely strong bands, while the support vectors will be a bit more muted. The effect of outlier removal should reduce the bands corresponding to outliers in the training kernel, *i.e.*, the kernel of training subjects with outlier ablation. More importantly, however, we can also look at the kernel of the test subjects, in which the outliers cannot be identified and ablated (because their labels are not known). In the test kernel (*i.e.*, kernel function between examples in their training role, and examples in their testing role,) the training outliers (horizontal bands) should be ablated, and the test outliers (vertical bands) should be even more extreme because the classifier does not put any emphasis on reducing their errors.

Figure 4.1 shows the kernels produced by Robust MKL. Figure 4.1 (a) and (c) display the uncorrected train and test kernel matrices created simply by summing-up the set of individual kernel matrices. Figure 4.1 (b) and (d) show the corresponding outlier-ablated train and test kernels. In 4.1(a), the outliers from both classes are highly visible, as indicated by the red ellipsoid. This effect is significantly attenuated with outlier detection in Figure 4.1(b), allowing a few other support vectors to appear.

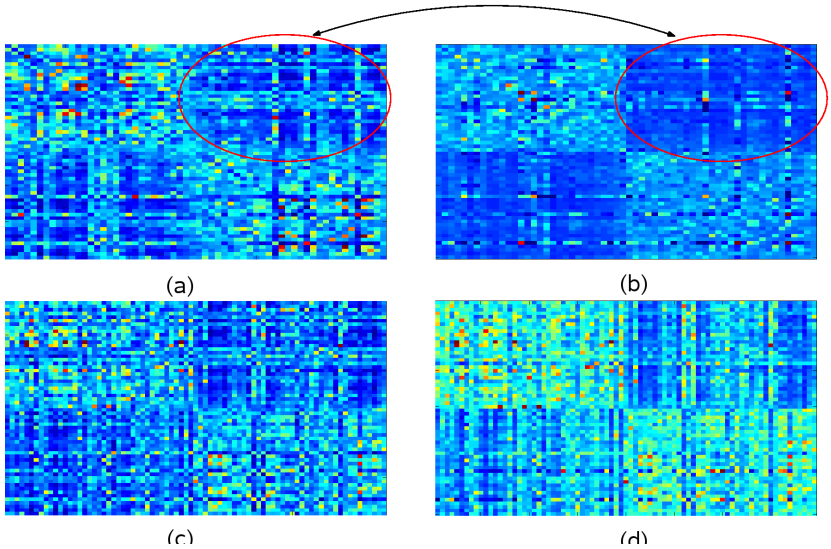


Figure 4.1: (a) Sum of base kernel matrices on training examples. (b) Robust MKL kernel matrix between training examples. Note that the two classes are clearly visible, and the vertical and horizontal lines corresponding to outliers are attenuated. (c) Sum of base kernel matrices on test examples. (d) Robust MKL kernel matrix between test examples. Notice that while there are vertical lines corresponding to outlier test examples, the horizontal lines remain largely attenuated.

Also note the "cleaner" separation between the two classes. Next, observe the effect of outlier ablation on unseen test subjects in Figure 4.1 (c) and (d). For this, the test kernel was constructed with the training examples as rows and test examples as columns. In the uncorrected case in Figure 4.1(c), the vertical lines correspond to unseen outlier subjects, whereas the horizontal lines are attenuated, indicating that in presence of training data, the non-outlier subjects have sharper contrast (causing an improved confidence in classification). Finally, the test kernel (after outlier detection) shown in Figure 4.1(d) shows a stronger within-class signal, and does not attempt to correctly classify the outliers, thereby discounting their effect on the decision boundary as desired (recall hinge loss from Equation (4.1)). It is important to note that the "cleaner" separation seen in the training kernel is carried through to the test subjects, and does not simply represent an overfitting of the data by removing the more difficult examples.

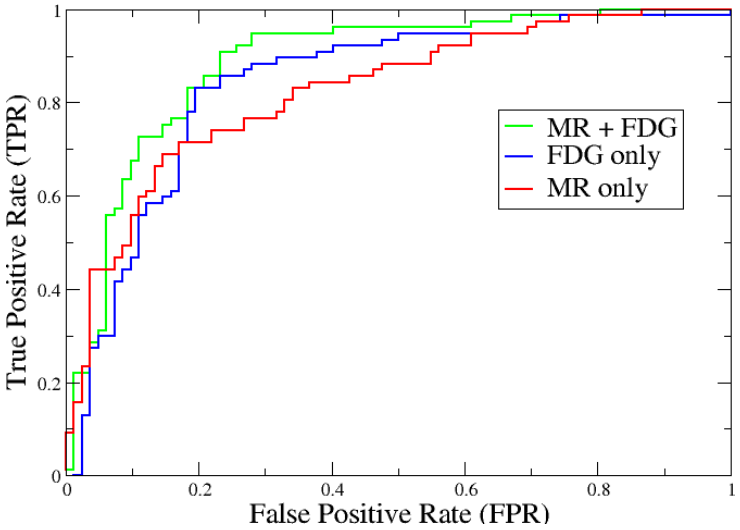


Figure 4.2: ROC curves for the single modal and multimodal classification using Robust MKL

Method	Accuracy	Sensitivity	Specificity	AUC
Robust MKL MR	75.27%	63.06%	81.86%	0.8248
Robust MKL FDG	79.36%	78.61%	78.94%	0.839
Robust MKL (multimodal)	**81.00%**	**78.52%**	**81.76%**	**0.885**

Table 4.2: Accuracy results for the single modal and multimodal classification using Robust MKL

ROC curves and accuracy results

Next, I evaluated the classification accuracy of Robust MKL for single modality classification, using MR and FDG scans individually as well as both these modalities in a combined setting. I used a set of eight kernels each (linear and Gaussian with varying values of σ) for MR and FDG PET: 16 in all. Feature selection was performed using a simple voxel-wise t-test, and thresholding based on the p-values. I used 10-fold cross-validation with 25 realizations (*i.e.*, separate cross-validation runs), and report

the average accuracy, sensitivity, and specificity. Results are summarized in Table 4.2 and Figure 4.2. As expected, we can clearly see that Robust MKL with MR and FDG PET data outperforms the accuracy obtained using only one imaging modality (even when we use multiple kernels with each image type). The area under the curve (AUC) is 0.885 suggesting that it is an effective method for AD classification.

Interpretation of discriminative brain regions

I evaluated the relative importance of various brain regions selected by the algorithm, and whether these regions are consistent with clinically accepted distributions of AD pathology. When using linear kernels, it is possible to recover weights corresponding to the original (voxel-wise) features. This way, the classifier weights correspond to individual voxels, and therefore can be interpreted as distributions of weights on corresponding brain regions. Figure 4.3 shows the calculated weights for Gray Matter Probability (GMP) and FDG-PET images. For GMP, we see the hippocampus and hippocampal gyri are featured prominently, along with middle temporal regions. For FDG-PET, the posterior cingulate cortex and parietal lobules bilaterally are featured prominently. These results validate the Robust MKL method since the selected regions are all known to be affected in AD patients [Jack Jr. et al., 2000, Minoshima et al., 1997].

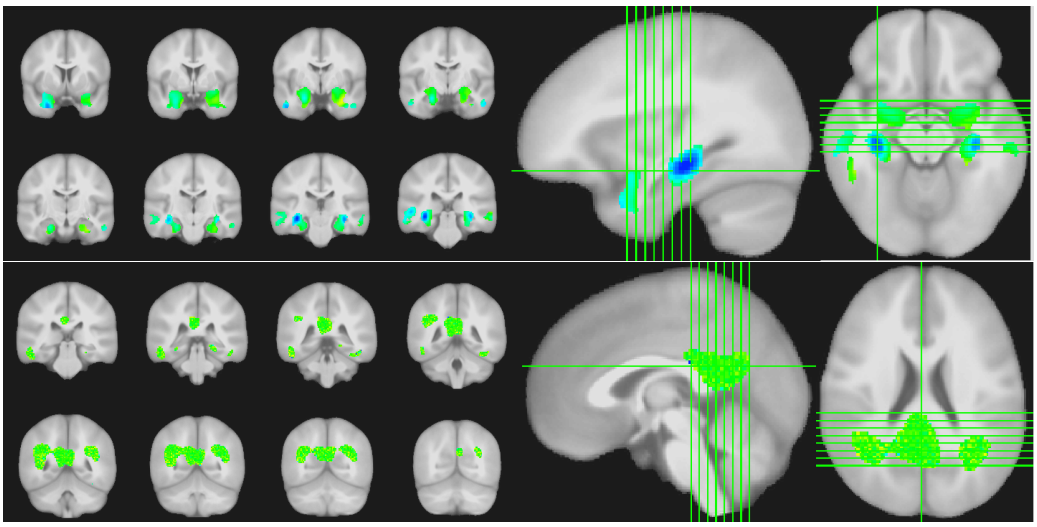


Figure 4.3: (Top) Classifier weights for gray-matter probability images shown overlaid on a template. (Bottom) Classifier weights for FDG-PET images shown overlaid on a template. The images (left) show the discriminative regions as a mosaic. The images (right) are provided for 3D localization.

Chapter 5

Exploiting Modality-modality Interactions

In this chapter I discuss an extension to the MKL framework which incorporates knowledge of the interactions between various modalities available to a learning algorithm. More generally, it can also be beneficial to regularize the interactions between kernels, either in a domain-driven (*i.e.*, using "out-of-band" information,) or in a data-driven sense, *i.e.*, using only empirical measures of interaction. In MKL, model complexity is controlled using various norm regularizations on the kernel combination coefficients, or sub-kernel weights. However, existing methods neither regularize nor exploit potentially useful information pertaining to how kernels in the input set interact; that is, higher-order kernel-pair relationships that can be easily obtained via unsupervised (similarity, geodesics), supervised (correlation in errors), or domain knowledge driven mechanisms (*e.g.*, which features were used to construct the kernel?). In this chapter I will show that by substituting the norm regularizer with an arbitrary quadratic function determined by the square, symmetric matrix $\mathbf{Q}$, one can impose a desired covariance structure on sub-kernel weight selection, and use this as an inductive bias when learning the concept. This formulation significantly generalizes the widely used 1- and 2-norm MKL objectives. I will discuss ramifications in terms of learning bounds (*i.e.*, Rademacher complexity), and explore the model's utility for exploiting aggregate information from several distinct imaging modalities through AD vs. control classification experiments on ADNI data, as well as on several benchmark data sets. Experimental results show that the new model outperforms the state of the art (p-values $\ll 10^{-3}$) in the AD classification task.

5.1 Q-MKL

Kernel learning methods (such as Support Vector Machines) are conceptually simple, strongly rooted in statistical learning theory, and can often be formulated as a convex optimization problem. As a result, SVMs have come to dominate the landscape of

supervised learning applications in bioinformatics, computer vision, neuroimaging, and many other domains. A standard SVM-based 'learning system' can generally be thought of as a composition of two modules [Guyon and Elisseeff, 2003, Gehler and Nowozin, 2009b, Zhang et al., 2011a]:

1. feature pre-processing;

2. a core (linear) learning algorithm.

The design of a kernel encompasses the usual notions of feature preprocessing and may involve using different sets of extracted features, dimensionality reduction tools and methods, or parameterizations of the kernel functions. Each of these alternatives produces a distinct kernel matrix. (Or a distinct kernel function if using a method which only instantiates the kernels on-the-fly.) While much research has focused on efficient methods for the latter (*e.g.*, support vector learning) step, specific choices of feature pre-processing are frequently a dominant factor in the system's overall performance as well, and may involve significant user effort.

Multi-kernel learning [Lanckriet et al., 2004, Sonnenburg et al., 2006, Rakotomamonjy et al., 2008] transfers a significant part of this burden from the user to the algorithm. Rather than selecting a single kernel, MKL offers the flexibility of specifying a large set of kernels corresponding to the many options (*i.e.*, kernels) available, and additively combining them to construct an optimized, data-driven Reproducing Kernel Hilbert Space (RKHS) – while *simultaneously* finding a max-margin classifier. MKL has turned out to be very successful in many applications: on several important Vision problems (such as image categorization), some of best known results on community benchmarks come from MKL methods [Gehler and Nowozin, 2009a, Yang et al., 2009]. In the context of our primary motivating application, the current state of the art in multi-modality neuroimaging-based Alzheimer's Disease (AD) prediction [Klöppel et al., 2008, Vemuri et al., 2008] is achieved by multi-kernel methods [Zhang et al., 2011a], where each imaging modality spawns a kernel, or set of kernels.

In allowing the user to specify an arbitrary number of base kernels for combination, MKL provides more expressive power, but this comes with the responsibility to regularize the kernel mixing coefficients so that the classifier generalizes well. While the importance of this regularization cannot be overstated, it is also a fact that commonly

used ℓ_p norm regularizers operate on kernels separately, without explicitly acknowledging dependencies and interactions among them. To see how such dependencies can arise in practice, consider our neuroimaging learning problem of interest: the task of learning to predict the onset of AD. A set of base kernels $K_1, \ldots, K_M$ are derived from several different medical imaging modalities (MRI; PET), image processing methods (morphometric; anatomical modelling), and kernel functions (linear; RBF). Some features may be shared between kernels, or kernel functions may use similar parameters. As a result we expect the kernels' behaviors to exhibit some correlational, or other cluster structure according to how they were constructed. (See Figure 5.1 (a) and related text, for a concrete discussion of these behaviors in our problem of interest.) We will denote this relationship as $\mathbf{Q} \in \mathbb{R}^{M \times M}$. Next, I provide a technical summary of $\mathbf{Q}$-MKL, before moving to a presentation of the experimental evaluations.

From MKL to Q-MKL

Refer to Chapter 2 for a more complete coverage of existing MKL methods. Here I briefly review the standard MKL model of Kloft et al. [2011].

MKL Models

Adding kernels corresponds to taking a direct sum of Reproducing Kernel Hilbert spaces (RKHS), and scaling a kernel by a constant c scales the *axes* of it's RKHS by $\sqrt{c}$. In the MKL setting, the SVM margin regularizer $\frac{1}{2}\|\mathbf{w}\|^2$ becomes a weighted sum $\frac{1}{2}\sum_{m=1}^{M} \frac{\|\mathbf{w}_m\|^2_{\mathcal{H}_m}}{\beta_m}$ over contributions from RKHS's $\mathcal{H}_1, \ldots, \mathcal{H}_M$, where the vector of mixing coefficients, β scales each respective RKHS [Kloft et al., 2011]. A norm penalty on β ensures that the units in which the margin is measured are meaningful (provided the base kernels are normalized). The MKL primal problem is given as

$$\begin{aligned} \min_{\mathbf{w}, b, \beta \geqslant 0, \xi \geqslant 0} \quad & \frac{1}{2}\sum_{m}^{M} \frac{\|\mathbf{w}_m\|^2_{\mathcal{H}_m}}{\beta_m} + C\sum_{i}^{n} \xi_i + \|\beta\|_p^2 \\ \text{s.t.} \quad & y_i \left(\sum_{m}^{M} \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m} + b \right) \geqslant 1 - \xi_i, \end{aligned} \tag{5.1}$$

where $\phi_m(\mathbf{x})$ is the (potentially unknown) transformation from the original data space to the m^{th} RKHS, $\mathcal{H}_m$. As in SVMs, we turn to the dual problem to see the role of kernels:

$$\begin{aligned} \max_{0 \leqslant \alpha \leqslant C} \quad & \alpha^T \mathbf{1} - \frac{1}{2}\|\mathbb{G}\|_q, \quad \mathbb{G} \in \mathbb{R}^M \\ & \mathbb{G}_m = (\alpha \circ \mathbf{y})^T K_m (\alpha \circ \mathbf{y}), \end{aligned} \tag{5.2}$$

where $\circ$ denotes element-wise multiplication, and the dual q-norm follows the identity $\frac{1}{p} + \frac{1}{q} = 1$. Note that the primal norm penalty $\|\beta\|_p^2$ becomes a dual-norm on the vector $\mathbb{G}$. At optimality, $\mathbf{w}_m = \beta_m (\alpha \circ \mathbf{y})^T \phi_m(X)$, so the term $\mathbb{G}_m = (\alpha \circ \mathbf{y})^T K_m (\alpha \circ \mathbf{y}) = \frac{\|\mathbf{w}_m\|_{\mathcal{H}_m}^2}{\beta_m^2}$ is the vector of *scaled* classifier norms. This shows that the dual norm is tied to how MKL measures the margin in each RKHS.

The Q-MKL model

The key characteristic of **Q**-MKL is that the standard ℓ_p-norm penalty on β and the corresponding dual-norm penalty on classifier magnitudes in Equation (5.2) is substituted with a more general class of positive semi-definite penalty functions, expressed as $\beta^T \mathbf{Q} \beta$. We first present the formalization and then provide a discussion to justify the design. The primal model is given as:

$$\begin{aligned} \min_{w, b, \beta \geqslant 0, \xi \geqslant 0} \quad & \frac{1}{2} \sum_m^M \frac{\|\mathbf{w}_m\|_2^2}{\beta_m} + C \sum_i^n \xi_i + \beta^T \mathbf{Q} \beta \\ \text{subject to} \quad & y_i \left(\sum_m^M \langle \mathbf{w}_m, \phi_m(\mathbf{x}_i) \rangle_{\mathcal{H}_m} + b \right) \geqslant 1 - \xi_i. \end{aligned} \tag{5.3}$$

It is easy to see that if $\mathbf{Q}$ is $\mathbf{1}^{n \times n}$, we obtain the standard (squared) 1-norm MKL as a special case. On the other hand, setting $\mathbf{Q}$ to $\mathbb{I}^{n \times n}$ (identity), **Q**-MKL reduces to 2-norm MKL.

5.2 The Case for Q-MKL

Extending the MKL regularizer to arbitrary quadratics $\mathbf{Q} \succeq 0$ significantly expands the richness of the MKL framework; yet we can show that for reasonable choices of $\mathbf{Q}$, this actually *decreases* MKL's learning-theoretic complexity. In the following I present three motivating intuitions which provide some insight as to why it is a good idea to use this type of regularization, and provide some simple schemes for constructing a $\mathbf{Q}$ matrix. Subsequently, I present a more formal argument analyzing the Rademacher complexity of $\mathbf{Q}$-MKL.

(Intuition 1) Spectral Clustering and Laplacian. Consider $\mathbf{Q}$ to define the graph Laplacian of a similarity function on the base kernels, *i.e.*, which captures dependencies among kernels as edge weights in a graph with kernels as nodes. The choice of the similarity measure is mostly unrestricted (see Table 5.1 for some examples) because taking the graph Laplacian ensures that the regularizer will be convex (graph Laplacians are diagonally dominant). Using the Laplacian as $\mathbf{Q}$, we can expect that there will be a few small eigen-values, and their eigen-vectors will correspond to *clusters of kernels*. The small eigen-values mean that β, and hence the learned kernel K^*, will be biased towards spectrally-derived clusters of kernels, offering a regularization based on higher order interactions between base kernels.

(Intuition 2) Error covariances. Consider a case in which we are boosting weak learners, each trained from a single base kernel, as in [Gehler and Nowozin, 2009a]. Boosting theory requires that (in the optimal case) the weak learners' *errors* will be uncorrelated so that when combined, their errors will cancel [Rudin et al., 2004]. However, this is often violated in practice, and also in the multi-modality AD classification problem described above. By incorporating an estimate of the degree of correlation between the base kernels' contribution to the total error, we can instead boost *orthogonal components* from this correlational structure, which better satisfies the independence assumption.

(Intuition 3) SVM parameter correlations. Joachims et al. [2001] derived a theoretical generalization error bound on kernel combinations that depends on the degree

of redundancy between support vectors in SVMs trained on base kernels individually. Using this type of correlational structure, we can derive a $\mathbf{Q}$ function between kernels to automatically select a combination of kernels that will maximize this bound. This type of $\mathbf{Q}$ function can be shown to have lower Rademacher complexity (see below) while simultaneously decreasing the error bound from [Joachims et al., 2001], which does not directly depend on Rademacher complexity.

The common thread among these intuitions is that there is something towards which, or away from which, we would like to bias the kernel mixing weights, β. We should expect that in most cases the eigen-basis that determines this bias will be apparent from the particular characteristics of the problem, but it is nevertheless possible that a boosting-like behavior can be encouraged by estimating various types of interactions between kernels from the training data, and regularizing based on these correlations. The hope is that if we bias β *away* from the major eigen-functions, this will counteract the tendency of β to align with any regularities in the data that are incidental to the classification task, at the expense of slightly increasing the optimization complexity. Another intriguing possibility is that by leveraging the *differences* between the supervised and unsupervised interactions, we may be able to derive a better estimate of the true error covariances, without the confounding influence of data artifacts or normalization issues. However, there is a more formal analysis, detailed next, which shows that an appropriately chosen $\mathbf{Q}$ matrix gives a lower Rademacher complexity model class.

Virtual Kernels, Rademacher Complexity and Renyi Entropy

If we decompose $\mathbf{Q}$ into its component eigen-vectors, we can see that each eigen-vector defines a linear combination of kernels. This observation allows us to analyze $\mathbf{Q}$-MKL in terms of these objects, which I will refer to as *virtual kernels*. I first show that as $\mathbf{Q}^{-1}$'s eigen-values decay, so do the traces of the virtual kernels. Assuming $\mathbf{Q}^{-1}$ has a bounded, non-uniform spectrum, this property can then be used to analyze (and bound) $\mathbf{Q}$-MKL's Rademacher complexity. I then offer a few observations on how $\mathbf{Q}^{-1}$'s Renyi entropy [Renyi, 1961] relates to these learning theoretic bounds.

Virtual Kernels

In the following assume that $\mathbf{Q} \succ 0$ and has eigen-decomposition $\mathbf{Q} = V\Lambda V$, with $V = \{v_1, \cdots, v_M\}$. First, observe that because $\mathbf{Q}$'s eigen-vectors provide an orthonormal basis of $\mathbb{R}^M$, $\beta \in \mathbb{R}^M$ can be expressed as a linear combination in this basis with γ as its coefficients: $\beta = \sum_i \gamma_i v_i = V\gamma$. Substituting in $\beta^T\mathbf{Q}\beta$ we have

$$\begin{aligned}\beta^T\mathbf{Q}\beta &= (\gamma^T V^T)V\Lambda V^T(V\gamma) \\ &= \gamma^T(V^TV)\Lambda(V^TV)\gamma \\ &= \gamma^T\Lambda\gamma = \sum_i \gamma_i^2\lambda_i \end{aligned} \tag{5.4}$$

This simple observation offers an alternate view of what $\mathbf{Q}$-MKL is actually optimizing. Each eigen-vector v_i of $\mathbf{Q}$ can be used to define a linear combination of kernels, which I will refer to as virtual kernel $\widetilde{\mathbb{K}}_i = \sum_m v_i(m)K_m$. Note that if $\widetilde{\mathbb{K}}_i \succeq 0,\ \forall\, i$, then they each define a valid RKHS. This can be ensured by choosing $\mathbf{Q}$ in a specific way, if desired. This leads to the following result:

Lemma 5.1. *If $\widetilde{\mathbb{K}}_i \succeq 0, \forall i$, then $\mathbf{Q}$-MKL is equivalent to 2-norm MKL using virtual kernels instead of base kernels.*

Proof. Let $\mu_i = \gamma_i\sqrt{\lambda_i}$. Then $\beta^T\mathbf{Q}\beta = \|\mu\|_2^2$, as shown in Equation (5.4), and:

$$\begin{aligned}K^* &= \sum_m \beta_m K_m \\ &= \sum_m^M \sum_i^M \gamma_i v_i(m) K_m \\ &= \sum_i^M \mu_i \lambda^{-\frac{1}{2}} \sum_m^M v_i(m) K_m \\ &= \sum_i^M \mu_i \widetilde{\mathbb{K}}_i,\end{aligned}$$

where $\widetilde{\mathbb{K}}_i = \lambda^{-\frac{1}{2}} \sum_m^M v_i(m)K_m$ is the i^{th} virtual kernel. The learned kernel K^* is a weighted combination of virtual kernels, and the coefficients are regularized under a

squared 2-norm. □

Rademacher Complexity in MKL

With this result in hand, we can now evaluate the Rademacher complexity of **Q**-MKL by using a recent result for p-norm MKL. We first state a theorem from [Cortes et al., 2010], which relates the Rademacher complexity of 2-norm MKL to the traces of its base kernels.

Theorem 5.2. *[Cortes et al., 2010] The empirical Rademacher complexity on a sample set S of size n, with M base kernels is given as follows (with $\eta_0 = \frac{23}{22}$),*

$$\mathfrak{R}_S(\mathcal{H}_M{}^p) \leqslant \frac{\sqrt{\eta_0 q \|\boldsymbol{u}\|_q}}{n} \tag{5.5}$$

where $\boldsymbol{u} = [\mathrm{tr}(K_1), \cdots, \mathrm{tr}(K_M)]^T$ and $\frac{1}{p} + \frac{1}{q} = 1$.

The bound in Equation (5.5) shows that the Rademacher complexity $\mathfrak{R}_S(\cdot)$ depends on $\|\mathbf{u}\|_q$, which is a norm on the traces of the base kernels. Assuming the base kernels are normalized to have unit trace, the bound for 2-norm MKL, (in which $p = q = 2$,) is governed by $\|\mathbf{u}\|_2 = \sqrt{M}$. However, in **Q**-MKL the virtual kernels' traces are not equal, and are in fact given by $\mathrm{tr}(\widetilde{\mathbb{K}}_i) = \frac{\mathbf{1}^T \nu_i}{\sqrt{\lambda_i}}$. With this expression for the traces of the virtual kernels, we can now prove that the bound given in Equation (5.5) is strictly decreased as long as the eigen-values ψ_i of $\mathbf{Q}^{-1}$ are in the range $(0, 1]$. (Adding 1 to the diagonal of **Q** is sufficient to guarantee this).

Theorem 5.3. *If $\mathbf{Q}^{-1} \neq \mathbb{I}^{M \times M}$ and $\widetilde{\mathbb{K}}_i \succeq 0 \ \forall i$, then the bound on Rademacher complexity given in Equation (5.5) is strictly lower for **Q**-MKL than for 2-norm MKL.*

Proof. By Lemma 5.1, we have that the bound in Equation (5.5) will decrease if $\|\mathbf{u}\|_2$, the norm on the virtual kernel traces, decreases. As shown above, the virtual kernel traces are given as $\mathrm{tr}(\widetilde{\mathbb{K}}_i) = \sqrt{\psi_i}\mathbf{1}^T\nu_i$, meaning that:

$$\begin{aligned}\|\mathbf{u}\|_2^2 &= \sum_{i}^{N} \psi_i (\mathbf{1}^T \nu_i)^2 \\ &= \sum_{i}^{N} \psi_i \mathbf{1}^T \nu_i \nu_i^T \mathbf{1} \\ &= \mathbf{1}^T \mathbf{Q}^{-1} \mathbf{1}.\end{aligned}$$

Clearly, this sum is maximal for $\psi_i = 1,\ \forall i$, which is true if and only if $\mathbf{Q}^{-1} = \mathbb{I}^{M \times M}$. This means that when $\mathbf{Q} \neq \mathbb{I}^{M \times M}$, the bound in (5.5) is strictly lower. □

Lemma 5.1 requires that each virtual kernel $\widetilde{\mathbb{K}}_i \succeq 0$. This is so that a positive combination of virtual kernels will be guaranteed to define a valid RKHS, which is necessary to ensure the equivalence between $\mathbf{Q}$-MKL and 2-norm MKL on virtual kernels. This in turn allows us to apply the result from [Cortes et al., 2010]. We can ensure this is the case by choosing $\mathbf{Q}$ as follows. Note that there may be other ways of guaranteeing this condition; the procedure below is given as a demonstration that such cases do exist.

We will construct $\mathbf{Q}$ by constructing its eigen-vectors directly, after which we may choose the eigen-values arbitrarily so as to be non-uniform. Let V be the matrix of eigen-vectors as columns, such that $\mathbf{Q} = V^T \Lambda V$, where Λ is an arbitrary, non-uniform, diagonal matrix. We begin by setting $V = \mathbb{I}^{M \times M}$. Next, we arbitrarily choose a pair of base kernels K_1 and K_2, and find the minimum $c \in \mathbb{R}$ such that $K_1 + cK_2 \succeq 0$ and $K_2 + cK_1 \succeq 0$. We then put c in $V(1,2)$, and $-c$ in $V(2,1)$, and renormalize the first 2 columns of V. This way, the two updated columns of V are normalized, orthogonal, and define p.s.d. virtual kernels. Let this updated V be denoted as $V_{1,2}$. If desired, we may construct $V_{i,j}$ for all other pairs of kernels K_i and K_j in a similar fashion, and combine the resulting orthonormal matrices.

Note that while this procedure does indeed guarantee that the virtual kernels are p.s.d., it is rather restrictive. In practice, such a $\mathbf{Q}$ matrix is not likely to differ substantially from the identity matrix. We therefore provide the following result, which frees us from this restriction and has more practical significance.

Theorem 5.4. **Q**-*MKL is equivalent to the following model:*

$$\begin{aligned} \min_{\boldsymbol{w},\mathfrak{b},\mu,\xi\geqslant 0} \quad & \frac{1}{2}\sum_{\mathfrak{m}}^{M}\frac{\|\boldsymbol{w}_{\mathfrak{m}}\|^2_{\mathcal{V}_{\mathfrak{m}}}}{\mu_{\mathfrak{m}}} + C\sum_{\mathfrak{i}}^{n}\xi_{\mathfrak{i}} + \|\mu\|_2^2 \\ \textit{s.t.} \quad & y_{\mathfrak{i}}\left(\sum_{\mathfrak{m}}^{M}\langle \boldsymbol{w}_{\mathfrak{m}}, \phi_{\mathfrak{m}}(\boldsymbol{x}_{\mathfrak{i}})\rangle_{\mathcal{H}_{\mathfrak{m}}} + \mathfrak{b}\right) \geqslant 1-\xi_{\mathfrak{i}}, \\ & \mathbf{Q}^{-\frac{1}{2}}\mu \geqslant 0 \end{aligned} \tag{5.6}$$

where $\phi_{\mathfrak{m}}()$ *is the feature transform mapping data space to the* $\mathfrak{m}^{\mathfrak{th}}$ *virtual kernel, denoted as* $\mathcal{V}_{\mathfrak{m}}$.

Proof. The result follows from the observation that the two problems are equivalent up to a change of variables. □

While the virtual kernels themselves may be indefinite, recall that $\mu = \mathbf{Q}^{\frac{1}{2}}\beta$, and so the constraint $\mathbf{Q}^{-\frac{1}{2}}\mu \geqslant 0$ is equivalent to $\beta \geqslant 0$, which guarantees that the combined kernel will be p.s.d. This formulation is slightly different than the 2-norm MKL formulation, however it does not alter the theoretical guarantee of [Cortes et al., 2010], providing a stronger result.

Renyi Entropy

Theorem 5.3 points to an intuitive explanation of where the benefit comes from as well, if we analyze the Renyi entropy [Renyi, 1961] of $\mathbf{Q}^{-1}$. Renyi entropy significantly generalizes the usual notion of Shannon entropy, [Jenssen, 2010, Girolami, 2002, Erdogmus and Principe, 2002]. Renyi entropy has applications in statistics, statistical mechanics, [Lenzi et al., 2000], and many other fields, and has recently been proposed as an alternative to PCA [Jenssen, 2010]. The quadratic Renyi entropy of a probability measure is given as:

$$H(p) = -\log\int p^2(\mathbf{x})d\mathbf{x}.$$

Now, if we use a kernel function (i.e., $\mathbf{Q}^{-1}$) and a finite sample (*i.e.*, base kernels) as a kernel density estimator (*c.f.* [Ong et al., 2005],) then with some normalization we can derive an estimate of the underlying probability $\hat{p}$, which is a distribution over base

kernels. We can then interpret its Renyi entropy as a complexity measure on the space of combined kernels. Equation (5.2) in [Girolami, 2002] relates the virtual kernel traces to the Renyi entropy estimator of $\mathbf{Q}^{-1}$ as $\int \hat{p}^2(\mathbf{x})\,d\mathbf{x} = \frac{1}{N^2}\mathbf{1}^T\mathbf{Q}^{-1}\mathbf{1}$,[1] which leads to a nice connection to Theorem 5.3. This view informs us that setting $\mathbf{Q}^{-1} = \mathbb{I}^{M\times M}$, (i.e., 2-norm MKL), has *maximal* Renyi entropy because it is maximally uninformative; adding structure to $\mathbf{Q}^{-1}$ concentrates $\hat{p}$, reducing both its Renyi entropy, *and* Rademacher complexity together.

This series of results suggests an entirely new approach to analyzing the Rademacher complexity of MKL methods. The proof of Theorem 5.3 relies on decreasing a norm on the virtual kernel traces, which we now see directly relates to the Renyi entropy of $\mathbf{Q}^{-1}$, as well as with decreasing the Rademacher complexity of the search space of combined kernels. It is even possible that by directly analyzing Renyi entropy in a multi-kernel setting, this conjecture may be useful in deriving analogous bounds in*e.g.*, Indefinite Kernel Learning [Kowalski et al., 2009], because the virtual kernels are indefinite in general.

Function	$\mathbf{Q}(i,j) =$	Arguments	Uses
Covariance	$\frac{\langle K_i, K_j\rangle}{\|K_i\|\|K_j\|}$	$K_{1,\ldots,M}$	Unsupervised covariance (matrix cosine)
Eigen-space alignment	$\frac{1}{N}\sum_k^N \lvert v_{ik}^T v_{jk}\rvert$	$K_{1,\ldots,M}$	Mean cosine of the first N eigen-vectors
Histogram intersection	$\|\min(K_i, K_j)\|_1$	$K_{1,\ldots,M}$	Interpret kernels as histograms; $\min(\cdot,\cdot)$ applied entry-wise.
Training covariance	$\mathrm{err}(K_i, \mathbf{y})^T \mathrm{err}(K_j, \mathbf{y})^T$	$K_{1,\ldots,M}, \mathbf{y}$	Covariance of training errors
α covariance	$\alpha_i^T \alpha_j$	$K_{1,\ldots,M}, \mathbf{y}$	Covariance of SVM α parameters

Table 5.1: $\mathbf{Q}$-functions, their arguments, and their uses.

[1] Note that this involves a Gaussian assumption, but [Erdogmus and Principe, 2002] provides extensions to non-Gauss kernels.

Designing Q-functions

In some settings, including multi-modality AD classification, domain knowledge of how the kernels were constructed is sufficient to fully populate the $\mathbf{Q}$ matrix. However, when this is not the case we require a method for inferring the proper $\mathbf{Q}$ matrix empirically from the data. The design of $\mathbf{Q}$ depends on the *type* of interaction structure (and the corresponding bias) we wish to impose – recall that the value of $\mathbf{Q}$-MKL is that it replaces the task of designing the kernel with designing a similarity matrix *between* kernels. (*c.f.* hyperkernels in [Ong et al., 2005], which are similar in concept, but very different in form and especially in implementation-level details, requiring a Semi-Definite Program to be solved, or a problem which has a number of parameters that is quartic with respect to the number of examples.)

The ideas relating to the Laplacian above are simple yet effective, and the user can always revert to 1- and 2-norm MKL with an uninformative $\mathbf{Q}$. Just as with kernel design in SVMs, many possibilities are available for $\mathbf{Q}$.

Consider the behavior of $\mathbf{Q}$ matrices. Viewed in terms of individual entries, positive off-diagonal entries will *penalize* putting weight on both of the kernels, while negative entries *encourage* doing so. At a macro-level, each eigen-vector of $\mathbf{Q}$ has an associated cost – it's corresponding eigen-value. Thus, the eigen-decomposition of $\mathbf{Q}$ gives an indication of the bias imbued by a particular $\mathbf{Q}$-matrix: β is more likely to resemble the least eigen-vectors, inversely proportional to their eigen-values – this provides the spectral clustering perspective. Thus, we could in principle construct a $\mathbf{Q}$-matrix by choosing the eigen-vectors directly, however it is easier to take the graph Laplacian of an arbitrary similarity matrix which we would like β to resemble. This has the benefit that the similarity matrix need not be positive definite, as the Laplacian of any graph is guaranteed to be positive semi-definite. Alternatively, recall that current multi-kernel methods seek to maximize the margin in the combined RKHS, however overly focusing on the margin can lead to an increase in error variance (*i.e.*, taken over training sets as the random variable) [Shivaswamy and Jebara, 2010]. This suggests choosing $\mathbf{Q}$ to control this variance directly. A simple way of doing so is to use the *covariances of the training errors of kernels*, which clusters the kernels in terms of their training error. Similarly, for unsupervised similarity measures, there are many options. A brief, but by no means exhaustive, list of $\mathbf{Q}$ matrices is shown in Table 5.1.

Relative Margin

Before describing **Q**-MKL 's optimization strategy, I discuss an interesting extension to the **Q**-MKL model. Several interesting extensions to the SVM and MKL frameworks have been proposed that focus on *relative* margin methods [Shivaswamy and Jebara, 2010, Gai et al., 2010] which maximize the margin *relative* to the spread of the data. An intuitive justification for this approach is that if the spread of the data is large relative to the margin, then there is an implied uncertainty in the classifier's future output, and it is more likely that unseen examples will fall on the wrong side of the margin owing to this uncertainty. In particular **Q**-MKL can be easily modified to incorporate the Relative Margin Machine (RMM) model [Shivaswamy and Jebara, 2010] by replacing Module 1 as in Equation (5.7) with the RMM objective. Our alternating optimization approach (described next) is not affected by this addition; however, the additional constraints would mean that SMO-based strategies would not be applicable.

Optimization

In order to employ **Q**-MKL in practical settings, a core engine to optimize Equation (5.3) must first be developed. Most MKL implementations make use of an alternating minimization strategy which first minimizes the objective in terms of the SVM parameters, and then with respect to the sub-kernel weights, β. Since the MKL problem is convex, this method leads to global convergence [Rakotomamonjy et al., 2008, Kloft et al., 2011] and minor modifications to standard SVM implementations are sufficient. **Q**-MKL generalizes the norm regularizer on β to arbitrary positive semi-definite quadratic functions, so the feasible set is the same as for MKL, while the objective is generalized to a larger class of convex functions. This directly gives:

Property 1. *The* **Q***-MKL model in Equation (5.3) is convex.*

The optimization strategy I developed for **Q**-MKL broadly follows this strategy. But as will become clear shortly, interaction between the sub-kernel weights makes the optimization of β more involved than [Sonnenburg et al., 2006, Kloft et al., 2011], and requires new solution mechanisms.

One may consider this process as a composition of two modules: one which solves for SVM dual parameters (α, b) with fixed β, and the other for solving for β with fixed SVM parameters. In each iteration we alternate between the following two problems:

(Module 1)

$$\max_{0 \leqslant \alpha \leqslant C} \alpha^T \mathbf{1} - \alpha^T Y K Y \alpha \quad \text{s.t.} \quad \alpha^T y = 0 \tag{5.7}$$

(Module 2)

$$\min_{\beta \geqslant 0} \sum_m \frac{\|\mathbf{w}_m\|^2}{\beta_m} \quad \text{s.t.} \quad \beta^T \mathbf{Q} \beta \leqslant 1 \tag{5.8}$$

Module 1 in Equation (5.7) reduces to an instance of SVM; however, Module 2 in Equation (5.8) is a quadratically constrained problem in a unusual form. Notice that (5.8) appears difficult because the optimization variables appear in the *denominator*. Further, the objective is a sum of ratios – fortunately, however, the numerators are constant, so the problem is not an instance of fractional programming, which is NP-Hard in general[]. Secondly, the inequality is expressed as a quadratic constraint. This makes (5.8) a challenging problem to solve with standard QP solvers. My solution is based on the observation that an optimal solution will be a stationary point. We can write the gradient in terms of Lagrange multiplier δ:

$$\frac{\|\mathbf{w}_m\|^2}{\beta_m^2} - \delta(\mathbf{Q}\beta)_m = 0, \quad \forall m \in \{1, \cdots, M\}. \tag{5.9}$$

We now need only to eliminate δ so that the non-linear system will be limited to quadratic terms in β, allowing us to use a non-linear system solver. Let

$$\mathbf{W} = \text{Diag}(\|\mathbf{w}_1\|_{\mathcal{H}_1}^2, \ldots, \|\mathbf{w}_M\|_{\mathcal{H}_M}^2)$$

and

$$\beta^{-2} = (\beta_1^{-2}, \ldots, \beta_M^{-2}).$$

We can then write $\mathbf{W}\beta^{-2} = \delta(\mathbf{Q}\beta)$. Now, solving for β (on the right-hand side) gives

$$\beta = \frac{1}{\delta} \mathbf{Q}^{-1} \mathbf{W} \beta^{-2} \tag{5.10}$$

Because $\mathbf{Q} \succeq 0$, at optimality the constraint $\beta^T \mathbf{Q} \beta \leqslant 1$ must be active, that is, the

constraint must be at equality. So, we can plug in the above identity to solve for δ,

$$1 = \left(\frac{1}{\delta}\mathbf{Q}^{-1}\mathbf{W}\beta^{-2}\right)^{\mathsf{T}} \mathbf{Q} \left(\frac{1}{\delta}\mathbf{Q}^{-1}\mathbf{W}\beta^{-2}\right) \tag{5.11}$$

$$\delta^2 = \mathbf{W}\beta^{-2}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{W}\beta^{-2} \tag{5.12}$$

We directly obtain

$$\delta = \sqrt{(\mathbf{W}\beta^{-2})^{\mathsf{T}}\mathbf{Q}^{-1}(\mathbf{W}\beta^{-2})} \tag{5.13}$$

$$= \|\mathbf{W}\beta^{-2}\|_{\mathbf{Q}^{-1}}, \tag{5.14}$$

where the Lagrange multiplier δ effectively normalizes $\mathbf{W}\beta^{-2}$ according to $\mathbf{Q}^{-1}$, meaning that the relative magnitude of $\|\mathbf{w}_m\|^2\beta^{-2}$ must be equal to the normalized covariance of β with β_m. Module 1 is easy to solve with any SVM method. Any nonlinear root solver (*e.g.*, GNU Scientific Library) is sufficient to find the value for β (Module 2). Putting these parts together, we have the following algorithm for optimizing $\mathbf{Q}$-MKL:

Algorithm 1. Q-*MKL*

Input: Kernels $\{K_1, \cdots, K_M\}$*;* $\mathbf{Q} \succeq 0 \in \mathbb{R}^{M\times M}$*; labels* $y \in \{\pm 1\}^N$.

Outputs: α*,* b*,* β

$\beta^{(0)} = \frac{1}{M}$; $t = 0$ *(iterations)*

while *not optimal* ***do***

$K^{(t)} \leftarrow \sum_m \beta_m^{(t)} K_m$

$\alpha^{(t)}, b^{(t)} \leftarrow SVM(K^{(t)}, C, y)$ *(**Module 1**, Equation (5.7))*

$W_{mm} = \alpha^{(t)\mathsf{T}} K_m^{(t)} \alpha^{(t)} (\beta_m^{(t)})^2$

$\beta^{(t+1)} \leftarrow \arg\min (Problem(5.8))$ *(**Module 2**, Equation (5.8))*

$t = t + 1$

end while

It can be shown that $\mathbf{Q}$-MKL can be solved optimally by noting that Module 2 can be precisely optimized at each step, and in practice, on the order of a few tens of iterations are all that is required.

Finally, a remark on normalizing the scale of β: If $\mathbf{Q}$ has any eigen-values $\lambda_i \approx 0$, then β is effectively unregularized along the direction of the corresponding eigen-

vectors ν_i, which has the effect of allowing the scale of β to grow quite large. If so, then this property will scale the kernel matrices in such a way that the C parameter may lose its interpretation. That is, the C parameter reflects a trade-off between regularizer and loss. However, under standard regularizers the regularizer also controls the units in which the loss is measured, meaning that the two are not completely unconnected. Therefore, if the regularizer does not control the units in which the loss is measured, then this role falls to the loss function itself, at which point the C parameter's meaning is altered. Since β is constrained to be nonnegative, eigen-vectors must be nonnegative as well in order to be fully unconstrained – which is guaranteed to be the case when $\mathbf{Q}$ is a graph Laplacian.

There are two separate approaches to combatting this problem: one is to set $\mathbf{Q} = \mathbf{Q} + \mathbb{I}^{M \times M}$, which effectively adds a $\|\beta\|_2^2$ regularizer term. It also happens that this will guarantee that the eigen-values of $\mathbf{Q}$ are greater than one, and hence the eigen-values of $\mathbf{Q}^{-1}$ are all less than one, which is required for the theoretical guarantees of Section 5.2. Alternatively, one could add $\varepsilon \mathbf{1}^{M \times M}$ to $\mathbf{Q}$ to directly penalize the least eigen-vector when $\mathbf{Q}$ is a graph Laplacian. These two cases correspond to adding an additional 1- or 2-norm regularizer on top of the $\mathbf{Q}$-norm. A second approach is to scale β to unit 1- or 2-norm at each iteration, which affects only the scale of the combined kernel. This can be thought of as mixing penalty-based and constraint-based regularizers because this is the behavior of a projected-gradient method for constrained optimization. In practice, this approach did not affect convergence.

Successive approach using Newton's method. Note that, in practice, widely available methods for root-finding are sufficient. However, I also implemented a successive method based on Newton's method [Nocedal and Wright, 1999] which may be more appropriate for certain applications. Notice that in general, while the number of kernels may be large, it is unlikely that the size of $\mathbf{Q}$ (quadratic in the number of kernels) will dominate the total size of all the kernels, which are quadratic in the number of examples. If so (as in a majority of computer vision and neuroimaging problems), it may be advantageous to employ second-order methods to solve Equation (5.8) for β in terms of $\mathbf{w}$. Newton's method iterates with the following update: $\beta \leftarrow (\beta - \mathrm{H}^{-1}\mathrm{g})$

where the Hessian H and gradient g are

$$\mathrm{H} = \left[\mathbf{Q} + 2\ \mathrm{Diag}\left(\frac{\|\mathbf{w}\|_{\mathrm{m}}^2}{\beta^3}\right)\right]$$
$$\mathrm{g} = \left(\mathbf{Q}\beta - \frac{\|\mathbf{w}\|_{\mathrm{m}}^2}{\beta_{\mathrm{m}}^2}\right)$$

In order to compute these functions we need only the $\|\mathbf{w}_{\mathrm{m}}\|_2^2$ term for each sub-kernel, which is given as $\|\mathbf{w}_{\mathrm{m}}\|_{\mathcal{H}_{\mathrm{m}}}^2 = \frac{1}{2}\beta_{\mathrm{m}}^2\left((\alpha \circ \mathbf{y})^{\mathsf{T}}\mathrm{K}_{\mathrm{m}}(\alpha \circ \mathbf{y})\right)$. Again, standard SVM implementations can be used. This mechanism comes with a pitfall: there is no guarantee that $\beta \geqslant 0$ at the optimum, in which case we must substitute $\beta_{\mathrm{m}} \leftarrow 0,\ \forall \beta_{\mathrm{m}} < 0$, essentially projecting β back into the nonnegative orthant (here, gradient and Hessian terms must also be set to zero where the corresponding $\beta = 0$ to rule out infinite values). Nonetheless, this method works well experimentally, and Algorithm 1 generally converges in about 10 iterations.

5.3 Experiments

I performed extensive experiments to validate the $\mathbf{Q}$-MKL model, examine the effect its regularization scheme has on β, and to assess its advantages in the context of AD classification. In the first set of experiments I evaluate $\mathbf{Q}$-MKL 's performance on benchmark UCI datasets [Frank and Asuncion, 2010]. I include these experiments in order to show that the proposed regularization scheme does not worsen its performance relative to existing MKL models, and in some cases the data-driven regularizer may even improve performance. Note that while there are some interesting theoretical guarantees which show that $\mathbf{Q}$-MKL employs a stronger regularizer, this does not by itself ensure greater accuracy. Rather, it is the ability to program into $\mathbf{Q}$ properties of the data which are known beforehand or through domain knowledge which allow us to generate a stronger regularizer for the same amount of training set error. This is an important validation because if one supposes that the purpose of regularization is to push the model *away* from sampling artifacts in the data, then a data-driven regularizer might defeat this purpose. The UCI results show that this is not the case, and in fact there can, in some cases, be a benefit to using a purely data-driven regularizer.

In the main experiments, I demonstrate in the concrete setting of neuroimaging analysis how domain knowledge can be adapted to improve the algorithm's performance. My focus on a practical application is intended as a demonstration of how domain knowledge can be seamlessly incorporated into a learning model, giving significant gains in accuracy.

UCI datasets

I begin with an evaluation of several $\mathbf{Q}$-MKL regularizers on standard UCI repository datasets [Frank and Asuncion, 2010]. In order to facilitate comparison, I followed the methods of the SimpleMKL experiments [Rakotomamonjy et al., 2008]. Briefly, I used the same five repositories, (*Liver*, *Pima Diabetes*, *Ionosphere*, *WPBC Breast Cancer*, and *Sonar*), whitened each feature by mean centering and normalizing to unit standard deviation, and normalized kernels to unit trace. The C parameter was set to 100. For kernels I used polynomials of degree one through three, and Gaussians with ten different bandwidths: $\{1, 2.5, 5\} * 10^{-1,0,1}$, and 100. I used 4-fold cross validation with 20 iterations to approximate the 70% training sets used in [Rakotomamonjy et al., 2008]. For each data set, I repeated the entire process with several different $\mathbf{Q}$-functions: (Pseudo-) Inverse and Graph-Laplacian of matrix covariance, training error covariance, and covariance of training-set SVM parameters α. For comparison, I also used Identity (2-norm MKL) and $\mathbf{1}^{\mathrm{M}\times\mathrm{M}}$ (1-norm MKL). Accuracy, sensitivity, specificity and area under ROC curve are shown in Table 5.3.

Several trends can be seen from these results: First, 1-norm MKL significantly underperforms the other methods in two of the data sets (*Liver, Pima*), while slightly over-performing on one of them (*Sonar*). The greater variance of 1-norm MKL's predictive performance is likely attributable to the sparsity it encourages at the kernel level. Next, note that on *Pima, Ionosphere*, and *Breast Cancer* 2-norm MKL is comparable to the two $\mathbf{Q}$-functions used (on *Sonar* Train-Error covariance slightly underperforms the other two) while on *Liver* the two $\mathbf{Q}$-functions significantly outperform 2-norm MKL. From this, we can conclude that, in general, $\mathbf{Q}$-MKL does not induce as significantly varying of a risk as 1-norm does, while performing competitively (or more favorably) than 2-norm MKL. It is interesting to note that the UCI datasets are essentially *unimodal*; there is no *a priori* information that tells us how the kernels are

related. However, if we *do* have extra information of this kind, then $\mathbf{Q}$-MKL provides a direct way of using it, as I describe next.

Multi-modality AD prediction

Next, I performed multi-modality AD prediction experiments using all available kernels. Recall that several different types of imaging modalities are available, each of which highlights a different aspect of disease pathology; MR provides structural information, while FDG-PET assesses hypo-metabolism. Further, in practice we may use several image processing pipelines. Yet, due to the inherent similarities in how the various kernels are derived, there are clear cluster structures / behaviors among the kernels, which we would like to exploit using $\mathbf{Q}$-MKL .

The experimental setup is the same as in [Hinrichs et al., 2011], and is described in Section 4.1. For $\mathbf{Q}$-matrices, I used the Laplacian of covariance between single-kernel α parameters (recall the motivation from Joachims et al. [2001] in Section 5.2) plus a block-diagonal representing clusters of kernels derived from the same imaging modalities. This $\mathbf{Q}$ matrix was designed to balance between biasing towards (1) *clustering* β according to unsupervised similarity (*i.e.*, sample covariances), and between kernels derived from the same modality on the one hand, and (2) *inducing sparsity* in clusters of kernels having highly correlated errors.

I used 10-fold cross-validation with 30 realizations, for a total of 300 folds. Accuracy, sensitivity, specificity and area under ROC curves were averaged over all folds. For comparison I also examined 1-, 1.5-, and 2-norm MKL. Results are shown in Table 5.2; The first observation we can make is that $\mathbf{Q}$-MKL had the highest performance overall, reducing the error rate from 12.5% to 11.2%. The Null hypothesis stating that the differences are not significant can be rejected for $\alpha = 0.001$. We can interpret p-norm MKL methods such that their primary benefit is that they effectively *filter out* uninformative kernels, leaving behind the most informative ones. In this set of experiments, however, the kernels used in these experiments were all derived from neuroimaging data, and were thus highly reliable. $\mathbf{Q}$-MKL's performance suggests that it is better able to combine kernels in a way that boosts the power of the combined classifier.

Regularizer	Acc.	Sens.	Spec.
$\|\beta\|_1$-MKL	0.864	0.771	0.931
$\|\beta\|_{1.5}$-MKL	0.875	**0.790**	0.936
$\|\beta\|_2$-MKL	0.875	**0.789**	0.938
Correlation	0.874	**0.785**	0.939
Eigen-space	0.875	0.786	0.937
Histogram	0.874	**0.788**	0.934
Lap.(diag)	0.884	0.785	**0.955**
Lap.(Cov) + diag	0.874	**0.785**	0.939
Cov_α	**0.884**	0.780	0.942
Lap.(Cov_α)	0.884	0.785	**0.955**
Lap.(Cov_α) + diag	**0.888**	0.786	**0.956**

Table 5.2: Comparison of **Q**-MKL & MKL. Bold numerals indicate methods which did not differ from the best at the 0.01 level using a paired t-test. Lap. = "Laplacian"; diag = "Block-diagonal".

Anatomical analysis

As in other experiments, the equivalent brain regions (computed by training with linear kernels only) are shown in Figure 5.5. Warm colors have positive weight, meaning that intensity in these regions is indicative of health, while cool colors are indicative of pathology. In the FDG-PET images (a-b), we see the posterior cingulate and lateral parietal lobules bilaterally, which are known to be hypometabolic in AD. In the MR-derived images, we see the hippocampus and surrounding gyri bilaterally, and Cerebro-Spinal-Fluid (CSF) voxels are associated with AD, *i.e.*, signs of increased CSF are associated with AD.

Virtual kernel analysis

As blocks of kernels derived from the same imaging modalities are expected to be highly correlated, we next turn to an analysis of the covariance structures found in the data empirically as a concrete demonstration of the type of patterns towards which the **Q**-MKL regularizer is biasing β. Recall that the eigen-vectors of a **Q** matrix can show which covariance patterns are encouraged or discouraged, in proportion to

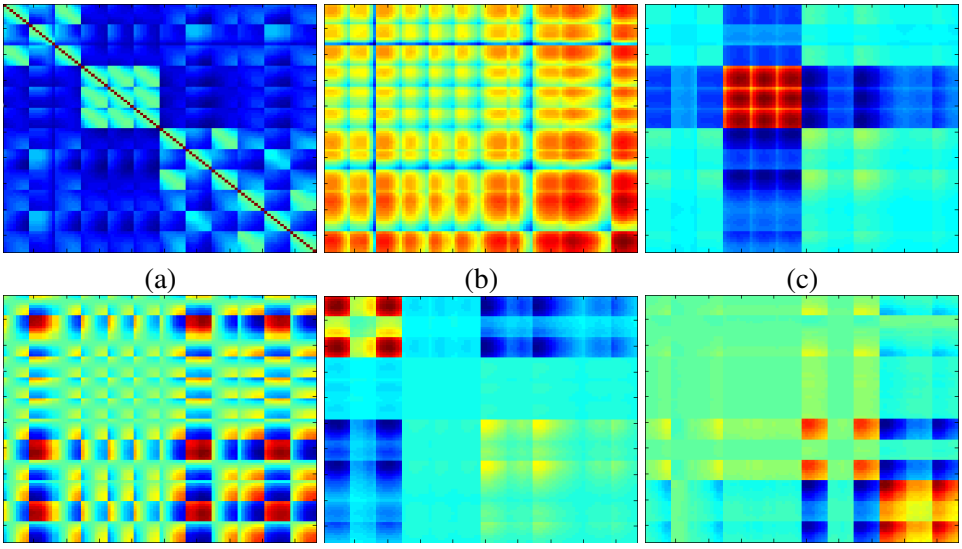


Figure 5.1: Covariance $\mathbf{Q}$ (before graph Laplacian) used in AD experiments (a); the contribution from the three least eigen-vectors (b-d); and the outer product of β from $\mathbf{Q}$-MKL (e). Note the block structure in (a) relating to the imaging modalities and kernel functions. TBM-based kernels have a strong block structure in (b). Quadratic kernels show a surprising covariance pattern among different modalities in (d). This pattern for quadratic kernels is missing in (e), showing the implicit group sparsity structure imposed by $\mathbf{Q}$-MKL.

their eigen-values. In the following, we compare the least 3 eigenvalues from several empirical $\mathbf{Q}$ matrices in the ADNI data.

Covariance Q

The $\mathbf{Q}$-matrix (before graph Laplacian) and the three least eigen-vectors (after) are shown as outer products in Figure 5.1. In Figure 5.1(a), a strong block-structure is visible among the four imaging modalities as expected, though more subtle patterns are also visible. The significant interaction between the two blocks (lower right) is due to the common FDG-PET processing pipeline. Within each of the four blocks, there are three smaller blocks corresponding to each kernel type (linear, quadratic, Gaussian). Note that there is significant interaction between the two blocks in the

lower right. These kernels were derived from FDG-PET scans at two timepoints, using the same image processing pipeline, as opposed to the two MR-derived modalities which used different image analysis methods, which explains the greater degree of covariance between the blocks. Next, we note that the second block from the top left, composed of the TBM-derived kernels, is more "solid", which is reflected in the second eigen-vector, in Figure 5.1(c) – *i.e.*, **Q**-MKL has detected that there is little variation among the various TBM-derived kernels, and automatically merged them into a single cluster. Similarly, in Figure 5.1(d), we see a surprising cluster structure among the quadratic kernels (excluding TBM). The optimal weights β are shown as an outer product in Figure 5.1(e). Note that the pattern of quadratic kernels in Figure 5.1(d) is largely absent from β because those kernels were removed as a group. In all eigen-vectors we can see a strong 8×8 pattern corresponding to the groups of 8 kernels that differ only in terms of the level of feature selection, which is desired.

Histogram Q

Next, for comparison we present the histogram **Q** (Table 1, entry 3). Note that patterns remarkably similar those of the covariance **Q** appear in the least two eigen-vectors, but the third shows a strong connection between the linear and Gaussian VBM kernels, and a slight negative correlation with the linear and Gaussian FDG-PET kernels.

Eigen-space alignment Q

Next we present the eigen-space alignment **Q** (see Table 5.1, entry 2). See Figure 5.3. Note that some patterns are similar to the covariance and histogram intersection **Q** matrices, some are different – note that both VBM- and TBM-derived kernels show a block structure, while the interaction between quadratic kernels is even more prominent. An interesting trend is that the fewer the number of features shared between kernels, (or the more in some cases) the less alignment there is between their eigen-spaces, regardless of the kernel function used; see especially Figure 5.3(b,c,d).

Training error covariance

Lastly, we present a similar analysis of the **Q**-function derived from training error covariance. See Figure 5.4. As with the unsupervised **Q**-functions, there are several

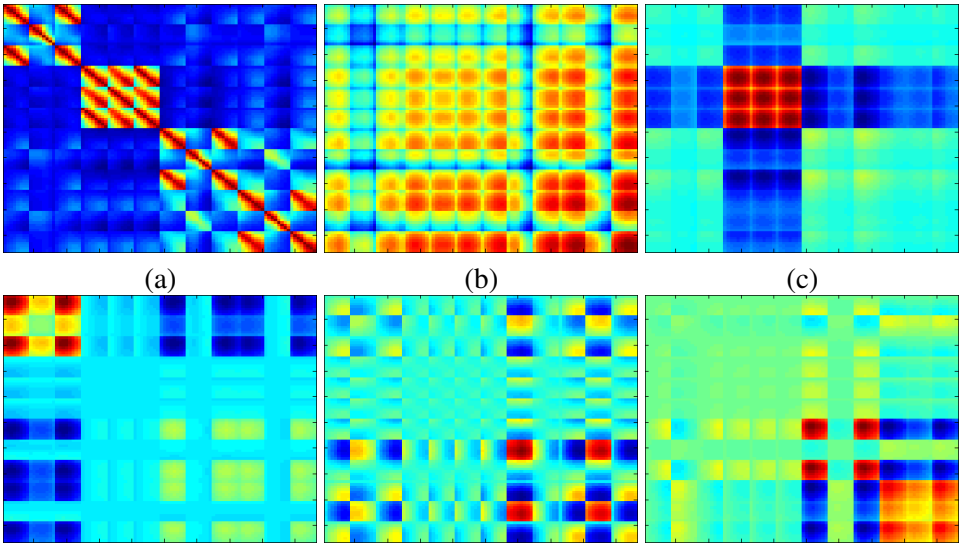


Figure 5.2: (a) Histogram $\mathbf{Q}$ matrix. (b–f) Least 5 eigen-vectors, represented as outer products.

block-structures corresponding to similarities in the construction of the kernels, *e.g.*, among VBM, TBM, quadratic, and linear and Gaussian kernels. This is expected, because where features are correlated we can expect there to be some error correlation as well. Notice, however, that in Figure 5.4(a) there are some interesting differences. Particularly, some of the FDG-PET quadratic kernels are more strikingly anti-correlated with the rest of the kernels, and within the quadratic kernels there appear to be some sub-clusters as well. (See Figure 5.4(a,b,e).) Moreover, overlapping features seem to be less of a dominant factor – note the "flatter" appearance of the blocks, with less of a "gradient" moving from upper-left to lower-right within the blocks. (*c.f.* Figures 5.1, 5.3.) An intriguing possibility is that by leveraging the *differences* between the supervised and unsupervised interactions, we may be able to derive a better estimate of the true error covariances, without the confounding influence of data artifacts or normalization issues.

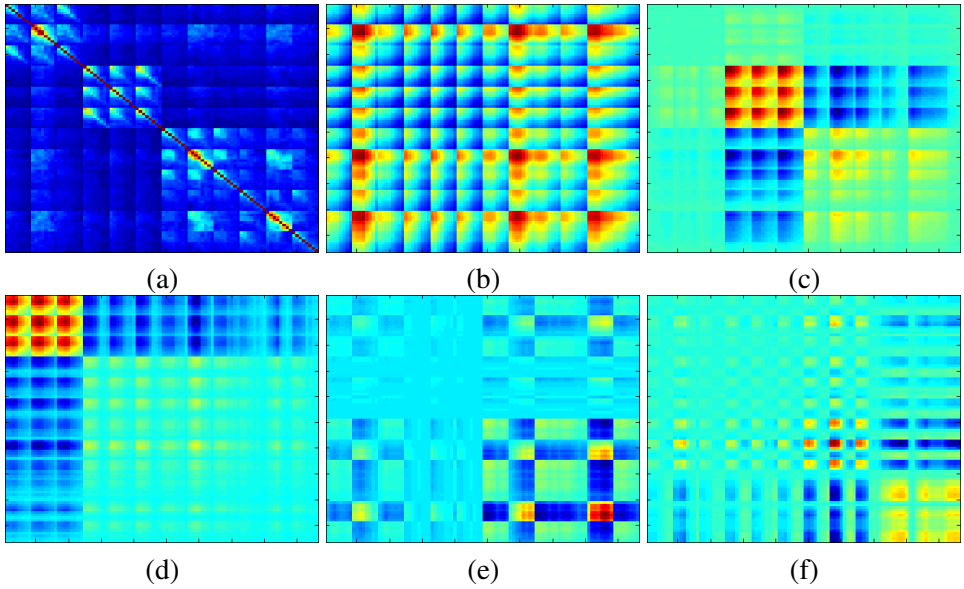


Figure 5.3: (a) Eigen-space alignment $\mathbf{Q}$ matrix. (b-f) Least 5 eigen-vectors, represented as outer products.

5.4 Conclusions

MKL is an elegant method for aggregating multiple data views, and is being extensively adopted for a variety of problems in machine learning, computer vision, bioinformatics, and neuroimaging. $\mathbf{Q}$-MKL extends this framework to account for and exploit higher-order interactions between kernels – derived from supervised, unsupervised, or domain-knowledge driven – as shown in Figure 5.1. Note that $\mathbf{Q}$-MKL is not only concerned with selecting or discarding groups of kernels, but also with choosing the right weighted combination of kernels. This flexibility can impart greater control over how the model utilizes cluster structure among kernels, and effectively encourage cancellation of errors wherever possible. I have presented a convex optimization model to efficiently solve the resultant model, and presented experiments on the challenging problem of identifying Alzheimer's disease based on multi modal brain imaging data (obtaining statistically significant improvements), as well as on benchmark datasets. In the next chapter I will describe advances in Alzheimer's Disease research made possible by

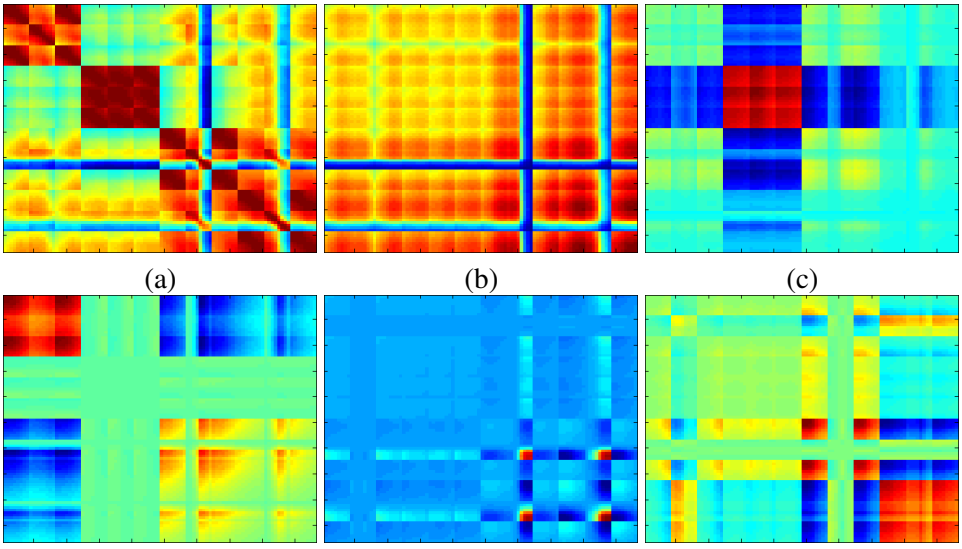
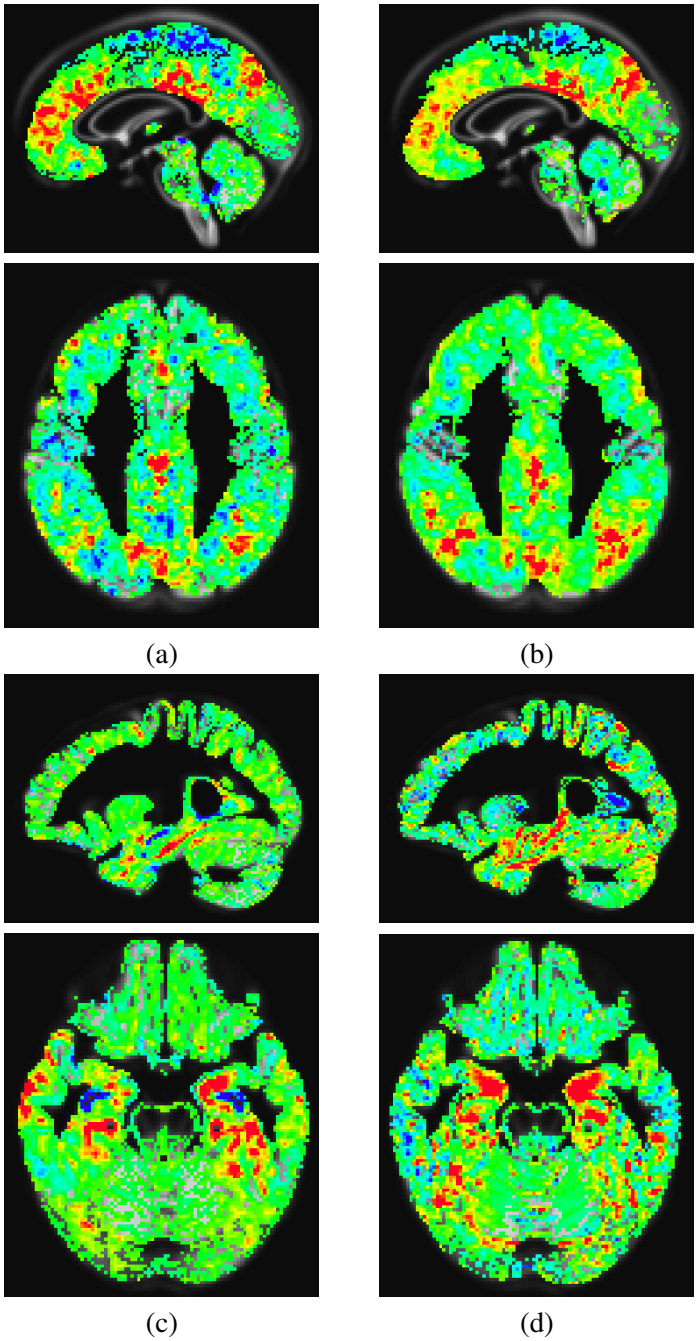


Figure 5.4: (a) Training error covariance $\mathbf{Q}$ matrix. (b–f) Least 5 eigen-vectors, represented as outer products.

MKL and other learning algorithms.



(a) (b) (c) (d)

Figure 5.5: Relevance maps in each modality: FDG-PET at baseline (a); FDG-PET at 2-year follow up (b); Tensor-based Morphology (TBM) (c); and Gray Matter density maps (d). In the FDG-PET-based modalities (a-b) we can see the posterior cingulate cortex and precuneus; in the MR-based modalities (c-d) we can see parahippocampal structures and CSF.

Liver

Regularizer	Accuracy	Sensitivity	Specificity	Area under ROC
Inv(Cov)	**0.717** $\pm$ 0.04	**0.530** $\pm$ 0.09	0.852 $\pm$0.05	**0.760** $\pm$ 0.01
Lap(Cov)	**0.722** $\pm$ 0.04	**0.530** $\pm$ 0.08	0.860 $\pm$0.05	**0.761** $\pm$ 0.01
Inv(Err)	**0.716** $\pm$ 0.04	**0.530** $\pm$ 0.07	0.851 $\pm$0.05	**0.760** $\pm$ 0.01
Lap(Err)	**0.719** $\pm$ 0.04	**0.529** $\pm$ 0.08	0.859 $\pm$0.04	**0.765** $\pm$ 0.01
$\mathbb{I}^{M\times M}$	0.701 $\pm$0.05	**0.518** $\pm$ 0.10	0.835 $\pm$0.06	0.750 $\pm$0.02
$\mathbf{1}^{M\times M}$	0.644 $\pm$0.07	0.271 $\pm$0.25	**0.913** $\pm$ 0.09	0.717 $\pm$0.05
Err	0.682 $\pm$0.04	0.477 $\pm$0.08	0.833 $\pm$0.06	0.731 $\pm$0.02

Pima

Regularizer	Accuracy	Sensitivity	Specificity	Area under ROC
Inv(Cov)	0.761 $\pm$0.03	0.871 $\pm$0.03	0.559 $\pm$0.06	0.821 $\pm$0.01
Lap(Cov)	**0.767** $\pm$ 0.03	0.878 $\pm$0.03	0.562 $\pm$0.06	0.822 $\pm$0.01
Inv(Err)	**0.764** $\pm$ 0.03	0.873 $\pm$0.03	**0.564** $\pm$ 0.06	0.822 $\pm$0.00
Lap(Err)	0.763 $\pm$0.03	0.877 $\pm$0.03	0.554 $\pm$0.06	0.823 $\pm$0.00
$\mathbb{I}^{M\times M}$	**0.766** $\pm$ 0.02	0.866 $\pm$0.03	**0.582** $\pm$ 0.06	0.821 $\pm$0.00
$\mathbf{1}^{M\times M}$	0.651 $\pm$0.03	**0.998** $\pm$ 0.02	0.007 $\pm$0.06	**0.993** $\pm$ 0.03
Err	**0.771** $\pm$ 0.02	0.890 $\pm$0.03	0.551 $\pm$0.05	0.829 $\pm$0.00

Ionosphere

Regularizer	Accuracy	Sensitivity	Specificity	Area under ROC
Inv(Cov)	0.939 $\pm$0.03	0.866 $\pm$0.06	**0.980** $\pm$ 0.02	**0.980** $\pm$ 0.01
Lap(Cov)	0.927 $\pm$0.06	0.829 $\pm$0.18	**0.982** $\pm$ 0.02	0.973 $\pm$0.02
Inv(Err)	0.940 $\pm$0.02	0.869 $\pm$0.06	**0.980** $\pm$ 0.02	0.981 $\pm$0.00
Lap(Err)	0.934 $\pm$0.05	0.845 $\pm$0.15	**0.982** $\pm$ 0.02	**0.976** $\pm$ 0.02
$\mathbb{I}^{M\times M}$	**0.948** $\pm$ 0.03	**0.887** $\pm$ 0.07	**0.982** $\pm$ 0.02	**0.982** $\pm$ 0.00
$\mathbf{1}^{M\times M}$	**0.940** $\pm$ 0.07	**0.866** $\pm$ 0.18	**0.983** $\pm$ 0.02	**0.974** $\pm$ 0.02
Err	**0.950** $\pm$ 0.02	**0.897** $\pm$ 0.05	**0.981** $\pm$ 0.02	**0.982** $\pm$ 0.00

WPBC

Regularizer	Accuracy	Sensitivity	Specificity	Area under ROC
Inv(Cov)	**0.759** $\pm$ 0.05	**0.025** $\pm$ 0.05	0.989 $\pm$0.02	0.625 $\pm$0.03
Lap(Cov)	**0.762** $\pm$ 0.06	**0.025** $\pm$ 0.05	0.992 $\pm$0.02	0.616 $\pm$0.03
Inv(Err)	**0.763** $\pm$ 0.04	**0.028** $\pm$ 0.05	0.992 $\pm$0.02	0.616 $\pm$0.03
Lap(Err)	**0.755** $\pm$ 0.06	**0.027** $\pm$ 0.05	0.984 $\pm$0.03	0.605 $\pm$0.03
$\mathbb{I}^{M\times M}$	**0.756** $\pm$ 0.06	**0.033** $\pm$ 0.06	0.982 $\pm$0.03	0.614 $\pm$0.03
$\mathbf{1}^{M\times M}$	**0.762** $\pm$ 0.05	0.000 $\pm$0.00	**0.999** $\pm$ 0.00	**0.957** $\pm$ 0.09
Err	**0.764** $\pm$ 0.05	0.006 $\pm$0.02	**1.000** $\pm$ 0.00	0.534 $\pm$0.04

Sonar

Regularizer	Accuracy	Sensitivity	Specificity	Area under ROC
Inv(Cov)	0.834 $\pm$0.05	0.789 $\pm$0.10	0.878 $\pm$0.08	0.917 $\pm$0.01
Lap(Cov)	0.816 $\pm$0.06	0.764 $\pm$0.16	0.863 $\pm$0.07	0.896 $\pm$0.04
Inv(Err)	0.783 $\pm$0.08	0.670 $\pm$0.18	0.899 $\pm$0.13	0.888 $\pm$0.03
Lap(Err)	0.817 $\pm$0.06	0.765 $\pm$0.13	0.869 $\pm$0.07	0.901 $\pm$0.04
$\mathbb{I}^{M\times M}$	0.836 $\pm$0.06	**0.805** $\pm$ 0.10	0.868 $\pm$0.08	0.923 $\pm$0.01
$\mathbf{1}^{M\times M}$	**0.858** $\pm$ 0.05	**0.817** $\pm$ 0.08	0.898 $\pm$0.07	**0.947** $\pm$ 0.01
Err	0.756 $\pm$0.08	0.523 $\pm$0.19	**0.973** $\pm$ 0.06	0.881 $\pm$0.03

Table 5.3: Performance measures for several $\mathbf{Q}$ functions on UCI datasets. Bold numerals indicate measures which are not significantly different from the maximum under a paired t-test. Lap = Laplacian; Inv = Inverse; Cov = Matrix covariance (Table 5.1, row 1); Err = Training error covariance.

Chapter 6

Machine Learning Approaches to Scientific Investigation of Alzheimer's Disease

---

By acquiring scans of cohorts of subjects undergoing a pathological condition of interest as well as healthy controls, scientists can examine in detail the effects of pathology by separating *individual variation* from *group-wise variation*. This way, if the groups are properly controlled then they will systematically differ only on the basis of disease, allowing hypotheses relating to disease processes to be tested. Traditionally this has been done by way of standard univariate models which can test whether means vary between groups relative to measures of their variances. More recent investigations have begun to move beyond the simple case by utilizing Statistical Parametric Mapping (SPM), in which a group statistic is computed at every voxel which can then be interpreted in terms of known anatomical and functional regions.

6.1 Predictive Multi-modality Markers of Neurodegeneration

Once classification with high accuracy has been demonstrated, the next task, and the more challenging one, is to predict which MCI subjects will progress to AD, and which ones will remain stable as MCI subjects. The methodology I adopted to explore the applicability of this approach is to first train a multi-modality classifier on AD and control subjects using MKL, and then compute the outputs of that classifier on MCI subjects. The output of this classifier is then called a Multi-Modality Disease Marker (MMDM). I then conducted a series of statistical analyses showing that while the accuracy of MMDMs in discriminating progressing MCI subjects from non-progressing is not as good as in the AD vs. control case, my results are nevertheless competitive with the state of the art. MMDMs based on baseline only (left) and longitudinal (right)

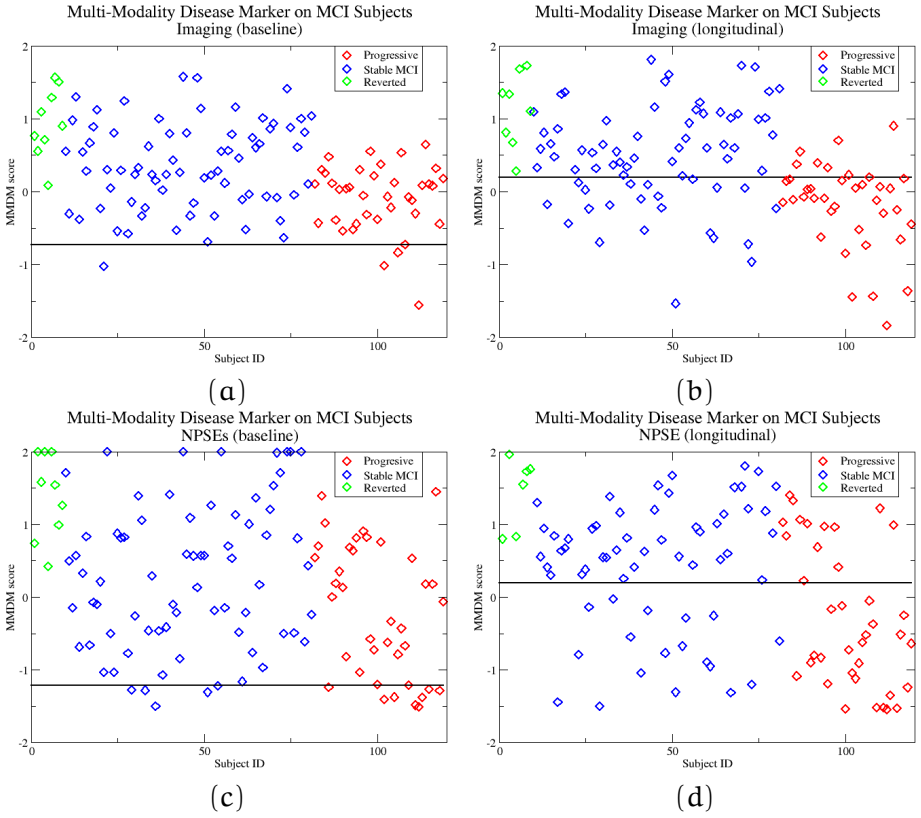


Figure 6.1: MMDMs applied to the MCI population. Subjects which remained stable are shown in blue; subjects which progressed to AD are shown in red; subjects which reverted to normal cognitive status are shown in green. In each figure, a line giving maximal leave-one-out accuracy is shown. Note that in some cases, the best accuracy can be achieved by simply labeling all subjects as the majority class. In some cases, MMDM scores were truncated to ± 2 so as to preserve the relative scales. On the left (a,c) are shown MMDMs based on information available at baseline. Note the homogeneity of the groups, leading to poor separability. Imaging-based MMDMs are shown at the top (a), while MMDMs based on NPSEs are shown below (c). On the right (b,d) are shown MMDMs based on all modalities available at 24 months. Note the improved separability between the progressing (red) and stable (blue) MCI subjects. Note that the imaging-based marker above (b) shows slightly greater separation of the 2 groups.

imaging data are shown in Figure 6.1. MMDMs based on neuropsychological scores are shown below for comparison. Note that while cognitive tests constitute a form of ground-truth (or at least, they are highly confounded with AD diagnosis, which is based on similar cognitive tests,) in the predictive setting of classifying progressing MCI subjects from non-progressing, they are less useful than imaging-based markers. Demonstrating this is an important goal of the ADNI.

6.2 Discovery of Anomalous Subjects

In the course of performing classification experiments it became apparent that most of the errors were coming from a small set of about 10% – 12% of the subjects, including both AD subjects and controls. We termed the group of anomalous subjects Group II, and the remainder, (*i.e.*, inliers,) as Group I. Note that these anomalous subjects included both AD cases and controls. (This became the motivation for developing outlier-robust methods described in Section 4.2.) Group-wise analyses showed some startling differences between all four groups: Group I AD / controls, and Group II AD / controls. For instance, Group II AD subjects had *greater* average total brain volume than the Group I control subjects. This is a truly startling result, considering that every AD subject in Group II scored below the dementia threshold on the MMSE and other diagnostic neuropsychiatric tests, yet, as a group, they showed more gray matter on average than even the healthy controls. Certain neuropsychological measures also correlated well with Group I/II status, which strongly suggests that AD diagnosis *in vivo*, while being highly reliable, is nevertheless imperfect. Clearly, in a small, but nevertheless significant fraction of subjects, there are important confounding factors which have an impact on analyses based on gray matter alone, and future studies may make use of this observation.

The rationale for conducting this analysis is that it is well known that AD-related neurodegenerative pathology is heterogeneous Thompson et al. [2001]. In addition, while the ADNI dataset is based on the most rigorous quality control protocol possible barring access to gold standard diagnostics such as biopsy or post-mortem analysis, there is some expectation that subjects will be misclassified. This may be because of the difficulty in distinguishing AD from other types of dementia such as Fronto-Temporal Dementia (FTD) or Lewy bodies Klöppel et al. [2008]. Further confounding the

situation is the possibility of comorbidity of AD with other neurodegenerative and neurovascular diseases such as stroke or multi-infarcts.

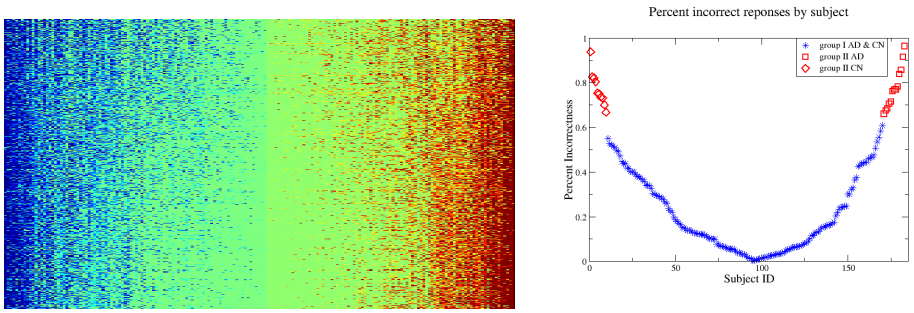
Identification of outlying participants.

The criterion I used in order to find this group was based on the extent to which the gray matter levels in disease-specific regions seemed to contradict the label given each subject, *i.e.*, AD or CN. In order to do this, I selected the 2000 most significant voxels in terms of p-values derived from a t-test, and examined the weak classifier predictive outputs on those voxels. (See Chapter 3 for a detailed description of the weak classifier methodology.) These outputs are shown in Figure 6.2 (a). Each column corresponds to a single example, and each row to a single weak classifier. The columns, i.e. subjects, are ordered from those having the most false negatives at the left, to those having the most false positives at the right. The color indicates the degree of incorrectness, with blue indicating false negative, green correct response, and red false positive response, respectively. We can clearly see that there are two "bars" at either end, consisting of subjects which are given the wrong label by nearly the entire set of weak classifiers. Subjects for which more than 65% of the weak classifiers gave incorrect outputs were placed in group II (Note that this closely matched the "bars" in Figure 6.2 (a)). This gave 10 controls, and 13 AD subjects. Figure 6.2 (b) shows the percentage of weak classifiers giving incorrect outputs on each subject. The labeling of anomalous subjects in this manner is not simply an artifact of the weak classifiers, but reveals a systemic pattern of deviation from the mean in each group. Evidence from hippocampus volume measures yields a similar labeling. (Hippocampus volume measures were computed by other groups, and were provided along with the ADNI data.) That is, the set of subjects more than one standard deviation away from the group mean, (of hippocampal volume), is almost identical to the set of examples placed in group II as above.

Characteristics of group II controls. I found that in several respects the group II controls were very similar to group I AD subjects.

- The first observation was that the group II controls had *significantly less* total brain volume, even relative to group I AD subjects: 8.8×10^5 (group II CN) [1] compared to 1.02×10^6 (group I CN) and 9.48×10^5 (group I AD) with

[1] Units are mm^3.



(a) (b)

Figure 6.2: (a) Weak classifier outputs for the 183 members of the MR population, ordered by the number of weak classifiers giving incorrect outputs. Each column corresponds to an individual subject, and each row corresponds to one of the 2000 selected voxels; columns are ordered by the number of weak classifiers giving incorrect outputs. Color indicates type and degree of incorrectness; blue corresponds to false negative, red to false positive, and green indicates correct response. Note the relatively sharp boundaries between the red and blue bands at either end – these are the members of group II. (b) Percent of weak classifiers giving incorrect responses for the same subjects.

p-values $< 10^{-9}$.

- *All* regions (where manual tracing-derived volume measures are provided in the ADNI dataset) were significantly smaller in group II controls compared to group I controls (p-values $< 10^{-3}$). Regional volumes for group II controls were closer to the respective measures from group I AD subjects.

- The ventricles in group II controls were *not* significantly smaller than controls in group I, which indicates that the above variations cannot be attributed to smaller brain sizes alone (and suggests possible atrophy).

- The hippocampal volume measures showed even larger variations in controls between groups I and II.

- VBM analysis between group II controls and group I AD subjects gave *no* discriminating regions and only isolated voxels.

Biomarker (AD subjects)	Group I	Group II	Z-test p-value
Mini-Mental State Exam (MMSE)	21.5 (3.04)	22.94 (2.84)	0.08
Tau-protein	111.94 (51.77)	151.88 (88.34)	0.0147
Logical Memory - Immediate Recall	3.13 (2.18)	4.91 (3.338)	$\sim 10^{-3}$
Logical Memory - Delayed Recall	0.48 (0.8)	3.13 (2.54)	$\sim 10^{-16}$
Boston Naming - Spontaneous Correct Responses	19.69 (6.95)	25.49 (4.70)	$\sim 10^{-3}$
Audio Visual	1.1 (1.08)	1.99 (2.15)	0.0374
Brain volume (UCSD)	948005.03 (84947.07)	1025001.3 (79868.99)	$\sim 10^{-3}$
L. Hippocampal volume (UCSD)	2706.69 (382.98)	3446.61 (573.23)	$\sim 10^{-10}$
R. Hippocampal volume (UCSD)	2813.38 (432.2)	3713.32 (368.21)	$\sim 10^{-12}$
L. Entorhinal cortex volume (UCSD)	2.44 (0.46)	3.03 (0.36)	$\sim 10^{-5}$
R. Entorhinal cortex volume (UCSD)	2.50 (0.46)	3.18 (0.42)	$\sim 10^{-7}$
L. Hippocampal volume (UCSF)	1518.45 (246.11)	1996.95 (426.44)	$\sim 10^{-10}$
R. Hippocampal volume (UCSF)	1498.39 (334.53)	2163.35 (341.04)	$\sim 10^{-14}$

Table 6.1: Comparison of relevant biomarkers in group I AD and group II AD. MMSE is included for reference; all other biomarkers listed are significantly different between groups at at least the 0.05 level.

- VBM analysis also revealed a significant gray matter density deterioration (p-values $< 10^{-6}$) in the hippocampus and parahippocampal gyri for group II controls, when compared to controls in group I.

Characteristics of group II AD subjects.

AD subjects in group II similarly resembled group I controls.

- The mean total brain volume of group II AD subjects was almost identical to that of group I controls ($\approx 1.02 \times 10^6$ in both groups). By comparison, the mean total brain volume of group I AD subjects was 9.48×10^5.

Biomarker (CN subjects)	Group I	Group II	Z-test p-value
Mini-Mental State Exam (MMSE)	28.98 (0.8)	29.19 (0.69)	0.33
Ventricles volume (UCSD)	38788.18 (23264.37)	40085.85 (13514.94)	0.84
Brain volume (UCSD)	1023746.53 (86217.87)	880452.33 (75572.03)	$\sim 10^{-9}$
L. Hippocampal volume (UCSD)	3599.87 (383.32)	3116.90 (301.58)	$\sim 10^{-5}$
R. Hippocampal volume (UCSD)	3791.06 (422.58)	3159.28 (359.84)	$\sim 10^{-7}$
L. Mid temporal volume (UCSD)	2.58 (0.17)	2.45 (0.12)	$\sim 10^{-3}$
R. Mid temporal volume (UCSD)	2.6 (0.20)	2.48 (0.21)	0.0454
L. Inf. temporal volume (UCSD)	2.64 (0.15)	2.49 (0.14)	$\sim 10^{-4}$
R. Inf. temporal volume (UCSD)	2.60 (0.19)	2.47 (0.25)	$\sim 10^{-2}$
L. Fusiform volume (UCSD)	2.39 (0.17)	2.25 (0.16)	$\sim 10^{-3}$
R. Fusiform volume (UCSD)	2.36 (0.17)	2.25 (0.18)	$\sim 10^{-2}$
L. Entorhinal cortex volume (UCSD)	3.19 (0.30)	2.86 (0.36)	$\sim 10^{-4}$
R. Entorhinal cortex volume (UCSD)	3.34 (0.32)	3.02 (0.51)	$\sim 10^{-4}$
L. Hippocampal volume (UCSF)	2126.69 (267.67)	1795.54 (208.3)	$\sim 10^{-5}$
R. Hippocampal volume (UCSF)	2176.57 (275.65)	1781.65 (252.45)	$\sim 10^{-7}$

Table 6.2: Comparison of relevant biomarkers in group I CN and group II CN. MMSE is included for reference; all other biomarkers listed are significantly different between groups at at least the 0.05 level.

- In the hippocampus and entorhinal cortex the mean volume among group II AD subjects was nearly the same as that of group I controls: 7159.93 (UCSD) in group II AD subjects versus 7390.93 (UCSD) in group I controls for the hippocampus. By comparison, the same measures were 5520.07 (UCSD) in group I AD subjects. The mean entorhinal cortex volumes had a similar proportion.

- VBM analysis showed greater gray matter densities in the hippocampus for group II AD subjects compared to group I AD and hypertrophy in the thalamus relative to group I controls.

Cognitive status. While the image based biomarkers showed significant variations between groups I and II, the associated cognitive status and neuropsychological scores (e.g., MMSE) were relatively consistent. This is not surprising because cognitive status, especially the MMSE score, is used in clinical diagnosis. However, Group II AD subjects did show relatively small, but nevertheless significant group differences in tests measuring logical memory – both immediate and delayed recall, number of spontaneous correct responses given on the Boston Naming Test, and audio visual tests. In all of these, group II AD subjects scored higher indicating slightly healthier cognitive status (consistent with lower observed atrophy in the preceding discussion). Of these, the delayed recall was the most significantly different (p-value ≈ 0). There was no significant difference between the performance of group I and group II controls on any measure of cognitive status. Summaries of volume measures significantly differing between both groups I and II are presented in Tables 6.1 and 6.2.

Summary. It is important to note that confirmed diagnosis of AD is only possible post-mortem. Given the clinical nature of the ADNI data set, it is possible that some AD subjects in the cohort may have another form of dementia, or possibly depression, while some controls may have AD in the early stages, and have not yet begun showing signs of cognitive decline. The classification algorithm, however, assumes that every label in the training data is correct, and therefore tries to correctly classify every training example. In the presence of incorrectly labeled examples, however, it is difficult for a method to have a lower expected error rate than the fraction of mislabeled examples in the training set. Clearly, if our data set contains mislabeled examples [Wade et al., 1987, Schofield et al., 1995, Burns et al., 1990], an automated method may not be able to to outperform this limitation. I therefore developed the outlier-robust version of MKL presented in Chapter 4 in response to this issue. Looking to the future, characterizing this set will be useful for not only improving the accuracy of classification systems evaluated on this dataset, but may also suggest ways that the classifier can be modified to automatically handle them. The analysis above, and evaluations of classifier's performance with/without including group II have the potential to be a useful first step

in discovering mislabeled subjects that may not have been identified by the study's strict quality control protocols.

Outliers detected by MKL methodology

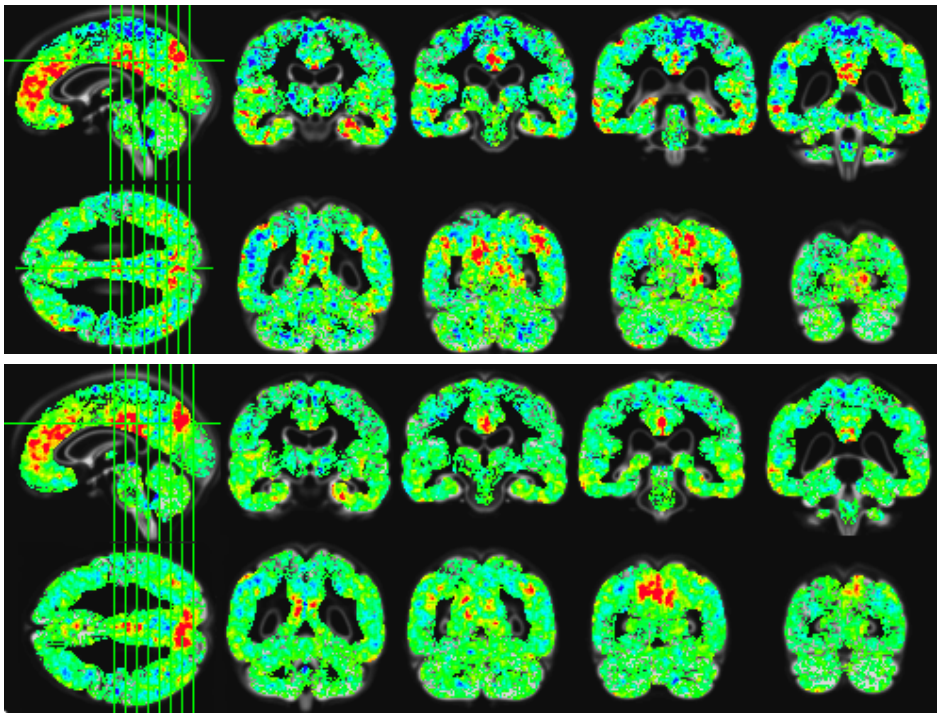


Figure 6.3: Voxel weights assigned by the MKL classifier (using only linear kernels) for FDG-PET baseline images. Top: Voxel weights with all subjects. Bottom: Voxel weights when the outlier subjects were removed. Note there are significant negative (blue) weights in heteromodal, frontal, parietal regions and temporal lobes on the top, but that these regions largely disappear when the outlier subjects are removed (bottom), giving weight patterns more consistent with the AD literature.

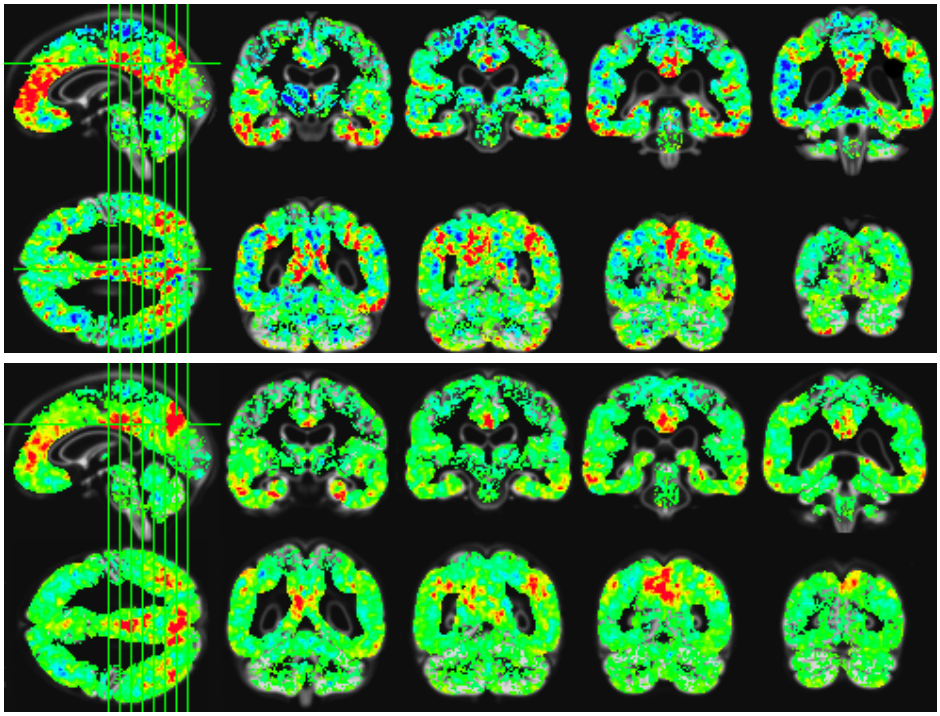


Figure 6.4: Voxel weights assigned by the MKL classifier (using only linear kernels) for FDG-PET images at 24 months. Top: Voxel weights with all subjects. Bottom: Voxel weights when the outlier subjects were removed. Note there are significant negative (blue) weights in heteromodal, frontal, parietal regions and temporal lobes on the top, but that these regions largely disappear when the outlier subjects are removed (bottom), giving weight patterns more consistent with the AD literature.

In [Hinrichs et al., 2011] I also performed an examination of the brain regions used in the MKL classifier. Note that this requires using linear kernels only, because non-linear kernels do not have an exact representation as a set of voxel (or feature) weights. On examination of these brain regions, several regions which should be associated with health, (*i.e.*, positive weights should be given to these regions, meaning that more gray matter or FDG-PET signal is indicative of a healthy subject,) were in fact being given negative weights. This indicates that there was, with high probability, there was a small set of AD cases who appeared healthy in these regions, but who were

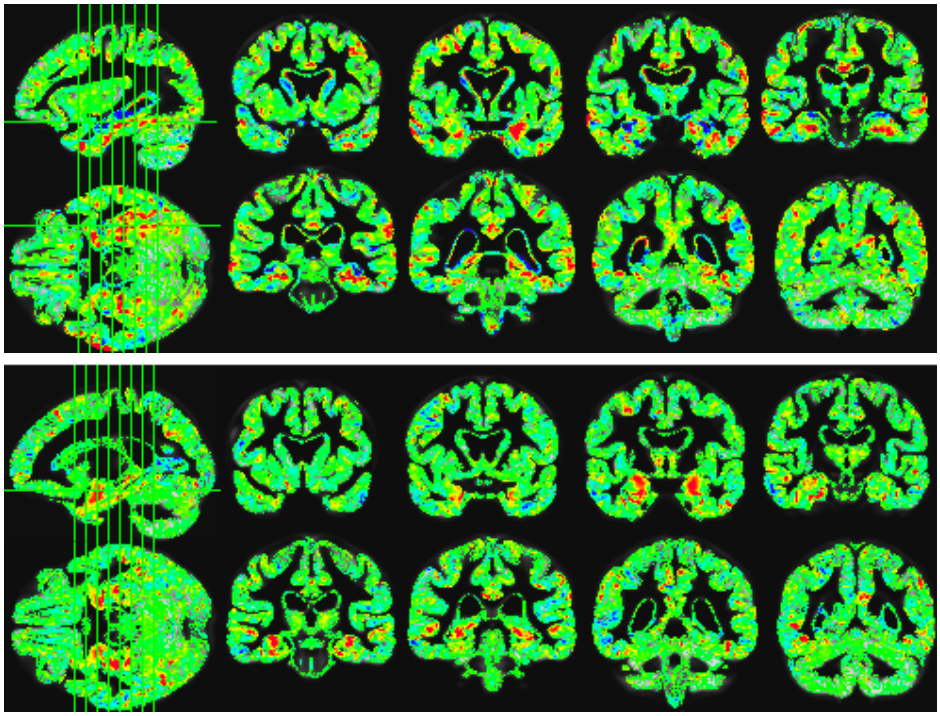


Figure 6.5: Voxel weights assigned by the MKL classifier (using only linear kernels) for GM density images at baseline. Top: Voxel weights with all subjects. Bottom: Voxel weights when the outlier subjects were removed. We again see perplexing negative weights in and around the hippocampus and surrounding regions which disappear when the outlier subjects are removed.

causing the algorithm to associate health in these regions with AD status. Likewise, a set of control subjects showing hypometabolism in those regions would have a similar effect. By selecting a small group of AD subjects who showed the greatest FDG-PET intensity in these regions which were given negative weights and retraining the classifier, I was able to suppress these artifacts, and generate more reasonable regions. The analysis of FDG-PET images yielded 5 outlier subjects, while analysis of GM density images yielded a further 4 subjects. Side-by-side comparisons are shown in Figures 6.3 through 6.6.

Starting with Figure 6.3, which shows the pattern of voxel weights for baseline FDG-PET images, we can see blue (negative) weights being assigned to heteromodal,

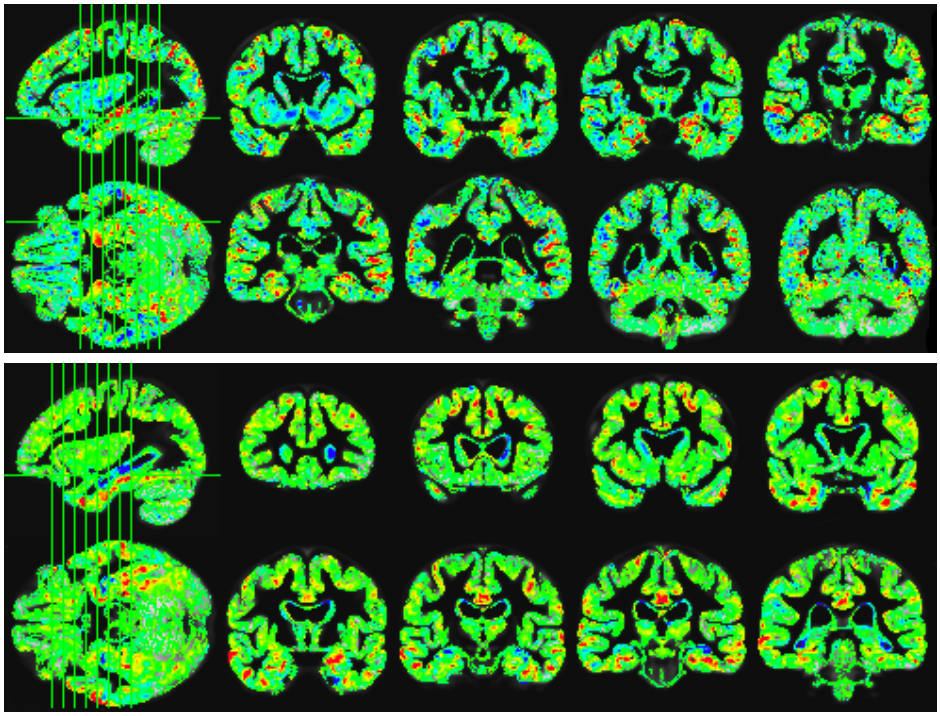


Figure 6.6: Voxel weights assigned by the MKL classifier (using only linear kernels) for TBM images. Top: Voxel weights with all subjects. Bottom: Voxel weights when the outlier subjects were removed. Comparing top with bottom, there is also an overall reduction of negative weights in gray matter regions as well as a concentration of negative weights to areas surrounding the ventricles, and CSF bordering the hippocampus and surrounding regions, when the outliers are removed.

frontal, parietal regions and temporal lobes, which is largely in conflict with the existing AD neuroimaging literature, as higher FDG-PET signal in these regions should be indicative of healthy status. As expected, when the outlier subjects were removed the selected regions assumed a more conventional pattern for AD, as shown in the bottom. A similar effect was observed in the FDG-PET images taken at 24 months, as shown in Figure 6.4. It should be noted as well that in FDG-PET images, there should in general be *no* strongly negative weights, as there is no known association with *hyper*metabolism in any brain region, and AD. Continuing to Figure 6.5, we again see troubling blue (negative) weights in and around the hippocampus – observe in

particular the second and third coronal slices from the right in the upper row, in both the top and bottom figures. On the bottom set of coronal slices, we see exactly the pattern of hippocampal health/atrophy which we should expect, especially in the second coronal slice from the right in the top row. Yet, in the top figure, there is a mix of red and blue weights which again defies intuition – there are no known hypertrophic effects of AD on gray matter, especially not in the vicinity of the hippocampus. Moreover, the *pattern* of weights in the top part of the figure does not resemble known AD patterns. Finally, in Figure 6.6, we see similar effects in the vicinity of the hippocampus in voxels which contain primarily gray matter. As TBM is a measure both of expansion and contraction, we should expect to see blue (negative) weights at the boundary between gray matter tissues and CSF, because atrophy in the hippocampus manifests as a retreat in the GM/CSF boundary. Thus, CSF voxels at that boundary will be seen to expand the most dramatically. Note also that in the bottom image in Figure 6.6, the pattern consists largely of smooth, contiguous regions of both positive and negative weights, which largely match the expected pattern of AD atrophy. Yet, in the top image, (that in which outliers were included,) the pattern is more varied, and (interpreting subjectively,) has a much stronger high spatial frequency component. This phenomenon can also be seen in Figure 6.5. We can interpret this pattern as the result of the algorithm searching futilely for a function which separates classes of objects which are inherently more difficult to separate, (because of the inclusion of outliers,) and is forced to rely more on sampling artifacts and extraneous signal.

Clearly, this is a *post-hoc* analysis, and so we must exercise caution in interpreting these results – given that the outliers were selected on the basis of whether they agree with predefined notions of how the classifier should look, then of course the classifier will be as expected when we remove them. Nevertheless, there are several remarks which can be made about these results. First, note the dramatic effect on the classifier's pattern that a few subjects can have. This is because a discriminative classification model does not attempt to model the distribution from which examples are drawn, but instead looks only to find a function which separates them. Also, Support Vector-based algorithms, (of which MKL is an example,) construct the separating hyperplane normal as a sparse combination of only difficult to classify examples, which adds to the sensitivity to outliers. Second, note that while the selection is made on the basis of features disagreeing with the target labels, in all four Figures the resulting classifier

after removing the outlier subjects is much more spatially smooth, which we have already shown to be an indicator of a genuine neurological effect, and not merely a sampling artifact.

These investigations formed a sidebar of the larger investigations of AD classification mechanisms, but they nevertheless highlight a way in which machine learning methods can potentially aid in scientific discovery, by identifying sub-populations in need of closer examination.

6.3 Clinical Trial Enrichment

As mentioned in section 2.3, a major goal of the ADNI is to encourage translational uses of AD research, particularly advances involving neuroimaging based markers. One possible avenue of inquiry, and a topic of much recent interest [Kohannim et al., 2010, Hua et al., 2009, 2010], is in reducing the sample sizes required for clinical trials of proposed treatments for AD. In this section I will discuss two avenues for approaching this problem: Multi-modality eligibility criteria based on MKL, and, custom outcome measures based on discriminative SVM models. These methods are appropriate for use in traditional clinical trial models, and serve as an enhancement, or extension. In Chapter 7 I discuss an entirely novel approach to clinical trial design, and statistical power analyses.

As methods of discriminating subjects who *already have AD* from controls have become more accurate, more recent efforts focus on the more difficult problem of discriminating MCI subjects from controls [Davatzikos et al., 2009, Querbes et al., 2009] and the hardest of all tasks – discriminating which MCI subjects will *progress* to AD [Hinrichs et al., 2011, Zhang et al., 2011a] and the patients that will remain stable (*i.e.*, they do not have pre-dementia of the Alzheimer's type). As described in Section 6.1, multi-modality learning methods such as MKL can provide highly predictive markers of which subjects will progress to AD. This issue is particularly relevant because the inclusion in clinical trials of a large subgroup of subjects who are *non*-converters, leads to significant heterogeneity. It is a serious problem in clinical trials which seek to assess the effects of a treatment on a homogeneous cohort, so as to to maximize the chance of detecting a statistically significant variation in how the placebo and treatment groups respond to the treatment.

In this work I considered two ways in which these predictive markers can *increase* sensitivity of clinical trials – thereby reducing the number of subjects required to detect a desired effect size. Key strategies are sample enrichment and custom outcome measures. First, consider the presence of a large number of MCI subjects who will *not* progress to AD (only 10–15% of MCI subjects convert annually). Even if the treatment under study is effective, it will have little or no measurable effect on subjects who do not suffer from the disease. For example, Visser et al. [Visser et al., 2005] suspected that a number of AD trials that could not identify significant effects of the treatment may have failed due to inclusion of non-AD MCI patients. In the absence of sensitive measures of change at milder degrees of impairment to identify such patients, a trial may potentially require a very large cohort to account for the variability, which may not always be feasible. Here, I present evidence that by excluding subjects whose predictive markers are not indicative of future decline, we can *enrich* the sample population – such an enriched cohort reduces the masking effect of non-progressing subjects. Second, consider the difficulty of using cognitive markers as an outcome measure – to test whether a treatment is effective, a common practice is to measure changes over time in various neuropsychological status measures such as the Mini-Mental State Exam (MMSE) [Petersen et al., 2005]. Unfortunately, such measures are subject to a large amount of inter- and intra-subject variation, and can change slowly over time. Thus, the use of such markers can result in large study cohorts, while recent results [Kohannim et al., 2010, Hua et al., 2009] have shown that with imaging-based outcome measures cohort sizes can be greatly reduced – by up to a factor of 8 [Hua et al., 2009]. Experimental results suggest it may be possible to move beyond these studies by using a *predictive* marker based on Multi Kernel Learning (MKL) methods rather than *summary statistics of atrophy* over entire Regions of Interest (ROIs).

Power calculation

The first step in designing a clinical trial is to determine an outcome measure most likely to vary as a function of the administered treatment. The second question is, the number of subjects (sample size) we need to recruit to observe (*e.g.*at 80% power) the induced variations in the outcome measure. This calculation is transparent to the actual treatment under study, and is fully determined by the *variance* and *effect size*

(difference in signal between placebo / treatment groups).

By performing a two-sample t-test on outcome measures taken from the two trial groups, we are comparing the separation of two sample means in terms of a Gaussian-distributed variable which represents the null hypothesis. Let δ be the difference between outcome means in the two populations, and σ^2 be the pooled variance of the outcome measure. The test statistic t is the ratio of the effect size δ to sample variance σ, $\frac{\delta}{\sigma\sqrt{2n}}$. For a desired Type I error rate of α, and Type II error rate of β, the requisite sample size can be calculated as:

$$n = \frac{2\sigma^2(t_{1-\alpha/2,n-1} + t_{1-\beta,n-1})^2}{(\delta\lambda)^2}, \tag{6.1}$$

where λ denotes the desired percentage of reduction in outcome measure [Hua et al., 2009]. Thus, if we desire a 25% reduction in atrophy (if atrophy is used as the outcome measure), then we set $\lambda = 0.25$.

Experimental Design

Summary

Experiments to assess the efficacy of the enrichment procedure above were conducted on an extensive dataset of different image types, cerebrospinal fluid measures, and cognitive scores acquired as part of ADNI. The goal was to calculate sample sizes required in a hypothetical placebo-controlled parallel clinical trial to observe a certain reduction in rate of atrophy (outcome measure) at a given power. To highlight potential gains, we provide power calculations both with and without the new inclusion criteria. Note that these calculations depend only on the mean and variance of the outcome measure (among the selected cohort).

Dataset and pre-processing

The data and preprocessing steps used in this work are the same as in [Hinrichs et al., 2011], and are described in 4.1. Voxel-Based Morphometry (VBM) and Tensor-Based Morphometry (TBM) processing pipelines were applied to MR data to extract baseline Gray Matter density and Jacobian Determinant maps. FDG-PET scans from baseline

and at 24 months were also included, for a total of four groups of images – which provided the kernels used in our MKL model (for the MKL-IC measure). TBM maps were used to compute outcome measures of interest (atrophy).

AD and control subjects were used for training the classifier, (*i.e.*, learning the disease pattern), and for feature selection (*i.e.*, for selecting Regions of Interest (ROI)). We then computed the classifier's output which provided the desired MKL Inclusion Criterion (MKL-IC). We rejected the 75% of subjects whose MKL-IC was least indicative of an AD-like pattern of atrophy, and preserved the remainder. This choice reflects a trade-off between boosting the power of a study, without requiring too many subjects to be screened only to be subsequently rejected.

Why TBM-derived outcome measures?

TBM has been shown to have excellent characteristics as an outcome measure [Hua et al., 2009, Kohannim et al., 2010] because by quantifying deformation between successive time points, it serves as a surrogate for atrophy. Since this is precisely the measure we expect to show variations in AD (as a result of the treatment), calculating sample sizes using TBM on our enriched sample is a reasonable assessment of its utility. For our evaluations, we computed t-statistics from each voxel using the AD and control population *only*, and then thresholded the voxels at $\mathrm{p} < 0.05$. A natural question is whether TBM can be used both for learning the MKL-IC (albeit from AD and controls) *and* as an outcome measure for the MCI group. Making this choice is similar to the common practice of using hippocampal volume measures as covariates and atrophy as an outcome measure [Schott et al., 2010]. However, if desired (and in the interest of being more conservative), one may prefer not to use such measures in both MKL-IC and in outcome measures. We will discuss both results in the following section.

Custom TBM measure

As in [Kohannim et al., 2010, Hua et al., 2009], we used mean TBM values in the chosen ROI, and explored the question of whether a *weighted* average over a ROI can also produce an informative outcome measure. We trained an SVM classifier using these voxels as features; (we first converted raw TBM determinants to annual rates of

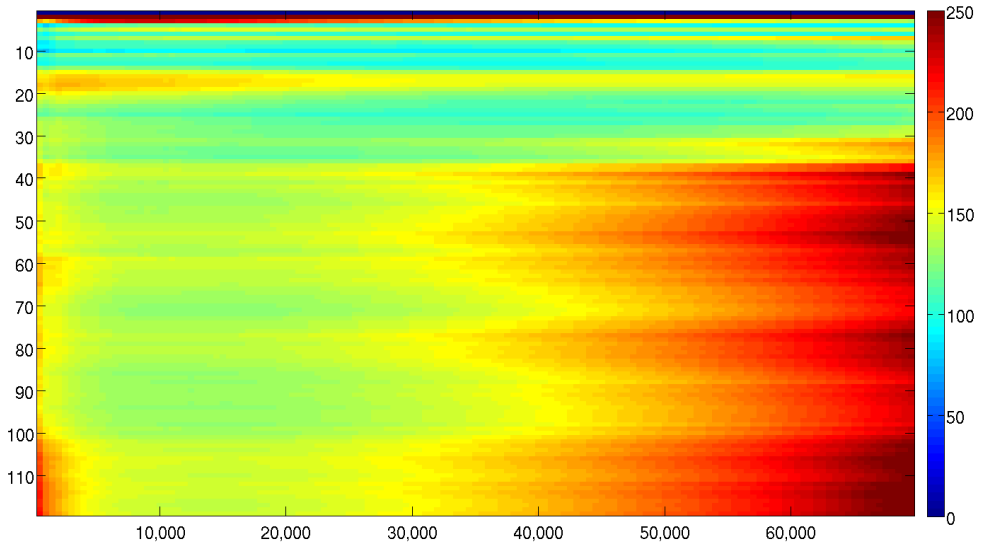


Figure 6.7: Sample cohort sizes as a function of number of TBM voxels (x-axis), and number of MCI subjects (y-axis).

change, and then negated voxels with negative mean change so that the SVM weights came out all positive). As a post-processing step, we normalized the SVM weights to sum to 1, as is the case when using a mean TBM value over an ROI. This outcome measure is designed to be as sensitive as possible not only to *global* gray matter loss, but specifically to atrophy *relative to a discriminative disease pattern*. We provide a brief intuition on the role of the number of voxels used, as well as the trade-off between screening out subjects for greater enrichment, versus controlling the number of subjects at screening.

Exploratory analysis

In addition to estimating required sample sizes for fixed parameter values (25% exclusion, TBM voxels with $p < 0.05$ used in computing the outcome measures), we also computed a map of sample cohort sizes for a range of voxel selection thresholds and number of subjects included (and excluded). (See Figure 6.7). This was exploratory in nature, and allows us to examine *qualitatively* (for this particular dataset), the choices available. Note the decreasing trend as the inclusion criteria become more strict (*i.e.*,

Outcome measure	Mean TBM			Custom-SVM	ADAS-Cog	MMSE	Schott et al. [2010]
Inclusion Criterion	**MKL-IC TBM**	**MKL-IC No TBM**	**MKL-IC Baseline**	**MKL-IC Baseline**	–	–	–
Power							
0.80	71	90	166	**88**	1,023	1,557	122
0.85	80	103	190	**100**	1,170	1,781	–
0.90	94	121	222	**117**	1,369	2,084	–

Table 6.3: Estimated sample cohort sizes for single modal and multimodal inclusion criteria. Non-TBM derived MKL-IC are shown in column 1. TBM / NO TBM refers to whether TBM-derived kernels were used in computing the MKL-IC. "Baseline" MKL-IC was derived *only* from data available at Month 0. Custom SVM is an SVM-derived outcome measure (weighted average over ROI). Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-Cog), mini-mental state examination (MMSE), and sum-of-boxes Clinical Dementia Rating (CDR-SB) outcomes are shown for comparison.

excluding more stable MCI subjects), highlighting the value of sample enrichment for improving detection of effects on atrophy.

Results and Discussion

Table 6.3 presents the main results of these evaluations. The primary concern is the number of subjects needed (per arm of a clinical trial) to detect a 25% reduction in atrophy as a result of treatment. Using more traditional clinical and cognitive single measures, as listed in the rightmost three columns, would require anywhere from 600 to over 2000 subjects *per arm* to detect the desired treatment effect (with power from 80% to 90%, type I error rate of 0.05). On the other hand, by using enriched samples and imaging-based outcomes, dramatic reductions are achievable. Even without using *any* longitudinal inclusion criteria, we can reduce sample sizes by a factor of 5 to 10. We also see that modest improvements can result from using an SVM-derived weighted statistical ROI, suggesting that even higher gains in sensitivity and power are possible with further development. If one uses longitudinal data, excluding TBM measures, still further improvements are possible. These results compared favorably

to recently reported findings [Schott et al., 2010, Hua et al., 2009, Kohannim et al., 2010] – note that this study uses *only* MCI participants, which are a more challenging group because atrophy effects are smaller, and variances are greater.

Chapter 7

Linear Outcome Measures in Clinical Trials

Following the evolution from previous chapters, I have described how machine learning methods can be adapted to better capitalize on spatial regularity and multiple modalities in neuroimaging analysis, with a particular view towards making predictions about the disease course of individual patients. The next step is to apply these predictions in designing clinical trials. Clinical trials are the most logical path towards making these developments translational for several reasons. First, there is no known cure, and the few treatments that are currently available are only effective at delaying the onset of AD for a short period. Thus, there is little benefit to having greater certainty about a patient's long term prognosis without a way of turning that information into more effective treatment options. Second, and perhaps more importantly, many pharmaceuticals and other treatments (*e.g.*, cognitive training, or targeted exercise programs,) are under development, yet, without the right statistical tools it may be very difficult to separate genuinely effective treatments from those which are not. Recall that while late-stage AD has a devastating and unmistakable impact on patients, in its early stages it is frustratingly difficult to make a certain prognosis. Tests of outward neuropsychological decline are limited in their effectiveness by the large degree of inter-subject variability, but also by a large *intra*-subject variability as well – *i.e.*, it is quite conceivable that a particular subject may perform at one level on a test on one day, and yet, for a variety of reasons, perform at another level on another day. Together, these forms of variability mean that it is often difficult to detect significant alterations in cognitive functioning with high confidence.

Neuroimaging offers a way of managing and mitigating these effects – though scanners do add a small amount of their own noise and variability to the data, technological development can reduce this burden; yet, there is *very* little intra-subject variability in neural tissues, (at least macroscopically,) except on decadal time-spans. Meanwhile, there *is* a large degree of inter-subject variability, but again this variability can be suppressed with the proper registration and normalization methods, ideally

preserving variation due to the disease itself. As mentioned in the previous chapter, several studies Kohannim et al. [2010], Hua et al. [2009] have recently shown that TBM and other measures of neural atrophy suffer far less from the kinds of variability described here.

In this Chapter I give an introduction to a novel clinical trial methodology, discuss ways of assigning significance levels to its results, and present detailed simulations which strongly suggest that it may be more effective than existing methodologies. This methodology is motivated by the observation that parametric univariate tests, such as the t-test used in most clinical trial methodologies, while optimal for detecting differences between groups in one dimension, do not make the best use of high dimensional data. That is, if the outcome measure is taken as the sum or mean of a sufficiently large number of *independent* covariates, (as is the case when we take the mean voxel intensity over a Region of Interest (ROI) to be the outcome measure,) then it is guaranteed to closely approximate a Gaussian distribution, in which the t-test is indeed the optimal measure of significance. However, aggregating voxels by taking a mean can actually obscure signal as much as enhance it. Consider that many linear learning models are designed to choose, with extreme care, a linear combination of weights which best amplifies a target signal hidden among a large number of covariates. The success and wide adoption of linear learning methods is indicative of just how much a difference can be made by the right choice of weights in such situations. As high-throughput data acquisition and processing technologies become more and more commonplace, this issue is likely to become a limiting factor in the sensitivity and statistical power of clinical trials. Yet, as I will show, the real issue is not so much the high dimensionality of this data, as it is the complex correlations and dependencies between covariates. Simulations presented below demonstrate that so long as covariates are (conditionally) independent, then the problem is effectively trivial, and solved: a simple t-test on the average is optimal. When, on the other hand, there are strong dependencies between covariates, the situation is very different. The principal insight that I intend to develop in this Chapter is that linear learning methods, especially the Support Vector Machine (SVM) model, are far better suited to this statistical task – and that there is are significant gains potentially waiting to be accrued in the sensitivity and power of clinical trial designs using neuroimaging data as primary end-points. Monte-Carlo simulation results presented below strongly

suggest that the proposed method has the potential to improve statistical sensitivity over naïve methods in neuroimaging trials, and the same train of reasoning applies to other settings as well.

In order to better account account for such highly correlated data, my proposed framework rephrases the question in terms of classifiability. That is, we can rephrase the question by asking whether there exists a linear classifier which can discriminate between participants who received the treatment, and those who did not. The key difference is that we include the issue of how to aggregate the data down to one dimension in the overall question of assessing the significance of observed differences, rather than first assuming a particular strategy, and only then assessing significance. We can then use the cross-validated accuracy of such a classifier as the test statistic, and calculate p-values by comparing the learned model's cross-validated accuracy against the Null hypothesis, i.e., the Binomial distribution with a 50% probability. Observing that a linear combination of Gaussian random variables is itself guaranteed to be Gaussian, then, if we can make the assumption that each covariate is individually Gaussian distributed, then we can be certain that the linear classifier's output will also be Gaussian. In such situations, it would then be ideal to use a t-test on the classifier's output to assess significance. However, in the absence of such an assumption, then the Binomial test may be preferable.

The t-test on unweighted averages, and SVM methodologies can be viewed as occupying opposite ends of a continuum. At one end, when voxels (or covariates, more generally,) are completely independent, then a t-test on an unweighted is indeed optimal for detecting differences in mean voxel intensities. At the other end of this continuum is the case where voxels are highly, (but not completely,) correlated. This means that the samples are effectively drawn from a approximately low-dimensional subspace, in which case the SVM is more effective. (Having highly correlated data essentially means that the covariance matrix has a few large eigen-values, and the rest quickly decay to near-negligible values. See Figures 7.2 and 7.3.) A simple diagram illustrating the intuition behind this phenomenon is given in Figure 7.1.

In the following, I first discuss statistical concepts employed in standard clinical trial methodologies so as to fully detail my motivation, before moving to a discussion of statistical power calculations for the proposed method. Subsequently, I report on Monte Carlo simulations which clearly demonstrate the need for methods which

are tailored to high-dimensional, highly-correlated data, such as neuroimaging, and potentially micro-array, GWAS, or other high-dimensional high-correlation data.

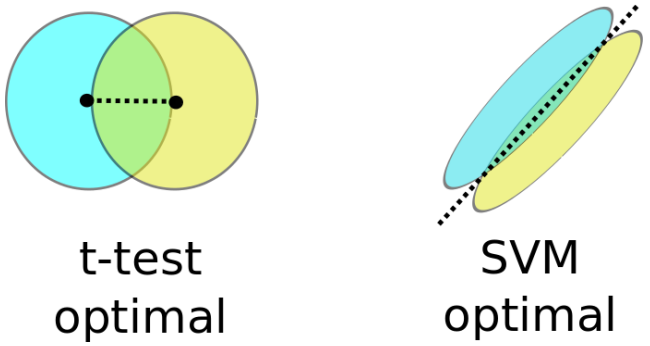


Figure 7.1: Two cases which illustrate, in a low-dimensional setting, the relative strengths of each method.

7.1 Outcome Measures, and Related Statistical Concepts

First, I define an outcome measure, and related concepts of statistical power, with particular emphasis on the neuroimaging setting. Suppose we are given a sample cohort of *trial* and *placebo* participants, which we denote as $X_{treatment}$ and $X_{placebo} \in \mathcal{X}$, where $\mathcal{X}$ is the space from which sample subjects are drawn, with some probability distribution $P(\mathcal{X})$. An outcome measure τ is a mapping $\tau : \mathcal{X} \to \mathbb{R}$, which implies that $P(\mathcal{X})$ will specify a distribution in $\mathbb{R}$ as well. This crucial observation allows us to frame the question of treatment effectiveness in terms of the distribution of $\tau(X_{treatment})$ as compared to that of $\tau(X_{placebo})$. That is, we are primarily interested in knowing whether the distributions $P(\tau(X_{treatment}))$ and $P(\tau(X_{placebo}))$ differ. Testing whether this is the case can be done by calculating sample statistics of $\tau(X)$, such as

$$t = \frac{|\mu_{trial} - \mu_{placebo}|}{\sigma\mathbb{Z}}$$

where σ is a pooled estimate of the standard deviation of $\tau(X)$ using both sample groups, μ_{trial} and $\mu_{placebo}$ are the sample means of the trial and placebo groups $\tau(X_{treatment})$ and $\tau(X_{placebo})$, and $\mathbb{Z}$ is a normalization constant. By comparing t

to a threshold δ according to a reference distribution we can test whether the means of the two groups significantly differ. Even more importantly, we can assign levels of confidence to the outcome of this comparison. That is, if the underlying distribution of trial subjects is such that $|\mu_{trial} - \mu_{placebo}| = 0$ then there exists some α such that $t < \delta$ with probability $1 - \alpha$; likewise, if $|\mu_{trial} - \mu_{placebo}| \neq 0$ then there exists some β such that $t > \delta$ with probability $1 - \beta$, which is the statistical power, or, the ability of this methodology to detect a real difference between sample groups $X_{treatment}$ and $X_{placebo}$ as measured by the outcome measure $\tau(X)$ and sample statistic t.

Traditionally, the t-statistic is interpreted as a measure of likelihood that two sample distributions have different means, as measured by the ratio of separation of means to average deviation; more broadly, we can interpret t as an inverse measure of the overlap between these two distributions. That is, large t-statistics imply that there is little overlap in the distribution functions under consideration. In the case of parametric, isotropic, unimodal distributions such as the Gaussian, it is sufficient to consider the displacements of means to determine whether distribution functions differ significantly, but for high dimensional, non-parametric distributions, we may wish to approach the question of distribution overlap more directly, in a more general sense which does not rely on Gaussian behaviors to induce separability. Note that it is common to assume that voxel intensities are Gaussian distributed, and that entire images can be thought of as coming from a multivariate Gaussian distribution.

Linear classification is where we attempt to fit a discriminating hyperplane between two groups of points in a (potentially high dimensional) space. While high dimensionality is not generally conducive to high accuracy in linear classifiers (the VC dimension of linear classifiers is proportional to the number of independent dimensions,) the high degree of correlation among voxels counteracts this effect by effectively reducing the dimensionality of the space. This makes linear classifiers well suited to neuroimaging classification tasks. A hyperplane is defined by its normal $\mathbf{w}$ and offset b, and the corresponding decision function is defined as $\text{sign}(\mathbf{w}^T x + b)$. In order to adapt this methodology to the problem of choosing an outcome measure, we can simply let $\tau(x) = \text{sign}(\mathbf{w}^T x + b)$, where $\mathbf{w}$ and b are chosen without observing x. The sample statistic then becomes $t^* = \frac{1}{N} \sum_i y_i \tau(x_i)$, where there are N subjects overall, and $y_i = 1$ for all trial subjects, and $y_i = -1$ for all placebo subjects. In other words, t^*

is the cross-validated empirical risk (test-set accuracy) of a linear classifier with trial group as the predicted labels. Note that by writing

$$t^* = \frac{1}{N_{trial}} \sum_{i:y_i=1} \tau(x_i) - \frac{1}{N_{placebo}} \sum_{i:y_i=-1} \tau(x_i)$$

we can see the role of t^* as a difference of means of two binomials, with the variances defined according to the error probabilities. (Under the Null hypothesis, with equal sized cohorts, the error probability is 0.5.) As a result of the process by which it is selected, (i.e., the training step,) the outcome measure τ is chosen so that the overlap between the distributions of $\tau(X_{treatment})$ and $\tau(X_{placebo})$ is minimized, leading to a higher t^*, but only in the event that the groups are indeed separable. What remains is to assign confidence levels α and β for a given threshold δ, which I discuss next.

7.2 A Motivating Example

In order to better motivate the remainder of this chapter, I first present a Monte-Carlo simulation which demonstrates the effect had by a dense, (*i.e.*, approximately low-rank) covariance matrix, as opposed to an isotropic (*i.e.*, diagonal, and high-rank) one. In particular, this first set of experiments examined whether or not this is sufficient to cause the univariate t-test on the unweighted mean voxel intensity to become underpowered relative to the SVM in a multivariate Gaussian setting. As we will see, this is indeed the case. In this experiment I used MCI participant data to compute the sample covariance matrix, Σ_{MCI}, from 1000 TBM voxels. See Figure 7.2 for the covariance matrix, and Figure 7.3 for its eigen-values. These voxels were selected by computing voxel-wise t-statistics between AD and CN groups, and choosing the lowest 1000. I modeled the simulated disease effect as the voxel-wise difference in means between the AD and CN groups, denoted as $\delta_{disease}$. To simulate the untreated disease, I added $\delta_{disease}$ to the voxel-wise mean of the MCI group, μ_{MCI}, to give $\mu_{placebo}$. Note that this is intended to exaggerate the disease effect somewhat, so as to more clearly establish the comparison between methodologies. Further simulations presented below will use a more subdued – and more realistic model of the disease effect likely to be experienced by MCI subjects. To model a putative treatment which reduces this disease effect by 25%, I added $0.75\delta_{disease}$ to μ_{MCI} to give $\mu_{treatment}$. The

means $\mu_{\mathrm{treatment}}$ and μ_{placebo} differ by an amount that corresponds roughly to a 25% reduction in disease-related atrophy, as opposed to *total* atrophy, which includes age-related atrophy as well; using the total atrophy, rather than the disease specific atrophy, to model the disease would overestimate the sensitivity of both methods. This observation will hold true for later simulations as well. Using Σ_{MCI}, $\mu_{\mathrm{treatment}}$, and μ_{placebo} I drew 100 samples from two multivariate Gaussian distributions, $X_{\mathrm{treatment}} \sim \mathbb{N}(\mu_{\mathrm{treatment}}, \Sigma_{\mathrm{MCI}})$, and $X_{\mathrm{placebo}} \sim \mathbb{N}(\mu_{\mathrm{placebo}}, \Sigma_{\mathrm{MCI}})$. I then used these samples to compare methodologies as in a real trial with a pre-selected "statistical ROI". Then, I repeated the entire process using an identity matrix instead of Σ_{MCI}, corresponding to the case where all covariates are uncorrelated.

To calculate the SVM methodology's test statistic, (*i.e.*, the cross-validated average accuracy), I used a 10-fold cross-validation procedure, i.e., holding aside 10% of examples – one fold – for testing, and training an SVM on the remainder. Accuracy was averaged over all 10 folds to give the final test statistic. I then calculated p-values according to a Binomial distribution, which is the Null distribution for this test. For comparison, I performed a univariate t-test on each fold using only the training subjects, and averaged the p-values over all 10 folds. In summary, each method had up to ninety training subjects which were used to select parameters, and test statistics were averaged over ten randomized samples (without replacement) of the entire sample population.

To evaluate each method in terms of its performance as a function of the number of subjects used in a hypothetical trial, I repeated the above process with an increasing subset of the training sample for each fold. That is, I first used only two virtual subjects per arm in each fold, and then four, and so on, up to ninety subjects per fold. Note that the SVM's accuracy, *i.e.*, the fraction of correctly labeled test subjects, is evaluated using all two hundred subjects through cross-validation, while the t-test is only performed on the training cohort, which can be quite small. To account for this, I used N_{tr} times the test-set accuracy, rounding down to the nearest whole number, as the parameter determining the number of "coin flips" when calculating the binomial p-value, where N_{tr} is the size of the training cohort. That is, I used all 200 virtual subjects to derive a low-variance estimate of the classifier's accuracy for each training set size, and then interpolated that accuracy to the size of the training set used in order to calculate a p-value. Thus, using the larger test set to estimate accuracy does not make

the SVM-based methodology more powerful relative to the univariate t-test method; in fact, using N_{tr} as the number of "coin-flips" when the accuracy values have been averaged over a larger number of trials is somewhat disadvantageous; consider that getting 55 heads out of 100 coin flips is far less significant than getting 550 heads out of 1000. However, this was not an issue, as the SVM-based methodology performed quite well in these simulations.

Results

First, observe that the covariance structure among voxels is fairly dense, and the main diagonal is almost imperceptible, as shown in Figure 7.2. As we might expect, the eigen-values decay rapidly, as shown in Figure 7.3. This provides an important validation of a critical underlying assumption: that voxel-wise covariates are not just correlated, they are extremely correlated. Note that the first eigen-value *alone* (9.16) accounts for 65% of the mass of the entire spectrum (14.13). It is worth pointing out that this is largely a result of the way that these voxels were selected – they represent a relatively small sampling of the entire set of voxels in the brain, and they are selected for their relevance to AD, which will contribute significantly to their overall correlation. However, this is precisely the property which we would like to enhance in a clinical trial or Neuroimaging study.

The results comparing the proposed methodology with the univariate t-test for the case where I used Σ_{MCI} are shown in Figure 7.4. The results for the case in which I used a diagonal covariance matrix are shown in Figure 7.5. In the first experiment using Σ_{MCI}, shown in Figure 7.4, the univariate t-test utterly failed to detect a significant difference; p-values are centered about 0.5, no matter how many samples were used. In contrast, the SVM methodology achieved significance at the $\approx 10^{-4}$ level with only twenty eight training samples per arm. Test-set accuracy quickly rose to 75%, and reached 88.5% when using the entire training set of ninety subjects per arm, giving a significance level of . In the second case, in which I used a diagonal covariance matrix, (*i.e.*., treating all covariates as completely independent,) the results were exactly the opposite. (Figure 7.5.) Test-set accuracy barely reached 60%, and was only consistently significant at the 5% level for seventy subjects per arm or above. In contrast, the t-test was able to show significant differences at the $\alpha < 10^{-6}$ level for

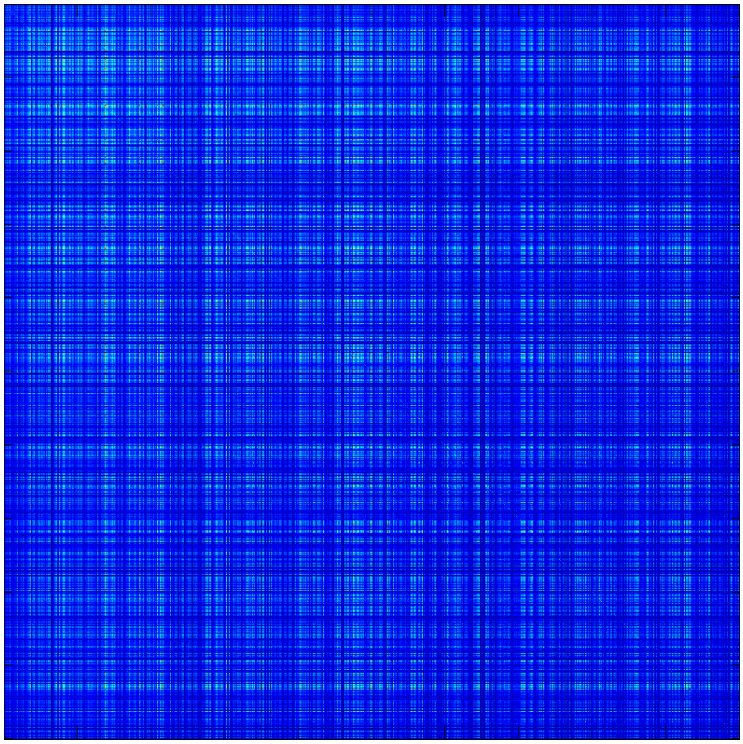


Figure 7.2: The MCI cohort sample covariance of the 1000 voxels selected. Note that the matrix is *not* diagonal, meaning that the voxel intensities are highly correlated.

as few as ten subjects, and decreased exponentially as more subjects were added.

This marked difference in outcomes between the methods can be understood in terms of their relative strengths and weaknesses. The SVM method is more effective in lower dimensional spaces, both theoretically and in practice, for several reasons. For one, having more training examples than input dimensions means that the instance space is better sampled than when the reverse is true. Another way of looking at the issue is that when the kernel space is high dimensional, each example can be thought of as having its own dedicated dimension outside of the span of the other examples – *i.e.*, the kernel is highly diagonal. This is a problem because the SVM is forced to use self similarity to classify each example and providing a large margin becomes trivially easy, yet, removing the diagonal would make the problem non-convex. Further, it is well known that VC dimension and Rademacher complexity grow with the dimensionality

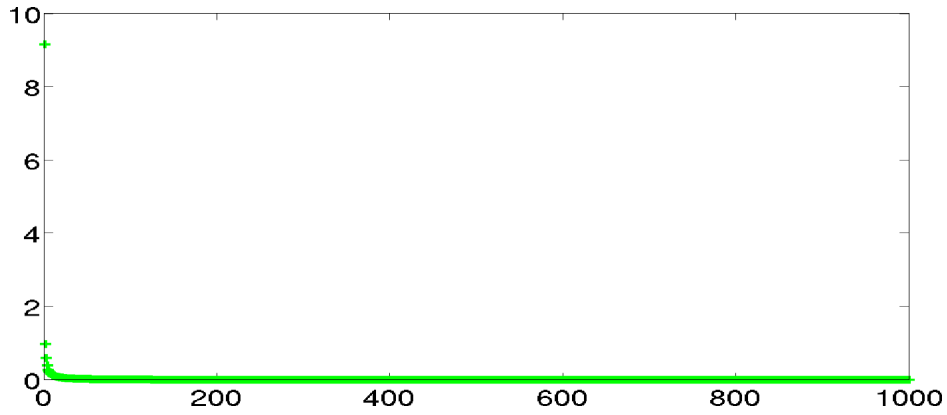


Figure 7.3: Eigen-values of the MCI cohort sample covariance matrix. Note the rapid decrease in magnitudes, and that the first single eigen-value alone accounts for nearly half of the variance of the entire distribution.

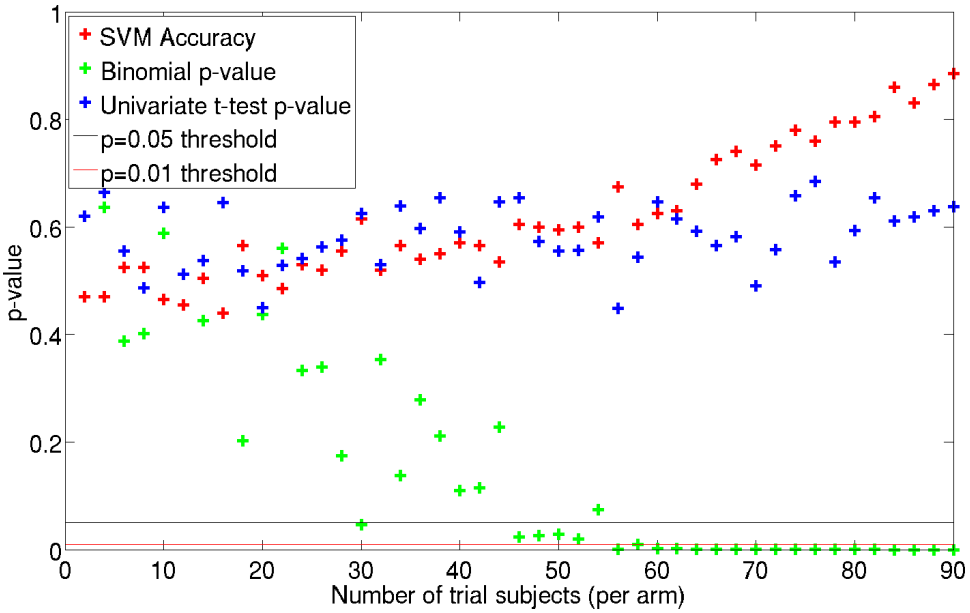


Figure 7.4: Experimental results with simulated Multivariate Gaussian data, using Σ_{MCI}. Cross-validated accuracy, corresponding p-values, and averaged t-test p-values as a function of the number of training samples per arm.

of the input space, and as these measures of complexity increase, the generalizability of the learned pattern classifier and its training accuracy and margin becomes more and more in doubt. Essentially, the SVM is looking for a very specific pattern, and in high dimensional, isotropically distributed data, it has too many potential options for it to confidently extract the right one.

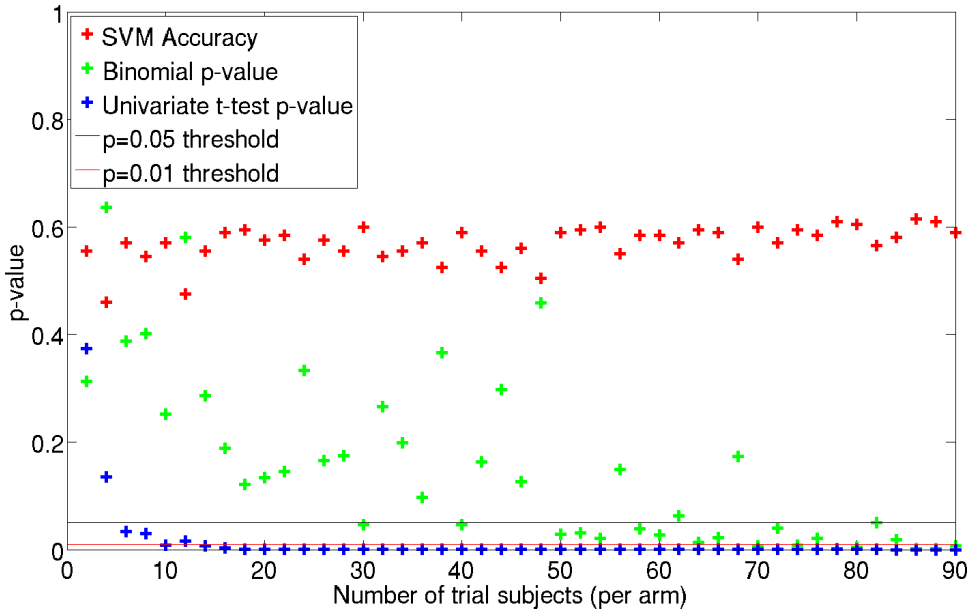


Figure 7.5: Experimental results with simulated Multivariate Gaussian data using $\Sigma = \mathrm{I}$. Cross-validated accuracy, corresponding p-values, and averaged t-test p-values as a function of the number of training samples per arm.

The univariate t-test, however, represents the opposite end of the spectrum. Consider that the mean of a large number of independent random variables has a distribution function which is equal to the convolution of the individual covariate distribution functions. The Central Limit theorem dictates that this distribution will not only be Gaussian for a large number of covariates, but more importantly, as the number of random variables increases, the distribution of their mean becomes narrower and narrower, provided the individual covariate distributions are unimodal and monotonically decreasing about the mean. In other words, where the SVM looks for as specific a pattern as it can, taking the mean over all of the covariates makes no specific assumption at all,

and the SVM's tendency to search for specific patterns makes it vulnerable to minute variations. On the other hand, in a low dimensional setting where the principal axes of variation (*i.e.*, Principal Components) differ markedly in terms of their contributions to the overall covariance, (*i.e.*, when the eigenvalues of the covariance matrix decay rapidly as in Figure 7.3,) then unless the optimal separation boundary is in the span of the first few eigen-functions, then the univariate mean method will fail to detect it. In contrast, the SVM's margin seeking behavior can be understood as pushing the discriminating boundary outside of the largest eigen-functions, and into the lesser eigen-fuctions, depending on how much each eigen-function contributes to the margin. (See Figure 7.1.)

As a final comment on the form of linear classifier, note that the SVM is not the only linearly discriminating method available to us. For instance, Fisher's Linear Discriminant Analysis (LDA) is the optimal method of discriminating between two Gaussian distributions having different covariances, and would work equally well in this setting – in these experiments, which are Gaussian by design, this is almost certainly the case. However, when distributions are not Gaussian, it is not as clear that LDA is always preferable. One of the strengths of the SVM is that it makes no assumption on the distribution of the data, *i.e.*it is a discriminative model, rather than a generative one. It is often assumed – and reasonably so – in the context of neuroimaging analysis that the observed data come from a Gaussian distribution, however neither is this always guaranteed to be the case. In fact, as will be discussed in the subsequent and final chapter, the use of permutation testing for estimating the Experiment-Wise error rate is largely driven by the desire to avoid making a Gaussian assumption. Regardless of the method of discriminating algorithm, or the assumptions it brings with it, however, the central purpose of this Chapter is to establish that there is a significant gain in information to be accrued by using cross-validated linear discriminant functions, trained on treatment vs. placebo labels, as a clinical trial end-point. Additional Monte-Carlo trial simulations described in the next Section serve to further this aim.

7.3 Power Calculations

Before we can accept the outcome of a trial, we must establish bounds on the probability that the trial's conclusion is incorrect. These calculations are different for the two cases involved, which I will discuss separately.

Type I error

The probability that $t^* > \delta$ when no real difference exists can be estimated using the Binomial distribution, (or permutation testing). That is, under the Null hypothesis that there is no detectable difference between groups, the output of a linear classifier will be Bernoulli distributed with $p = 0.5$, (or some other value if the treatment and placebo classes are not balanced,) and cross-validated accuracy will necessarily be Binomial distributed. Thus, we can simply calculate α as,

$$\alpha = \sum_{k}^{N} \left(\begin{array}{c} N \\ k \end{array} \right) 0.5^k (0.5)^{N-k},$$

where N is the number of subjects in the trial, and k is the number correctly classified by the algorithm. Note that this calculation assumes that there is absolutely nothing that a linear classifier can use to distinguish between arms of the study, so both arms *must* be well matched according to age, education, APOE type, and any other factor which may affect brain morphology.

Type II error

The calculation of β, the Type II error, is somewhat more involved. This is because one cannot estimate β without making some assumption on the type of differences to be detected. For instance, by using non-linear kernel functions to represent the data in an alternate RKHS, we can detect a wider variety of group differences, which linear classifiers may not be able to detect. However, as we cannot inspect the points in an arbitrary RKHS, or characterize their distribution in terms of observable variables, we lose the ability to calculate β. Thus, we will limit ourselves to using linear classifiers on TBM imaging data.

Type II error is the probability that $t^* < \delta$ given that H_0 is false, but this is a function of the true classification risk $1 - \delta^*$, which we do not know. The true risk is analogous to the notion of effect size in a two-sample t-test methodology, i.e., we have to assume a given level of separability in order to calculate the probability that t^* will exceed this level on test data. Our solution to this problem is that we will begin with the usual notion of "effect size", and translate it into a learning-theoretic bound on the true risk. The standard used by most ADNI studies is that there should be an 80% percent chance of detecting a 25% reduction in atrophy. We therefore require a way of translating the notion of a "25% reduction in atrophy" into a risk bound. To do so, we must generalize this notion slightly – note in particular that a treatment may have an effect on specific regions, as opposed to globally reducing the (AD-related) atrophy. For some voxels, the trial and placebo groups will have different distributions, so, provided the treatment is indeed effective, the two groups should have different means. While individual voxels may not have significant differences, (for a given sample), but in linear combination we can boost these differences.

The difficulty remains, however, that when we abandon the assumption that the primary end-point will have a parametric form, giving a closed-form solution for Type II error calculations, we are faced with a non-parametric setting in which there is no such corresponding expression. Yet, for a non-parametric problem, we can instead look for a non-parametric solution. As mentioned above, the point of focus is what a "25% reduction in atrophy" really means. The non-parametric way to answer this question is to train a linear classifier using AD and control subjects, (or stable and converting MCI subjects,) and treat the learned disease pattern as a model of the disease. Note that the SVM does not model the individual class distributions, however Fisher's LDA or Naïve Bayes do. Thus, the projection of a point onto the disease classifier's output space is a measure of atrophy, and a 25% reduction in AD-related atrophy would be a 25% reduction in the shift from the control distribution to the AD distribution in that output space. Subsequently, we need only to relate the Type II error rate to N, the number of participants in the study. Recall that the significance level is determined by the cross-validated error as a function of training examples. In order to find the number of subjects required in order to attain a desired significance level, we can plot learning curves on a simulated trial using the pattern trained on AD and control subjects as the disease model. This way, the desired significance level

will correspond to a particular point on the plot, and we can then read off the required number of trial participants needed in order to give the desired β level.

7.4 Monte Carlo Evaluations

Any analysis of this methodology would not be complete without showing that it can be effective in realistic scenarios. Theoretical calculations discussed above may be sufficient, but numerical simulations based on human participant data can demonstrate that this approach has merit in a real setting. In this section I describe a novel method of simulating clinical trials using ADNI neuroimaging data, as well as the results of those simulations. This methodology in some ways relaxes the Gaussian assumption of the previous section, in that rather than drawing virtual subjects as random samples from a multivariate Gaussian distribution, it proposes to use existing subjects, and model the disease trajectory as an affine translation of one distribution onto another. In this set of simulations, I used MCI participants who progressed to Alzheimer's Disease within 24 months to develop a model of the disease course, both in terms of the shift in voxel-wise means, *and* the change in their covariance pattern. Using this model, I then repeatedly simulated randomized clinical trials and compared the univariate t-test against my method. I discuss the details and results of these experiments next.

Direct Simulation trials

In this set of experiments, my aim is to simulate a real clinical trial as faithfully as possible, using only the scans that are available in the ADNI cohort. Observe that in AD, both the mean voxel intensity, *and* the pattern of covariances can vary as the disease progresses. Changes in the covariance may be caused by selective patterns of atrophy; that is, some voxels may become decorrelated with their neighbors in the presence of a systematic pattern of atrophy. In order to simulate this effect, I treated the two groups as having different covariance patterns, and the disease course as an affine warp from one distribution to the other. In this simulation, each subject follows a unique disease trajectory from the healthy group to the diseased group. I then model a 25% reduction in disease-related atrophy as a 25% shortening of this trajectory. This way, instead of drawing samples from multivariate Gaussian distributions, I used the

actual ADNI MCI participant data, and simulated *only* the disease pattern. Recall that if one were to treat *all* atrophy as being disease related then a "25% reduction" will significantly overestimate the effect of a treatment, giving unrealistic sample size, and power estimates. Secondly, instead of using AD and control subjects to estimate the effects of disease, I used stable MCI subjects, (those whose diagnosis remained unchanged after 24 months,) and converting MCI subjects, (those who converted to AD within 24 months), as this more closely resembles a clinical trial composed only of MCI subjects at risk of converting to AD. Recall that in the motivating experiments presented above, I used the control vs. AD difference of means as the disease model, so as to generate as strong a signal as possible for expository purposes. However, in a more realistic setting a clinical trial is likely to be focused on at-risk MCI patients, because these are exactly the patients who stand to benefit from a putative treatment. Moreover, for the stable MCI population, we cannot rule out the possibility that some of those subjects are themselves nearing a conversion to full dementia, which would have happened after the end of the study. This too is a more realistic setting. (*cf.* clinical trial enrichment methodologies, discussed in the previous Chapter, which represents the opposite case.) Thirdly, I used 10,000 voxels, rather than 1000. As before, these 10,000 voxels were chosen according to voxel-wise t-statistics, using AD and control participants only. There are several motivations for doing so: while the AD signal is strongest in the hippocampus and surrounding areas, roughly covered by the 1,000 strongest voxels, AD-related disease effects can also be seen in a broader, though somewhat more diffuse, pattern throughout the brain Klöppel et al. [2008], Cuingnet et al. [2011]. Finally, in order to investigate the relative strength of signal observed between GM and CSF voxels, I performed all of the following experiments separately for GM and CSF voxels only, (among the selected set of 10,000) as well as using all 10,000 voxels. The comparison is of interest because CSF voxels, particularly those bordering on GM regions, may give a clearer signal of atrophy. This is because CSF expansion seen at the boundary of gray matter regions is effectively an integration of all of the contraction happening at interior GM points. In other words, as a gyrus or other neuroanatomical structure shrinks, the inner-most voxels will appear to be unaffected, while the outer-most voxels will appear to be shrinking the fastest, and the surrounding CSF will be quite clearly expanding.

The procedure is as follows: First, I divided the subjects into two groups at random,

and computed the affine transformation as,

$$\mathbf{x}_{trans} = (\mathbf{x}_{original} - \mu_{stable})\Sigma_{stable}^{-1/2}\Sigma_{converting}^{1/2} + \mu_{convering},$$

where $\mathbf{x}_{original}$ is an unmodified stable MCI subject, $\mathbf{x}_{trans}$ is the same subject after the affine warp, μ_{stable} and $\mu_{converting}$ are the centroids of the stable and converting MCI populations, respectively, and Σ_{stable} and $\Sigma_{converting}$ are likewise the stable and converting MCI sample covariance matrices. Essentially, I centered the stable subjects, multiplied them by $\Sigma_{stable}^{-1/2}$, to make them isotropically distributed, multiplied them again by $\Sigma_{converting}^{1/2}$ to match the covariance pattern of the converting subjects, and finally I added back in the mean of the converting subjects. This is equivalent to finding a linear transformation which maps the distribution of the stable MCI group onto the distribution of the converting group. Thus, the vector difference between $\mathbf{x}_{trans}$ and $\mathbf{x}_{original}$ is the unique, individualized disease trajectory computed for each subject. For the simulated treatment group, I shorten this transformation by 25% by taking

$$\mathbf{x}_{treatment} = 0.25\ \mathbf{x}_{original} + 0.75\ \mathbf{x}_{trans},$$

and

$$\mathbf{x}_{placebo} = \mathbf{x}_{trans}.$$

Results are shown in Figures 7.6 and 7.7. I opted to display p-values in $-\log_{10}$ scale to better show the power of the proposed method. In Figure 7.6 are shown the results of this experiment when using all 10,000 voxels, and in Figure 7.7 are shown the same set of experiments using only CSF voxels. For comparison, the $-\log_{10}(0.05)$ threshold is shown in black, and the $-\log_{10}(0.001) = 3$ threshold in red. As in the previous simulations, the univariate mean t-test fails to reach even the 0.05 significance level, while the SVM test shows a significance of 10^{-6} with only 25 subjects per arm. This result is very competitive with anything reported so far in the literature (to my knowledge). Note also that with more and more training subjects, the SVM's accuracy continued to improve, giving p-values which decreased *exponentially* (seen as a linear trend in log scale).

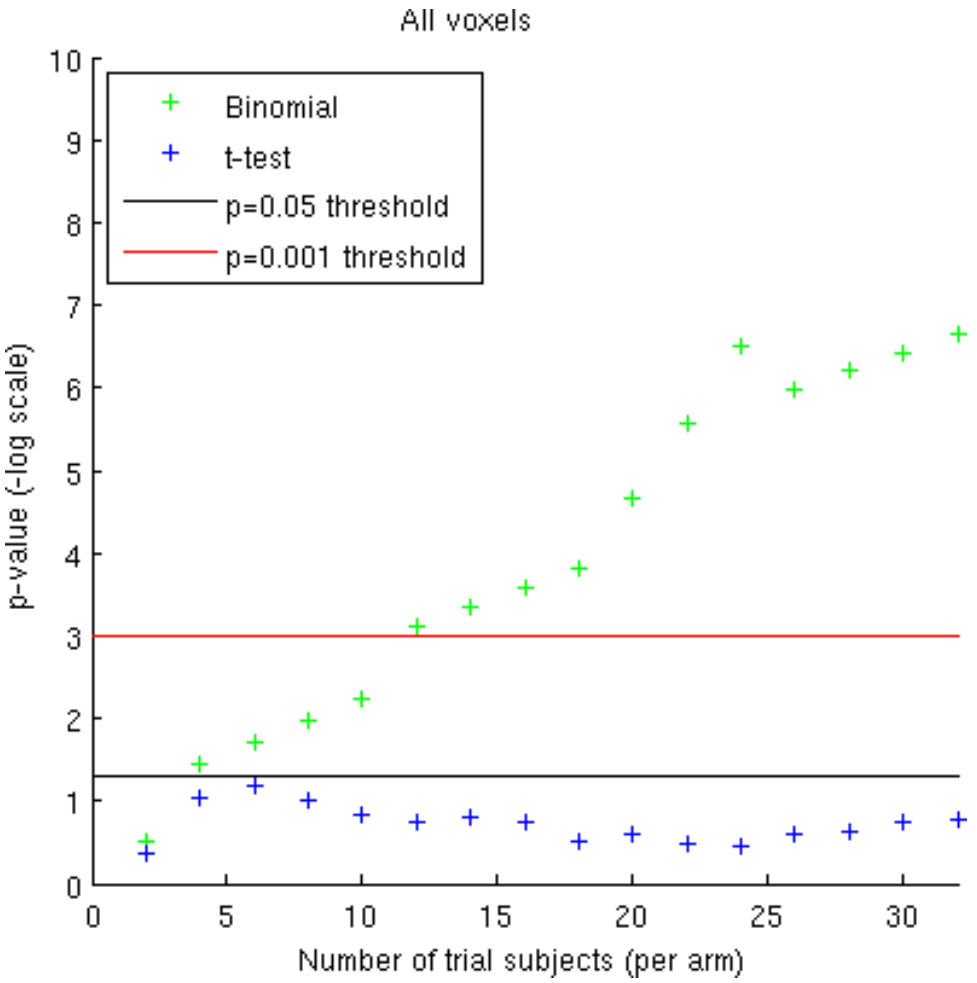


Figure 7.6: p-values from simulated trials using all voxels, mapping stable MCI to converting MCI.

7.5 Simulations Using the AD cohort

In addition to the simulations described above, I also simulated the disease course using the AD group. These experiments are shown in Figures 7.8 and 7.9. The results are largely the same, except that for some smaller samples sizes the t-test was able to show some significance at the 0.05 level, but never for the whole cohort. In the top row, I simulated the disease course by affinely warping the *entire* MCI group, including the

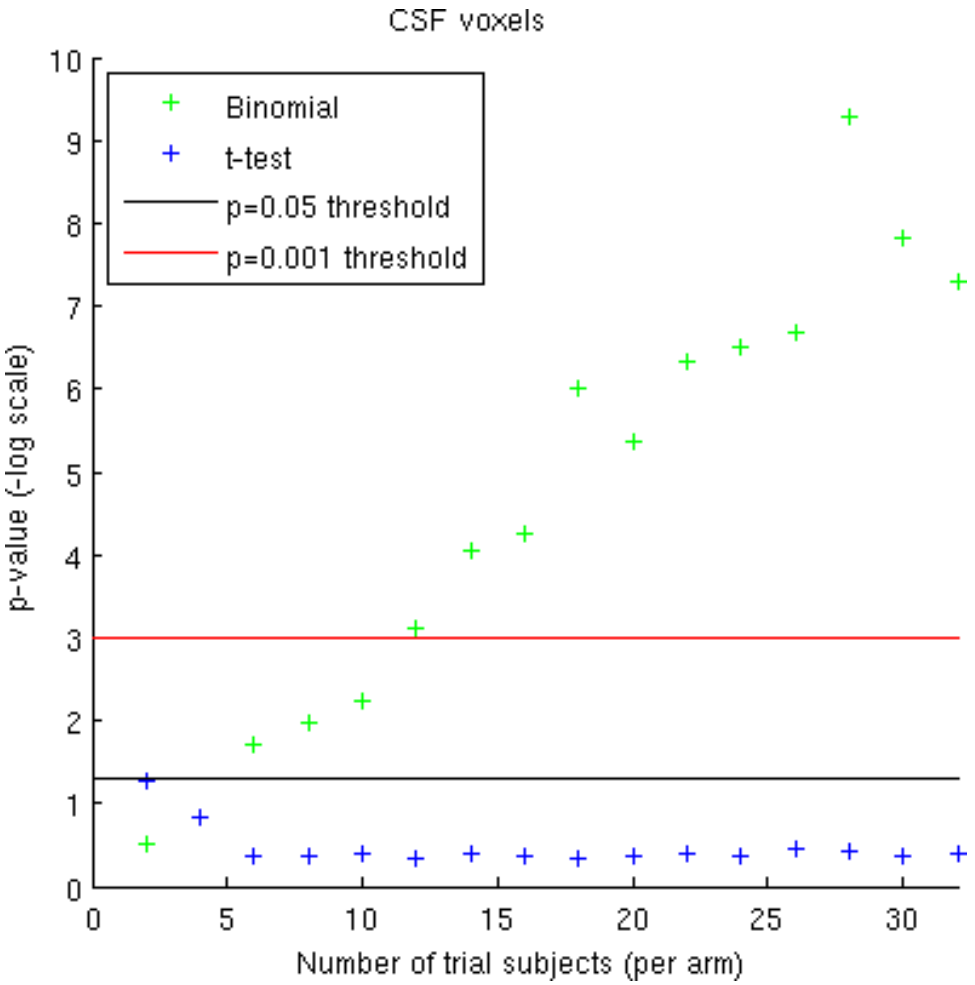


Figure 7.7: p-values from simulated clinical trials using CSF voxels only, mapping stable MCI to converting MCI.

converters, to the AD group. In the bottom row, I did the same, except using only the stable MCI subjects. As before, the left column shows the results when using ALL voxels, and the right column shows the results when using only CSF voxels.

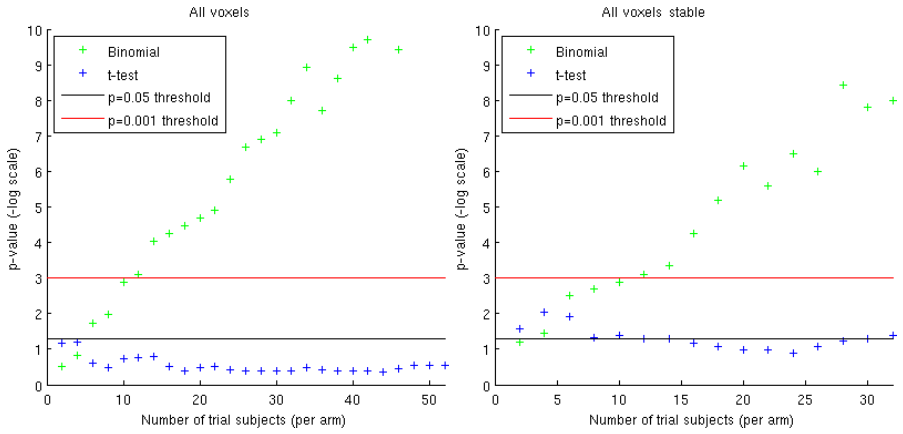


Figure 7.8: p-values from simulated trials using all voxels, mapping CN to AD.

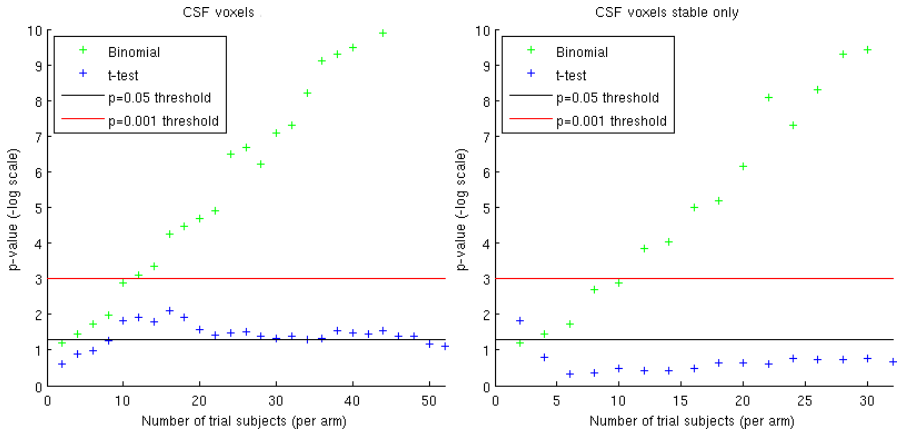


Figure 7.9: p-values from simulated clinical trials using CSF voxels only, mapping CN to AD.

7.6 Conclusions

There is an opportunity to dramatically increase the statistical power of clinical trials using neuroimaging-derived markers as primary end-points, by leveraging ideas from statistical learning theory to more directly address the question of treatment effect: rather than asking if a voxel-wise ROI mean differs in a way that is significantly different from chance, we could instead ask if there exists a linear discriminating function which can detect treatment effects in a way that is significantly different from

chance. My simulations show that the covariance structures inherent in neuroimaging data make uniformly weighted voxel means a non-ideal choice for aggregating high dimensional data, and that statistical learning methods such as the SVM are much better suited to the problem of detecting significant differences between trial populations.

Chapter 8

Future Directions and Open Questions

In this chapter I describe some ongoing work which is at varying stages of completion at this time. These projects are aimed at extending the works in this thesis both in theoretical, as well as in translational directions. If these ideas fully lead to fruition, they will almost certainly have a multiplicative effect on the utility of the contributions of this thesis.

8.1 Efficient Large-scale Permutation Testing via Matrix Completion

In previous chapters I have derived novel neuroanatomical markers of atrophy and used them in both identifying more sensitive outcome measures and specific inclusion criteria. By themselves, these methods show signs of having the potential to sharply reduce the required sample cohort sizes. Yet, the goal of a clinical trial is not merely to derive a yes/no answer that a treatment arm differs from a placebo arm. It is equally important to assess what the end results really mean — whether the neurodegenerative effect is as intended, and not a mere epiphenomenon, (*e.g.*, neuro-inflammatory side effects). To this end, once a difference is detected, we also need to assign statistical confidence levels both globally and at a voxel or feature level.

In this section, I examine the issue of testing for significant variations between groups in a high dimensional setting. That is, we are presented with two groups of high dimensional measurements (*i.e.*, each measurement consists of a high dimensional vector of random variables) and we wish to detect whether there is a significant difference in means for at least some of these variables. More generally, I am interested in finding out whether or not their density functions (PDFs) differ. For instance, in functional brain imaging, each measurement consists of a 3D image having as many as 10^6 voxels, and we are tasked with discerning whether some of these voxels show higher activation in one experiment group than in the other. When measuring only

one variable, a simple univariate t-test will often suffice; however, when the number of variables is large, multiple testing issues make point-wise comparisons difficult to interpret.

Multiple comparisons, or multiple testing, refers to the situation in which we are performing many tests on *independent* random variables, giving a large number of independent test statistics, which boosts the likelihood of observing a spurious result. In cases where each random variable – and hence its corresponding test statistic – is independent of all the others, we can calculate the Family-Wise (or Experiment-Wise) Type I error probability as a function of the Cumulative Distribution Functions (CDFs) of the individual test statistics. However, in the other extreme case, where all variables are 100% correlated, an uncorrected Type I error rate is correct since effectively there is only one random variable. For cases in between these extremes, there are many possible approaches to multiple comparison corrections.

As discussed in the previous chapter, another approach to the multiple comparisons issue is to avoid it altogether by performing one, and only one, test using all of the data. One such way is to use cross-validation to choose model parameters on one subset of the data, while evaluating test statistics on another. However, this is not so much a multiple comparisons *correction* as it is a multiple comparisons *avoidance*. In this section I will return to the more conventional setting in which we would like to examine in which *specific locations* a differential has been observed between clinical groups.

This issue is of vital importance: consider that if a cross-validated classification-based methodology detects a significant difference between treatment and placebo arms of a trial, we must still confirm that the difference is beneficial and that it relates to the disease under treatment. Subjectively, we can examine the pattern of voxel-wise weights that make up the linear classifier to see whether or not it is consistent with the existing literature on AD; however, this does not give us the ability to inspect individual voxels for significance. More broadly, we may propose a similar methodology for all neuroimaging studies that consider the possible existence and location of significant variations between groups of measurements. In the following I will propose a novel methodology currently under development which potentially offers a way of *efficiently* performing large-scale permutation tests by applying recently developed Robust Matrix Completion methods, giving a significant speedup over

existing methods while preserving the reliability of this method, and freedom from assumptions on the structure of the data. Next, I will review existing methodologies for attacking this problem, describe the proposed methodology and its motivation in detail, and present preliminary experiments demonstrating its effectiveness.

Multiple Comparisons correction methods

The simplest way to account for the effect of Multiple Testing is to use Bonferroni correction. However, this method makes several strong assumptions: because it is based on the Union Bound, in which the probability of the union of a number of events is no greater than the sum of the individual event-wise probabilities, Bonferroni correction is only exact in the case in which the events are *mutually exclusive*. In other words, Bonferroni correction is calculating the probability that any voxel or set of voxels will show a spurious significance, but it assumes that if this happens at any single voxel, then the same cannot be the case in any other voxel. This is obviously problematic when, rather than being mutually exclusive, or negatively correlated, or even independent, voxel-wise covariates are in fact highly correlated. In practice this can mask real, significant differences whenever the effect size is too small to provide a sufficiently high experiment-wise test statistic for the (usually small) number of subjects involved.

Another approach is to treat the voxel-wise statistics as variables in a Random Field, which allows a better characterization of their covariances. This type of model gives a better estimate of the experiment-wise p-value, which we can then use to filter out meaningful results from the expected amount of false positives. While several types of random field have been analyzed in this setting, Gaussian Random Fields (GRFs), which generalize the multivariate Gaussian to the case of an infinite number of variables, have received the most attention due to the ease of analysis they afford [Worsley et al., 1992, Worsley, 1994, Worsley et al., 1996]. Results from GRF theory show that the Euler characteristic number of a set (essentially, the number of contiguous components minus the number of holes) can be related to the expected supremum over the field. For the most commonly used global Null hypothesis (*i.e.*, that the mean of each voxel-wise distribution is equal between groups) this value yields the experiment-wise expected Type I error rate. However, this methodology

makes a Gaussian assumption, *i.e.*, that the distribution of voxel intensities for any image can be exactly characterized by a multivariate Gaussian distribution. While in practice individual voxels can be approximately Gaussian-distributed, their dependency structures can be somewhat more complex and cannot necessarily be captured via a measure of covariance.

A more direct way of addressing this question is to treat the issue not as a question of how to *correct* the measured p-values, but rather, of how to *interpret* them. In order to do this, we must have an unbiased estimator of the global (*i.e.*, joint) Null distribution of the test statistic. (In the following I will assume that this will be the t-statistic of each voxel for ease of exposition, however others, such as the coefficients of a GLM, etc., may be used as well, without changing the nature of the problem.) Once we have a good estimate of the Global distribution of the voxel-wise test statistics under the Null Hypothesis, we can then assess whether or not the observed range of statistics falls within this range, and, in particular, we can assess how likely the extremes are to be spurious, or not. If it were the case that each test statistic (*i.e.*, each voxel) were independent from all of the rest, then we can calculate the global Null distribution in closed form using the CDF of the test statistics. However, when covariates are dependent, this will not suffice, and so we must fall back on drawing samples from the Global Null distribution. A time-honored way of doing so is to randomly permute the groups many times, and for each permutation, re-compute the test statistic for each covariate. Aggregating these samples gives the global Null distribution; this method is known as Permutation Testing [Pesarin, 2001, Nichols and Hayasaka, 2003]. At the most basic level, permutation testing can be understood as a type of bootstrap sampling method [Wasserman, 2006], except that sampling is done without replacement.

Assuming a treatment effect is detected using the models described in the previous chapters, we can assign a reliable α-level to it by comparing the observed point-wise statistics with the permutation testing samples. This allows both localization of treatment effects and assignment of a global confidence level. A fact of crucial significance is that because we are interested specifically in the tails of the Null distribution, we require a very large number of iterations, often in the hundreds of thousands, each of which requires a pass over the entire imaging data set. This is the major driver of running time, which can run up to several days for large datasets, and hence there is significant benefit to reducing this cost. If a specific structural

redundancy can be identified, then it may be possible to speed up this process by exploiting it. As described next, the low-rank characteristics of this type of data may be able to serve this role.

Redundancy and rank in permutation testing

As in previous chapters, we may observe a fact of crucial significance: brain voxels are highly correlated with one another, even those from opposite ends of the brain, owing to global anatomical variations. This high degree of correlation means that for such voxels, if we know the t-statistic of one, we can predict those most correlated with it with high certainty. In the limit, as voxels become 100% correlated, they behave as a single t-statistic. Permutation testing, as currently used in neuroimaging studies [Singh et al., 2003, Nichols and Hayasaka, 2003], is completely unaware of this structure in the data. Our proposal is to make the above intuition rigorous using ideas from compressive sensing and matrix completion theory [Fazel et al., 2004, Recht et al., 2010, Candès and Recht, 2009, Candes and Tao, 2010], allowing for a more efficient process.

Just as the voxels are highly correlated, so are the rows of the permutation matrix $T \in \mathbb{R}^{m \times n}$, where m is the number of voxels tested per permutation, and n is the number of permutations performed. Thus, each permutation test fills a single column of T, where each row is the Null statistic computed for a particular voxel. Such strong correlations imply that these column-vectors are all tightly clustered in $\mathbb{R}^m$, *i.e.*, highly coherent. This leads to the central idea of this section – every time we fill in a column of T, much of this computation is highly redundant because T is inherently low-rank. Instead, if we were to randomly choose a small subset of entries, Ω, to populate, we could fill in the remainder of T by treating that process as a matrix completion problem. The question then becomes what type of low-rank structure do we wish to impose on this process?

While we do observe that the eigen-values of the sample covariance matrix (which relate to the singular values of the permutation matrix) decay rapidly, there is still the issue of what to do with the least singular values. Moreover, I have found that the singular values of the permutation matrix are spread out somewhat more broadly than those of the sample covariance matrix. Note that the sample covariance matrix cannot

be higher in rank than the minimum of the number of samples and the dimension of the ambient space, owing to the linearity of its construction; while the non-linearity of the construction of the t-statistics of the permutation matrix (especially in dividing by the pooled variance estimate) means that this is not necessarily the case. In practice, I found that the rank of this matrix is significantly higher than that of the sample covariance, albeit with very small values for the trailing singular values. To a reasonable approximation, this matrix has roughtly twice the rank as the sample covariance matrix. That is, when reconstructing from the first $2N$, where N is the number of subjects, singular vectors alone, the average entry-wise absolute residual was on the order of about 0.005, which is close to negligible for t-statistics. More encouragingly, this residual was almost always on the positive side, *i.e.*, it slightly over-estimated the distribution of the maximum, rather than systematically underestimating it, which is desirable for the intended application.

However, there is a way in which we can turn this behavior to our advantage. Consider that once the largest singular values are removed, the remainder of the spectrum is relatively flat. It is well known that among positive semi-definite matrices, there is a trade-off between sparsity and low-rank. That is, sparse matrices are high-rank (*e.g.*, the identity matrix is extremely sparse, and is full rank) while low-rank matrices are non-sparse in general. Thus, the relatively flat portion of the remaining spectrum, while much lower in magnitude than the few largest singular values, still contributes a sparse pattern to T, and this sparse pattern could be enough to upset the distribution of the maximum, which is our primary interest. For this reason, recently developed methods for sparse-plus-low-rank decomposition such as in He et al. [2012] are more likely to yield useful reconstructions. This is because they are able to exploit the "truly" low-rank behavior of the leading singular values, while treating the longer tail of smaller singular values as a higher ranked, but sparse contribution to the residual. Doing so will give a more faithful representation of the reconstructed permutation matrix, leading to a better estimated distribution of the maximum, as is demonstrated in the preliminary experimental results described below.

Let us therefore suppose that T can be decomposed as a product $T \approx UV^T + O$, such that the coefficients of the low-rank expansion are stored in $V \in \mathbb{R}^{n \times d}$ and the basis set, or "dictionary" of this expansion, is stored in $U \in \mathbb{R}^{m \times d}$, which is orthonormal and low-rank; d is the size of the dictionary, and $O \in \mathbb{R}^{m \times n}$ is a sparsely

populated residual that does not fit within the span of U. We can alternatively think of the sparsity of O as being a higher-rank, but lower mass, portion of this reconstruction rather than a "residual". Or, we can simply think of O as a sparse but otherwise unstructured component of the reconstruction. With this model, we need make no assumption on the Gaussian behavior of the distribution of T, which is the main motivation for performing permutation tests in the first place.

The reconstruction problem of solving for T as a constrained matrix completion problem is given as:

$$\begin{aligned} &\min_{U,V,O} \|O\|_{1,1} \\ &\quad \text{s.t. } [UV^T]_\Omega + O_\Omega = T_\Omega \\ &\qquad \mathrm{rank}(UV^T) \leqslant d \quad (d \text{ is a user-supplied constant}), \end{aligned}$$

where Ω is a set of entries of T which we have chosen at random to populate. The constraint on the rank of UV^T is crucial because otherwise the problem would be underdetermined, and any solution setting $T_\Omega = [UV^T]_\Omega, O = 0$ would be optimal. In general, *exact* rank-constrained optimization is computationally intractable, but as in the standard matrix completion problem [Candès and Recht, 2009, Candes and Tao, 2010], the rank constraint can be substituted with its tightest convex relaxation, the nuclear norm [Fazel et al., 2004, Recht et al., 2010] so as to make the problem efficiently solvable. The above model can transform how permutation testing is deployed within neuroimaging studies, by giving a reliable Family-Wise Error Rate, as well as power and sample size estimates, at drastically lower computational cost. Note that matrix completion is a generalization of compressive sensing to matrices, and much of the Restricted Isometry Property (RIP) theorems carry through [Recht et al., 2010]. Given that this is the case, it implies that T can be reconstructed to high fidelity even in the aggressive sampling regime (with information on just 5-10% of entries). Using an extensive array of first-order methods recently developed [Stich et al., 2012, He et al., 2012], this methodology can translate into a significant time savings in neuroimaging studies and clinical trials. Of particular interest is that in [He et al., 2012], where the rank d is a user-supplied constant and the algorithm returns a solution with exactly that many basis elements.

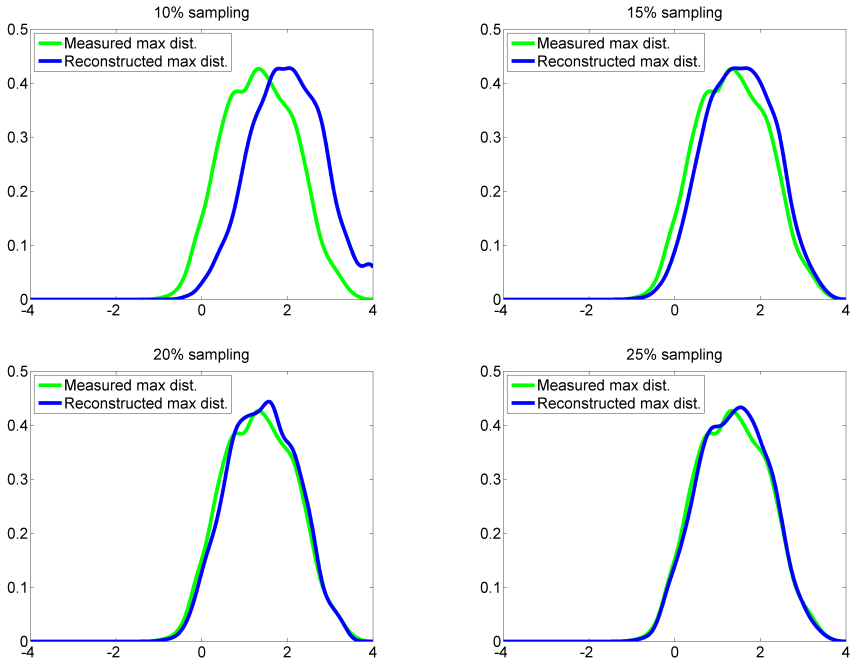


Figure 8.1: Distribution of the maximum t-statistic reconstructed from a varying sample as a percentage of all voxels.

Preliminary experimental results

Here I present preliminary experimental results demonstrating the effectiveness of the proposed methodology. The aim is to show that even though we have only sampled a small percentage of the full permutation table, we can nevertheless recover the object we desire: the distribution of the maximum Null t-statistic for each permutation trial. The procedure was as follows: I first computed a full permutation matrix for 1000 voxels over 2000 trials. Then, using the GRASTA method described in He et al. [2012], I reconstructed the full matrix using an incrementally increasing fraction of randomly selected entries, ranging from 10% to 25%. I then compared the distribution of the column-wise (*i.e.*, permutation-wise) maximum of the reconstructed permutation matrix with that of the full permutation matrix. Distributions were calculated by taking smoothed histograms, where smoothing was done with a Gaussian filter of width 0.1.

Results are shown in Figure 8.1. As the sub-sampled fraction goes from 10% to

25%, we can clearly see the reconstructed distribution, shown in blue, converging to the true distribution, shown in green. Moreover, the reconstruction is always conservative in that it slightly over-estimates the distribution of the maximum. Even for 15% sub-sampling, the reconstructed distribution matches the original quite closely, and at 25% it even matches the particular idiosyncratic features of the true distribution (bottom right). At this time, the reconstruction method requires more running time than the permutation sampling method itself, by a factor of about 4 for the 10% sampling case, and slightly more for the others. While the reconstruction accuracy is an encouraging sign that this problem can indeed be solved by such methods, the ultimate motivation is to speed up the process, which is not currently achievable. However, note that while the process of calculating the actual permutation tests can never be sped up, there is constant progress being made in efficient matrix completion algorithms, which strongly suggests that in the near future this methodology will indeed become the faster way of computing the permutation-wise maximum Null Statistic. Note in particular that while the latent rank of the permutation matrix is fixed, we can achieve much lower effective sampling rates by reconstructing a larger number of covariates at once, which may give the desired level of speed-up.

8.2 Ongoing Applications to Planned Clinical Trials

In the work described in Chapter 7, I used simulated clinical trials of Disease Modifying (DM) treatments in order to estimate the relative efficacy of the proposed method. However, until it has been validated on an real-world trial involving real human at-risk participants being given actual treatments, we cannot be fully certain that the method has as much merit as it appears to. Therefore, a clinical trial is currently in the planning stages that will include the proposed methodology in a retrospective analysis with the hope of validating its efficacy. Specifically, investigators at the Wisconsin Alzheimer's Disease Research Center (WADRC) are interested in using multi-modality imaging as well as other cognitive and biological measures in an MMDM-like framework (see Section 6.1) which can be used either in screening, as described in Section 6.3, or more interestingly, as the basis of a learning-based outcome measure as described in Chapter 7.

Participants in the planned trial are to be recruited from among a cohort of

90 subjects (63 controls, 27 MCI). This group has previously participated in the MERIT 220, PREDICT, and other studies conducted under the Wisconsin Registry for Alzheimer's Prevention (WRAP). All participants in this special cohort have had several imaging and other measures acquired at several visits. Imaging measures acquired include lognitudinal T_1, T_2 Fluid Attenuated Inversion Recovery (FLAIR), and Arterial Spin Labeled (ASL) MR images, taken at an intervals of approximately two years. In addition, family history, APOE genotype and vascular factors such as blood pressure (systolic and diastolic), HDL and total cholesterol levels, insulin and glucose levels, and Body Mass Index (BMI) are recorded for each of these visits. Further, a measure of change on White Matter (WM) hyperintensities (a type of vascular lesion in white matter found commonly in elderly populations) between the two visits is to be calculated. Of particular interest, this study will include a wider array of imaging and non-imaging modalities for each subject than are available in the ADNI study, and, the T_2-weighted FLAIR imaging is expected to be more sensitive to signs of early dementia than FDG-PET, due to the confounding relationship between WM hyper-intensities and AD.

During the planning stages, simulated trials will be conducted with the aim of estimating exactly how sensitive the proposed marker will be, and determining whether the existing cohort is large enough for this methodology to succeed. The methodology for deriving a test statistic and α-level will be largely as described in Chapter 7. Once the trial phase is completed, and all participant images have been normalized to a standard template, a series of classifiers will be trained with the task of discriminating between the treatment and placebo arms of the trial. The classifiers can include single-modality SVMs or $\mathbf{Q}$-SVM or multi-modality methods such as MKL or $\mathbf{Q}$-MKL . The $\mathbf{Q}$-SVM models will give a more interpretable voxel pattern of discrimination between groups, however, a $\mathbf{Q}$-MKL classifier will allow the combination of a larger set of modalities. Naturally, in a real trial we would specify one, and only one, classification model at the beginning of the trial, so as to avoid multiple testing issues. However, for the purposes of an exploratory, retrospective analysis, it is advantageous to examine a wider variety of classification methods.

The above steps having been completed, an α-level will be computed from the Null distribution of the classifier's cross-validated predictive error – *i.e.*, a Binomial distribution. This α-level can then be compared against the primary and other secondary

outcome measures of the trial to confirm or disprove the hypothesis that a learning-based outcome measure will be more effective than existing measures of cognitive status. If it is indeed the case that the learning-based outcome measure proves to be more sensitive to treatment effects than existing measures, then the next question will be to discern whether or not the treatment has had a beneficial effect. Various methods for doing so have been suggested in Section 7.3. Comparison and validation of these methods will be equal in importance to establishing a level of sensitivity.

As an exploratory analysis, this study will provide valuable data and insights into the efficacy of the proposed methodology, from an empirical perspective.

Chapter 9

Conclusion

Alzheimer's Disease is having a growing impact on society and, as a result, the search for effective treatments is receiving an ever increasing amount of attention. A significant roadblock to this effort is the difficulty in characterizing the exact relationship between neuropathology and observable cognitive status. Contributing factors include many confounds such as cardiovascular health, education level, genotype and family history, as well as the difficulty in characterizing cognitive status itself. From test-retest variability, to coarseness and ambiguity in the neuropsychological cognitive status measures such as delayed recall, auditory and visual learning, or other memory tasks, it remains difficult to put an exact number on a patient's degree of cognitive decline based solely on outward signs. Neuroimaging offers a much more precise and repeatable way of measuring the underlying pathological process which eventually leads to cognitive decline, but at the expense of greatly expanding the volume of data which must be analyzed before inferences can be drawn. When applied *en masse*, traditional univariate statistics such as t-statistics, linear regressions or measures of correlation have limited interpretability owing to multiple testing issues. These methods, while reliable in their original univariate context, were simply not designed in an era of massive, high-dimensional data sets, with complex interdependencies among covariates. In the last decade, the trend has been towards using machine learning methods as primary analysis tools in neuroimaging, largely as an acknowledgement of this fact. Yet, while such methods are indeed designed for high-dimensional settings, there is nevertheless further potential for improving the predictive performance of these methods by incorporating knowledge of the setting in which they are to be applied.

In this thesis, I have described machine learning methods designed to capitalize on the particular structural characteristics of neuroimaging data, making more accurate and more interpretable predictions about the form and progression of Alzheimer's Disease. Using data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), I have experimentally validated the power, and applicability of these methods

to the analysis of Alzheimer's Disease in neuroimaging contexts, showing significant improvements in terms of discrimination ability between diseased and healthy study participants; in terms of predictive ability as to which MCI patients will progress to AD; and in terms of the sensitivity of the proposed markers for use in clinical trials. Beyond the context of Alzheimer's Disease research, the algorithmic and modeling developments described above are relevant to the broader advancement of machine learning methods as well. For instance, the **Q**-MKL model described in Chapter 5 need not be treated solely as a neuroimaging method; for instance, it can be viewed as a way of implementing weighted combinations of sparse and non-sparse norm regularizers in MKL, among many other potential uses. Equally important is the fact that while improving predictive accuracy is indeed a major goal of this work, the real aim is to facilitate the use of machine learning methods as tools of scientific investigation. This means that *interpretability* of the learned model parameters is equally important; measures of sensitivity or precision merely serve to establish a level of confidence in these parameters. In this capacity as well, the models I have developed serve to enhance the scientific utility of discriminative learning models. Such enhancements may be brought about by providing more refined spatial models of atrophy; or providing more effective combinations of different imaging modalities; or by examining outlier subjects for signs of sub-group heterogeneity; or by allowing the primary statistical question under consideration to be phrased in new ways.

In the remainder of this chapter, I will describe the principal contributions of this thesis, and in Section 9.2 I conclude.

9.1 Contributions

The work described in this thesis falls under three main umbrella categories: Spatially regularized learning methods for neuroimaging; Multi-modality learning methods based on Multi-Kernel Learning (MKL); and applications of machine learning methods to clinical trials. Each of these is motivated, as well as evaluated, in the specific context of neuroimaging analysis of Alzheimer's Disease, however, there is nothing which inherently limits these methods to this context alone. So long as the right kind of structural assumptions can be justified, then the methodologies described throughout may be applied to other settings as well. In the following, I will briefly summarize

the key contributions of this work from a methodological, as well as investigative, perspective.

Structural biases for learning in the context of neuroimaging

In Chapter 3, I described several contributions to single-modality machine learning based neuroimaging analysis methods, which impose structural biases on the learned pattern classifier designed to capitalize on the known characteristics of neuroimaging data.

- Beginning In Section 3.2 I proposed a model, called Spatially Augmented Linear Program Boosting, which imposes a smoothness prior on the classifier, leading to both a more interpretable and a more accurate model.
- I then provided extensive and rigorous analyses of this method's predictive performance on ADNI data, as well as its ability to correlate with other markers of cognitive status.
- As a follow-on, I then proposed an alternative model that addresses the same issue of imposing smoothness, but using a less-harsh regularization scheme based on a Mahalanobis metric of voxel-wise similarity. This method, called **Q**-SVM, produced smoother, more interpretable disease patterns than the flat, single-valued patterns returned by SA-LP-Boost. Experiments on ADNI data confirmed that this method gives significant improvements over standard SVM models.

Multi-modality learning methods

Continuing in Chapter 4 I proposed several novel methods for combining multi-modality neuroimaging data using MKL-derived methods. These methods are motivated by the AD classification setting but are in principle applicable to other settings in which their underlying assumptions are met. As above, I rigorously evaluated these methodologies using ADNI patient data. These advances are detailed in Chapters 4 and 5.

- In Section 4.1, I examined in detail what effect the p-norm regularizer in existing MKL models has on discriminative accuracy for MKL models.

- In Section 4.2 I adapted a previously proposed outlier ablation method to the multi-kernel setting, in which outlier subjects are detected and systematically attenuated in terms of their contribution to the output classifier. The motivation for this is described in detail in Section 6.2. I presented a convex relaxation of the model which is easy to optimize, and rigorously evaluated its performance on ADNI data.

- In Chapter 5 I motivated and proposed a new class of MKL algorithms that make use of measures of interactions between kernels. That is, by measuring how much a pair of kernels varies in its contribution to the overall classification error, we can use this information in a regularizer to force greater diversity of information into the final classifier. If this information comes from outside of the training data itself, then the improvement in classification power is more tangible, but, in some cases, estimates of interaction from the training data alone were sufficient.

- I derived and implemented an optimization framework for this model which was shown to converge rapidly.

- I presented theoretical analyses showing that there are guaranteed improvements in learning generalization as long as certain assumptions are met.

Applications to clinical trials and other scientific questions

In Chapters 6 and 7, I described a number of contributions to the field of Alzheimer's Disease research, including the proposal and analysis of several new machine learning based analysis tools, as well as several significant contributions to clinical trial design.

- In Section 6.1 I examined how MKL can be used to generate predictive, multi-modality disease markers of AD, called Multi-Modality Disease Markers (MMDMs). I then evaluated their ability to predict which MCI subjects would progress to AD within a two-year time span.

- In Section 6.2 I detailed two separate analyses of outlier groups within the ADNI cohort. The identification of these subjects was done by examining their relative difficulty of classification and by their contribution to anomalous weight

patterns in the trained classifier. In the first analysis, the results showed that this sub-group of the AD cohort had more gray matter in certain regions than the controls cohort did on average and, conversely, the control outlier group more closely resembled the AD group. These results also carried through to highly significant variations in several cognitive measures as well. Because the identification of these outlier groups was made on the basis of discriminative measures, it is unclear whether they might have been identified by standard methods alone.

- In Section 6.3 I proposed a method of screening out low-risk subjects from clinical trials, using multi-modality imaging based predictive markers derived from an MKL classifier. When a clinical trial is conducted in which many participants cannot possibly benefit from the proposed treatment due to their not truly suffering from the disease being treated, then the real benefits of the treatment will be confined to a smaller pool of participants, which can mask the effect. Hence, this methodology can have a significant impact on both the size and sensitivity of clinical trials.

- In addition to screening methods, we can use learning methods to derive more sensitive outcome measures, (*i.e.*, end-points). That is, by using the AD and control cohort to train an AD-specific voxel-wise pattern of atrophy, I was able to increase the effect size, which led in turn to reduced estimates of required cohort sizes.

- Following this line of reasoning, I developed in Chapter 7 a methodology that uses a trained model of the actual treatment effect, rather than a predefined AD disease model, as an outcome measure. That is, if the treatment has an effect which is similar, but not identical, to the discriminative pattern given by an SVM or MKL classifier, then we may further increase the sensitivity of the outcome measure by training it directly from the trial cohort, using treatment and placebo arms as the classes to be discriminated.

- In order to simulate clinical trials using the proposed methodology, I devised a novel method of simulating clinical trials which both takes into account the change in voxel-wise means, as well as changes in their covariances.

9.2 Summary

This thesis has described in detail my contributions to the state of the art in Machine Learning, Neuroimaging analysis, and Alzheimer's Disease research. I have demonstrated in a variety of ways that whenever domain-specific information can be incorporated into learning models, there are tangible benefits to doing so. Such information can take the form of known relationships between features or covariates; relationships between views, modalities or kernels; estimates of the number of outlier subjects; or how we wish to use high dimensional data to answer a scientific question. In short, it is desirable to turn the challenges of high-dimensional data into assets. The key to doing so is to understand the underlying structure present in the data, and ensure that the learning model incorporates this understanding in the form of strong regularizers or priors. This gives the model selection algorithm the freedom to fit the data as needed, but only so long as it does not violate the known characteristics of the phenomenon under study. While this thesis makes several important contributions in this direction, there will always be questions left to answer, advancements in algorithmic optimization of parameters, and improvements in generalizability. In particular, I believe that the MKL framework can benefit from a more careful analysis of its boosting-like behaviors, (or breakdown thereof,) and that an improved model will expand upon previously demonstrated results. I also believe that machine learning methods will continue to grow in importance in fields that rely heavily on high-dimensional data with complex dependencies. Lastly, I believe that there is potentially a significant benefit to society in the development of modern statistical analysis tools for high dimensional data, and application to problems of large impact, of which neuroimaging analysis of Alzheimer's Disease is but a single example, to which this thesis aims to contribute.

Bibliography

Alzheimer's disease facts and figures. Technical report, Alzheimer's Association, 2007.

H. Arimura, T. Yoshiura, S. Kumazawa, K. Tanaka, H. Koga, F. Mihara, H. Honda, S. Sakai, F. Toyofuku, and Y. Higashida. Automated method for identification of patients with Alzheimer's disease based on three-dimensional MR images. *Academic Radiology*, 15(3):274–284, 2008.

N. Aronszajn. *Theory of reproducing kernels*. Defense Technical Information Center, 1950.

J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95–113, 113 2007.

J. Ashburner and K.J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.

F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.

F. R. Bach, G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *International Conference on Machine Learning*, 2004.

S. Bergsma, D. Lin, and D. Schuurmans. Improved Natural Language Learning via Variance-Regularization Support Vector Machines. In *Conference on Computational Natural Language Learning*, 2010.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006. ISBN 0387310738.

H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239–259, 1991.

A. Burns, P. Luthert, R. Levy, R. Jacoby, and P. Lantos. Accuracy of clinical diagnosis of Alzheimer's disease. *British Medical Journal*, 301(6759):1026, 1990.

S. M. Butler, J. W. Ashford, and D. A. Snowdon. Age, education, and changes in the Mini-Mental State Exam scores of older women: Findings from the nun study. *Journal of the American Geriatrics Society*, 44(6):675–671, 1996.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

E. J. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *International Conference on Machine Learning*, 2010.

R. Cuingnet, M. Chupin, H. Benali, and O. Colliot. Spatial and anatomical regularization of SVM for brain image analysis. In *Advances in Neural Information Processing Systems*, 2010.

R. Cuingnet, E. Gérardin, J. Tessieras, G. Auzias, S. Lehéricy, and M. O. Habert. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, 56 (2):766–781, 2011.

C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick. Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369, 2001.

C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of Aging*, 29(4):514–523, 2008a.

C. Davatzikos, S.M. Resnick, X. Wu, P. Parmpi, and C. M. Clark. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage*, 41(4):1220–1227, 2008b.

C. Davatzikos, F. Xu, Y. An, Y. Fan, and S. M. Resnick. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*, 132(8):2026–2035, 2009.

A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear Programming Boosting via Column Generation. *Machine Learning*, 46(1):225–254, 2002.

L. deToledo Morrell, T. R. Stoub, M. Bulgakova, R. S. Wilson, D. A. Bennett, S. Leurgans, J. Wuu, and D. A. Turner. MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging*, 25(9): 1197–1203, 2004.

B. C. Dickerson, I. Goncharova, M. P. Sullivan, C. Forchetti, R. S. Wilson, D. A. Bennett, L. A. Beckett, and L. deToledo-Morrell. MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. *Neurobiology of Aging*, 22(5):747–754, 2001.

S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins. MRI-Based Automated Computer Classification of Probable AD Versus Normal Controls. *IEEE Transactions on Medical Imaging*, 27(4):509–520, 2008.

D. Erdogmus and J. C. Principe. Generalized information potential criterion for adaptive system training. *IEEE Transactions on Neural Networks*, 13(5): 1035–1044, 2002.

Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39(4):1731–1743, 2008a.

Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos. Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *NeuroImage*, 41(2):277–285, 2008b.

M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference*, volume 4, 2004.

M. C. Fox and J. M. Schott. Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet*, 363(9406):392–394, 2004.

A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Conference on Computational Learning Theory*, 1995.

G. M. Fung and O. L. Mangasarian. A Feature Selection Newton Method for Support Vector Machine Classiacation. *Computational Optimization and Applications*, 28(2):185–202, 2004.

K. Gai, G. Chen, and C. Zhang. Learning Kernels with Radiuses of Minimum Enclosing Balls. In *Advances in Neural Information Processing Systems*, 2010.

P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *International Conference on Computer Vision*, 2009a.

P. V. Gehler and S. Nowozin. Infinite kernel learning. Technical Report 178, Max-Planck Institute for Biological Cybernetics, 2008.

P. V. Gehler and S. Nowozin. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 2836–2843, 2009b.

M. Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14(3):669–688, 2002.

M. Gönen and E. Alpaydin. Localized multiple kernel learning. In *International Conference on Machine Learning*, 2008.

M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

A. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learning ensembles. In *National Conference on Artificial Intelligence*, 1998.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

J. He, L. Balzano, and A. Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, and S. C. Johnson. Spatially augmented LPBoosting for AD classification with evaluations on the ADNI dataset. *NeuroImage*, 48(1):138–149, 2009.

C. Hinrichs, V. Singh, G. Xu, and S. C. Johnson. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*, 48:574–589, 2011.

X. Hua, A. D. Leow, N. Parikshak, S. Lee, M. C. Chiang, A. W. Toga, C. R. Jack Jr., M. W. Weiner, and P. M. Thompson. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. *NeuroImage*, 43(3):458–469, 2008.

X. Hua, S. Lee, I. Yanovsky, A.D. Leow, Y. Y. Chou, A. J. Ho, B. Gutman, A. W. Toga, C. R. Jack Jr., M. A. Bernstein, et al. 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 (4):668–681, 2009.

X. Hua, S. Lee, D. P. Hibar, I. Yanovsky, A. D. Leow, A. W. Toga, C. R. Jack Jr., M. A. Bernstein, E. M. Reiman, D. J. Harvey, J. Kornak, N. Schuff, G. E. Alexander, M. W. Weiner, and the Alzheimer's Disease Neuroimaging Initiative. Mapping Alzheimer's disease progression in 1309 MRI scans: Power estimates for different inter-scan intervals. *NeuroImage*, 51(1):63–75, 2010.

C. R. Jack Jr., R. C. Petersen, Y. Xu, P. C. O'Brien, et al. Rates of hippocampal atrophy correlate with change in clinical status in aging and AD. *Neurology*, 55 (4):484–490, 2000.

R. Jenssen. Kernel entropy component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):847–860, 2010.

T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In *International Conference on Machine Learning*, 2001.

S. C. Johnson, T. W. Schmitz, M. A. Trivedi, M. L. Ries, B. M. Torgerson, C. M. Carlsson, S. Asthana, B. P. Hermann, and M. A. Sager. The Influence of Alzheimer Disease Family History and Apolipoprotein E $\varepsilon 4$ on Mesial Temporal Lobe Activation. *Journal of Neuroscience*, 26(22):6069–6076, 2006.

S. C. Johnson, A. La Rue, B. P. Hermann, G. Xu, R. L. Koscik, E. M. Jonaitis, B. B. Bendlin, K. J. Hogan, A. D. Roses, A. M. Saunders, et al. The effect of TOMM40 poly-T length on gray matter volume and cognition in middle-aged persons with APOE$\varepsilon 3/\varepsilon 3$ genotype. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 2011.

I. T. Jolliffe. *Principal Component Analysis*. Springer New York, second edition, 2002. ISBN 1441929991.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Efficient and accurate ℓ_p-norm multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2009.

M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

S. Klöppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack Jr., J. Ashburner, and R.S. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689, 2008.

O. Kohannim, X. Hua, D. P. Hibar, S. Lee, Y. Y. Chou, A. W. Toga, C. R. Jack Jr., M. W. Weiner, and P. M. Thompson. Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiology of Aging*, 2010.

V. Kolmogorov and Y. Boykov. What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. In *International Conference on Computer Vision*, 2005.

M. Kowalski, M. Szafranski, and L. Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *International Conference on Machine Learning*, 2009.

G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

E. K. Lenzi, R. S. Mendes, and L. R. da Silva. Statistical mechanics based on Renyi entropy. *Physica A: Statistical Mechanics and its Applications*, 280(3): 337–345, 2000.

O. L. Mangasarian and E. W. Wild. Feature selection in k-median clustering. In *Conference on Data Mining, Workshop on Clustering High Dimensional Data and its Applications*, 2004.

S. Minoshima, B. Giordani, S. Berent, K. A. Frey, N. L. Foster, and D. E. Kuhl. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Annals of Neurology*, 42(1):85–94., 1997.

C. Misra, Y. Fan, and C. Davatzikos. 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(4):1415–1422, 2008.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. ISBN 0070428077.

L. Mukherjee, V. Singh, J. Peng, and C. Hinrichs. Learning Kernels for variants of Normalized Cuts: Convex Relaxations and Applications. In *IEEE conference on Computer Vison and Pattern Recognition*, 2010.

T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, 12:419–446, 2003.

J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Verlag, 1999. ISBN 0387987932.

C. S. Ong, A. Smola, and B. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1045–1071, 2005.

F. Orabona, L. Jie, and B. Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *IEEE conference on Computer Vision and Pattern Recognition*, 2010.

F. Pesarin. *Multivariate permutation tests: with applications in biostatistics*. Wiley, 2001. ISBN 0471496707.

R. C. Petersen, R. G. Thomas, M. Grundman, et al. Donepezil and vitamin E in the treatment of mild cognitive impairment. *New England Journal of Medicine*, 352(23):2379–2388, 2005.

J. L. Prince and J. M. Links. *Medical imaging signals and systems*. Pearson Prentice Hall, 2006. ISBN 0130653535.

O. Querbes, F. Aubry, J. Pariente, J. A. Lotterie, J. F. Demonet, V. Duret, M. Puel, I. Berry, J. C. Fort, and P. Celsis. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8): 2036–2047, 2009.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3): 471–501, 2010.

A. Renyi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

A. D. Roses, M. W. Lutz, H. Amrine-Madsen, A. M. Saunders, D. G. Crenshaw, S. S. Sundseth, M. J. Huentelman, K. A. Welsh-Bohmer, and E. M. Reiman. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *The Pharmacogenomics Journal*, 10(5):375–384, 2009.

C. Rudin, I. Daubechies, and R. E. Schapire. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *Journal of Machine Learning Research*, 5:1557–1595, 2004.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2): 197–227, 1990.

P. W. Schofield, M. Tang, K. Marder, K. Bell, G. Dooneief, R. Lantigua, D. Wilder, B. Gurland, Y. Stern, and R. Mayeux. Consistency of clinical diagnosis in a community-based longitudinal study of dementia and Alzheimer's disease. *Neurology*, 45:2159–2164, 1995.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002. ISBN 0262194759.

J. M. Schott, J. W. Bartlett, J. Barnes, K. K. Leung, S. Ourselin, and N. C. Fox. Reduced sample sizes for atrophy outcomes in Alzheimer's disease trials: baseline adjustment. *Neurobiology of Aging*, 31(8):1452–1462, 2010.

M. L. Schroeter, T. Stein, N. Maslowski, and J. Neumann. Neural correlates of Alzheimer's disease and mild cognitive impairment: A systematic and quantitative meta-analysis involving 1351 patients. *NeuroImage*, 47(4):1196–1206, 2009.

D. Shen and C. Davatzikos. Hammer: Heirarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11): 1421–1439, 2002.

P. Shivaswamy and T. Jebara. Maximum relative margin and data-dependent regularization. *Journal of Machine Learning Research*, 11:747–788, 2010.

N. Shock, R. Greulich, and R. Andres et al. Normal human aging: the Baltimore Longitudinal Study of Aging. Washington, DC: US Government Printing Office, 1984.

K. D. Singh, G. R. Barnes, and A. Hillebrand. Group imaging of task-related changes in cortical synchronisation using nonparametric permutation testing. *NeuroImage*, 19(4):1589–1601, 2003.

V. Singh, L. Mukherjee, and M. K. Chung. Cortical Surface Thickness as a Classifier: Boosting for Autism Classification. In *Medical Image Computing and Computer-Assisted Intervention*, 2008.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

S. U. Stich, C. L. Müller, and B. Gärtner. Optimization of convex functions with random pursuit. *Preprint*, 2012. arXiv:1111.0194v2.

P. M. Thompson, M. S. Mega, R. P. Woods, C. I. Zoumalan, C. J. Lindshield, R. E. Blanton, J. Moussai1, C. J. Holmes, J. L. Cummings, and A. W. Toga. Cortical Change in Alzheimer's Disease Detected with a Disease-specific Population-based Brain Atlas. *Cerebral Cortex*, 11(1):1–16, 2001.

V. N. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000. ISBN 0387987800.

P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack Jr. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage*, 39(3):1186–1197, 2008.

S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *IEEE conference on Computer Vision and Pattern Recognition*, 2008.

P. J. Visser, P. Scheltens, and F. R. J. Verhey. Do MCI criteria in drug trials accurately identify subjects with predementia Alzheimer's disease? *Journal of Neurology, Neurosurgery & Psychiatry*, 76(10):1348, 2005.

J. P. Wade, T. R. Mirsen, V. C. Hachinski, M. Fisman, C. Lau, and H. Merskey. The clinical diagnosis of Alzheimer's disease. *Neurology*, 44(1):24–29, 1987.

G. Wahba. *Spline models for observational data*. Society for Industrial Mathematics, 1990. ISBN 0898712440.

L. Wasserman. *All of nonparametric statistics*. Springer-Verlag New York Inc, 2006. ISBN 0387251456.

A. Wimo, L. Jonsson, and B. Winblad. An estimate of the worldwide prevalence and direct costs of dementia in 2003. *Dementia and Geriatric Cognitive Disorders*, 21(3):175–181, 2006.

K. J. Worsley. Local maxima and the expected Euler characteristic of excursion sets of χ^2, f and t fields. *Advances in Applied Probability*, 26:13–42, 1994.

K. J. Worsley, A. C. Evans, S. Marrett, P. Neelin, et al. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, 12(6):900–918, 1992.

K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4(1):58–73, 1996.

Z. Xiang, Y. Xi, U. Hasson, and P. Ramadge. Boosting with spatial regularization. In *Advances in Neural Information Processing Systems*, 2009.

G. Xu, D. G. Mclaren, M. L. Ries, M. E. Fitzgerald, B. B. Bendlin, H. A. Rowley, M. A. Sager, C. Atwood, S. Asthana, and S. C. Johnson. The influence of parental history of Alzheimer's disease and apolipoprotein E ε4 on the BOLD signal during recognition memory. *Brain*, 132(2):383, 2009.

L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *National Conference On Artificial Intelligence*, 2006.

J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *International Conference on Computer Vision*, 2009.

D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen. Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. *NeuroImage*, 2011a.

R. Y. Zhang, A. C. Leon, C. Chuang-Stein, and S. J. Romano. A new proposal for randomized start design to investigate disease-modifying therapies for Alzheimer disease. *Clinical Trials*, 8(1):5–14, 2011b.