1-1-2015

Converting Neuroimaging Big Data to information: Statistical Frameworks for interpretation of Image Driven Biomarkers and Image Driven Disease Subtyping

Bilwaj Krishnanand Gaonkar
University of Pennsylvania, bilwaj@gmail.com

Converting Neuroimaging Big Data to information: Statistical Frameworks for interpretation of Image Driven Biomarkers and Image Driven Disease Subtyping

Abstract
Large scale clinical trials and population based research studies collect huge amounts of neuroimaging data. Machine learning classifiers can potentially use these data to train models that diagnose brain related diseases from individual brain scans. In this dissertation we address two distinct challenges that beset a wider adoption of these tools for diagnostic purposes.

The first challenge that besets the neuroimaging based disease classification is the lack of a statistical inference machinery for highlighting brain regions that contribute significantly to the classifier decisions. In this dissertation, we address this challenge by developing an analytic framework for interpreting support vector machine (SVM) models used for neuroimaging based diagnosis of psychiatric disease. To do this we first note that permutation testing using SVM model components provides a reliable inference mechanism for model interpretation. Then we derive our analysis framework by showing that under certain assumptions, the permutation based null distributions associated with SVM model components can be approximated analytically using the data themselves. Inference based on these analytic null distributions is validated on real and simulated data. p-Values computed from our analysis can accurately identify anatomical features that differentiate groups used for classifier training. Since the majority of clinical and research communities are trained in understanding statistical p-values rather than machine learning techniques like the SVM, we hope that this work will lead to a better understanding SVM classifiers and motivate a wider adoption of SVM models for image based diagnosis of psychiatric disease.

A second deficiency of learning based neuroimaging diagnostics is that they implicitly assume that, `a single homogeneous pattern of brain changes drives population wide phenotypic differences'. In reality it is more likely that multiple patterns of brain deficits drive the complexities observed in the clinical presentation of most diseases. Understanding this heterogeneity may allow us to build better classifiers for identifying such diseases from individual brain scans. However, analytic tools to explore this heterogeneity are missing. With this in view, we present in this dissertation, a framework for exploring disease heterogeneity using population neuroimaging data. The approach we present first computes difference images by comparing matched cases and controls and then clusters these differences. The cluster centers define a set of deficit patterns that differentiates the two groups. By allowing for more than one pattern of difference between two populations, our framework makes a radical departure from traditional tools used for neuroimaging group analyses. We hope that this leads to a better understanding of the processes that lead to disease and also that it ultimately leads to improved image based disease classifiers.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Bioengineering

First Advisor
Christos Davatzikos

Keywords
Alzheimer's disease, biostatistics, heterogeneity analysis, machine learning, support vector machines

Subject Categories
Biomedical | Computer Sciences | Statistics and Probability

CONVERTING NEUROIMAGING BIG DATA TO INFORMATION: STATISTICAL FRAMEWORKS FOR INTERPRETATION OF IMAGE DRIVEN BIOMARKERS AND IMAGE DRIVEN DISEASE SUBTYPING

Bilwaj Gaonkar

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Christos Davatzikos, Ph.D.
Professor, Radiology

Graduate Group Chairperson

Jason Burdick, Ph.D., Professor, Bioengineering

Dissertation Committee

Russell T. Shinohara, Assistant Professor, Department of Biostatistics
Robert Nick Bryan, Emeritus Professor, Radiology
Paul Yushkevich, Associate Professor, Radiology
Li-San Wang, Associate Professor, Pathology

CONVERTING NEUROIMAGING BIG DATA TO INFORMATION: STATISTICAL FRAMEWORKS FOR INTERPRETATION OF IMAGE DRIVEN BIOMARKERS AND IMAGE DRIVEN DISEASE SUBTYPING

Dedicated to the Universe

ACKNOWLEDGEMENT

The PhD degree is the latest step of my educational sojourn which began at kindergarten. While I have been interested in research since childhood, I consider myself fortunate to have been inspired along my path by several illustrious and brilliant people.

I would like to thank my parents in India without whose love, support and encouragement this research would not have been possible. My parents, each in their own way, ignited and nurtured within me the spark of curiosity from an early age. This helped me develop an attitude of endlessly questioning the status quo and eventually culminated in the preparation of this dissertation.

Second, I would like to thank my advisor, Dr. Christos Davatzikos and his wife Parmpi Paraskevi for providing me with a loving and homely environment in the United States. They watched over me, helped me and took a personal interest in my success every step of the way. Just like my parents, they both showed sagacious forbearance, in the face of my perpetual vacillations in regards to both, research and life. For this I am eternally grateful to them.

Third, I would like to acknowledge the myriad contributions of the scientific community at large. The selfless work done by thousands of researchers over hundreds of years is what inspired and enabled me to pursue research as a career. Specific contributions that have lead up to this thesis are too numerous to count. However, the fact remains that none of the research presented here would be possible without the tireless efforts of scientific community in many places and many times.

Lastly this acknowledgement would be incomplete without thanking the many friends and colleagues that I met here in the course of my PhD studies. The peer driven learning environment at the center for biomedical image computing and in the University of Penn-

sylvania at large is a cornerstone of a fulfilling education. Hence, I feel blessed to be a part of this institution.

ABSTRACT

CONVERTING NEUROIMAGING BIG DATA TO INFORMATION: STATISTICAL FRAMEWORKS FOR INTERPRETATION OF IMAGE DRIVEN BIOMARKERS AND IMAGE DRIVEN DISEASE SUBTYPING

Bilwaj Gaonkar

Christos Davatzikos, Ph.D.

Large scale clinical trials and population based research studies collect huge amounts of neuroimaging data. Machine learning classifiers can potentially use these data to train models that diagnose brain related diseases from individual brain scans. In this dissertation we address two distinct challenges that beset a wider adoption of these tools for diagnostic purposes.

The first challenge that besets the neuroimaging based disease classification is the lack of a statistical inference machinery for highlighting brain regions that contribute significantly to the classifier decisions. In this dissertation, we address this challenge by developing an analytic framework for interpreting support vector machine (SVM) models used for neuroimaging based diagnosis of psychiatric disease. To do this we first note that permutation testing using SVM model components provides a reliable inference mechanism for model interpretation. Then we derive our analysis framework by showing that under certain assumptions, the permutation based null distributions associated with SVM model components can be approximated analytically using the data themselves. Inference based on these analytic null distributions is validated on real and simulated data. p-Values computed from our analysis can accurately identify anatomical features that differentiate groups used for classifier training. Since the majority of clinical and research communities are trained in understanding statistical p-values rather than machine learning techniques like the SVM, we hope that this work will lead to a better understanding SVM classifiers and motivate a

wider adoption of SVM models for image based diagnosis of psychiatric disease.

A second deficiency of learning based neuroimaging diagnostics is that they implicitly assume that, 'a single homogeneous pattern of brain changes drives population wide phenotypic differences'. In reality it is more likely that multiple patterns of brain deficits drive the complexities observed in the clinical presentation of most diseases. Understanding this heterogeneity may allow us to build better classifiers for identifying such diseases from individual brain scans. However, analytic tools to explore this heterogeneity are missing. With this in view, we present in this dissertation, a framework for exploring disease heterogeneity using population neuroimaging data. The approach we present first computes difference images by comparing matched cases and controls and then clusters these differences. The cluster centers define a set of deficit patterns that differentiates the two groups. By allowing for more than one pattern of difference between two populations, our framework makes a radical departure from traditional tools used for neuroimaging group analyses. We hope that this leads to a better understanding of the processes that lead to disease and also that it ultimately leads to improved image based disease classifiers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1. Background

Over the past few decades, there has been spectacular advancement in medical image acquisition technology. Due to the reduction in cost and improved scanning techniques there has been an explosive growth in the availability of in vivo neuroimaging data. This growth has lead to the collection of vast amounts of neuroimaging data by several different research groups and consortia around the world. Some of the examples include the Batimore longitudinal study of Aging or BLSA (Kawas et al., 1997), the Philadelphia neurodevelopmental cohort or PNC (Satterthwaite et al., 2014), the Alzheimer's Disease Neuroimaging Initiative or ADNI (Jack et al., 2008) and the Enhancing NeuroImaging Genetics through Meta-Analysis or ENIGMA (Thompson et al., 2014) cohorts. However, these are by no means unique. Several different groups, laboratories and consortia, across the world are collecting similar datasets in an attempt to better understand the structure and function of the living human brain. As such, each of these studies generates massive amounts of neuroimaging data.

Understanding the structure and function of the human brain using these massive neuro imaging data sets presents a fundamental challenge to existing traditional statistical analysis techniques. Each brain image contains an extremely large number of pixels/voxels. Relative

to the total number of voxels in an image, the total number of brain images in a given data set is always small. Thus, the dimensionality of the data, as measured by number of voxels per image, is far higher than the sample size, as measured by total number of images in the study. Traditional statistical techniques are mostly designed to analyze and visualize with the low dimension high sample size data. Hence analysis of high dimension low sample size data, such as the data generated by large scale neuroimaging studies, requires the development of fundamentally new statistical methodologies themselves.

Thus, the development of novel statistical analysis methodologies geared towards large scale neuroimaging studies is the primary topic of this thesis. Next, we present a short summary of the different technologies that must be understood before delving into the main topic of this work. While each of these technologies may themselves be considered an active topic of research, for the purposes of this thesis we utilize standard implementations that are peer reviewed and accepted by the medical image analysis community at large.

1.2. Key technologies for comparing images across populations

Brain imaging data as obtained from the scanner can come from one of many modalities. Further most of the raw data obtained from the scanner cannot be directly used for population wide statistical analysis. Depending on the modality and the quality of the data a sufficient degree of image pre-processing is necessary before delving into group comparisons. In this section we describe some of the common medical imaging modalities as well as commonly used preprocessing steps.

1.2.1. Medical imaging modalities

A diverse array of medical imaging modalities were developed in the last few decades of the twentieth century. We cannot describe the details of every single medical imaging modality out there in this thesis. As such design of imaging protocols in itself is still an active area of research. Nevertheless, we present a brief description of some of the more commonly used modalities in the appendix. One may broadly categorize medical image modalities as structural and functional. Structural imaging modalities such as T1-weighted MRI, T2-weighted MRI, FLAIR MRI, CT and diffusion tensor MRI (DTI) reveal anatomical aspects of the organ under study. Functional imaging modalities are measure temporally dynamic changes in tissue. Examples of such modalities include BOLD fMRI, contrast enhanced T1 MRI and PET. We have briefly described each of these modalities in the appendic. The work presented in this thesis contains experiments using T1-weighted MRI and BOLD-fMRI images. However, the methods presented are applicable to imaging data as well as other high dimensional data.

1.2.2. Image preprocessing

Image pre-processing is necessary step before imaging data can be used for group analysis. The exact nature and number of steps that may be required before the application of a specific algorithm may vary depending on the algorithm itself, the modality under study and the anatomy. Work presented here primarily deals with brain imaging data. Hence, we limit ourselves to brain imaging studies. In the appendix we present a brief description of preprocessing steps used to generate the data used in the thesis.

1.3. The evolution of population based neuroimaging analysis

With the widespread availability of various image acquisition technologies and standardization of acquisition and pre-processing protocols, it has become possible to collect neuroimaging data from entire populations. This has lead to several large scale neuroimaging studies targeted at comparing clinically abnormal populations to normal controls. The images involved are most often structural or functional MR scans. But they may be any other types of imaging data, including but not limited to contrast maps, connectivity maps or some measures derived from other more advanced imaging modalities. In either case, detection and description of imaging differences driving the clinical distinction between populations remains the central question of population neuroimaging. The sheer dimensionality of these data pose a challenge to traditional statistical analysis techniques. This has lead to the development of several methods attempting to address this challenge. The work presented in this thesis also addresses this problem. In order to place our work in context of current literature we present a review of prior art in this section.

1.4. Univariate morphometric methodologies

Given imaging data associated with two clinically different populations (eg. patients and controls) or with two functional conditions (eg. activation and baseline), the simplest question one may ask is, "Which brain regions differ between the two groups?" . Naturally, this lead to the development of a whole set of methods directed at addressing this question. We describe these methods next.

1.4.1. Region based morphometry

Traditional neuroscience dictates that specific anatomical structures in the brain are associated with specific functions. Thus, the first attempts at population based neuroimaging analyses worked off similar hypotheses. These approaches are collectively referred to as region of interest (ROI) based morphometry. In ROI based morphometric analysis, the volume of the whole brain or its subparts is measured by drawing regions of interest (ROIs) on brain images. The volumes of these regions are compared between subjects or across populations. (Giedd et al., 1996; Ge et al., 2002; Giedd et al., 1996) However, this is time consuming and can only provide measures of rather large areas. Smaller differences in volume may be overlooked and there is always the possibility of errors due to inconsistencies between manual segmentations of ROIs. Further, there exists substantial inter rater variability in the manual definition of ROIs. This makes reproducibility difficult. Thus, region based morphometric analysis was eventually superseded by other fully automated methods of analyzing population neuroimaging data.

1.4.2. Deformation based morphometry

Deformtion based morphometry (DBM)(Chung et al., 2001; Gaser et al., 1999; Cao et al., 1999) relies upon directly comparing two or more deformation fields using multivariate statistical tools such as the Hotelling's T^2 test. As such, this constitutes a voxel wise morphometric method. These methods as a whole and DBM in particular do not require apriori knowledge of the ROI to perform analysis. This has certain advantages in terms of being able to detect local structures. For instance, DBM can detect exactly what part of a particular ROI is responsible for most of the anatomical variation within a group. Thus, DBM has certain advantages over traditional ROI based approaches. However, one of the problems with comparing deformation vectors in a voxel wise fashion is that these measurements

incorporate translation in addition to growth. In the context of group difference analysis or the study of temporal variation in brain morphology, we are often more interested in increase/decrease in local volumes rather than displacements themselves. Since DBM aims to measure the relative position of two voxels before and after deformation, rather than volume changes, it provides a relatively indirect measure of volume changes. This lead to the evolution of tensor based morphometry, which looks at the Jacobian of the deformation field instead of the field itself. We describe this next.

1.4.3. Tensor based morphometry

Tensor based morphometry (TBM) uses is another voxel wise morphometric technique that uses the spatial derivatives of deformation fields instead of using the raw field displacements themselves. Since TBM is also a voxel wise technique it does not require apriori segmentation of regions of interest. It inherits the other big advantage of DBM as well, which is the ability to detect localized changes not detectable using anatomical landmarks or ROIs. Several tensorial measures may be constructed from the derivatives of deformation fields but perhaps the most interesting, and the most widely used is the Jacobian of the deformation field and it's determinant. If we assume $\mathbf{u}(\mathbf{x}) \in \mathbb{R}^3$ to be the three dimensional deformation vector associated with the point $\mathbf{x} \in \mathbb{R}^3$ in template space, then the Jacobian is a tensor that may be defined as:

$$J(x) = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_1}{\partial x_2} & \frac{\partial u_1}{\partial x_3} \\ \frac{\partial u_2}{\partial x_1} & \frac{\partial u_2}{\partial x_2} & \frac{\partial u_2}{\partial x_3} \\ \frac{\partial u_3}{\partial x_1} & \frac{\partial u_3}{\partial x_2} & \frac{\partial u_3}{\partial x_3} \end{bmatrix}$$

The nine components of this tensor may be used to measure morphological variability. The determinant of the Jacobian measures expansion or shrinkage of an infinitesimal volume element as it is deformed from the subject space to the template space. Thus, by comparing

Jacobian determinants across subjects we obtain a localized measure of volumetric variability of brain tissue across a population. This is the essence of tensor based morphometry. (Davatzikos et al., 1996; Thompson and Toga, 1998). Other tensorial measures that may be constructed for use with tensor based morphometry include but are not limited to the the strain and vorticity tensors (Chung, 2013). Any number of tensorial measures may be constructed from first or higher order derivatives of deformation fields. Each measure may be associated with a specific type of measurement. Any method that uses such measures to quantify morphological differences between groups may be considered under the title of tensor based morphometry.

1.4.4. Voxel based morphometry

In a loose sense the term voxel based analyses may refer to all methods that depend upon voxel wise comparison of neuroimaging based measurements. In this sense both deformation based and tensor based morphometric analyses are voxel based analyses. However, specifically voxel based morphometry refers to the voxel wise comparison of local concentrations of gray and white matters across populations (Davatzikos et al., 2001; Ashburner and Friston, 2000). The main advantage of using tissue density maps is that they are relatively robust to small registration errors. This is different from DBM or TBM, both of which rely on highly accurate registration of one brain to another. Given the tremendous anatomical variation of the human brain, perfect registration is almost always impossible. As described before, tissue density maps are robust to small registration errors since they are generated using a discrete approach. In a certain sense they are the discretized version of the Jacobian determinant described in the previous subsection. Because of the discrete approach forces local volume preservation between subject image and the tissue density map is essentially guaranteed even in the presence of registration errors. This cannot be said of the Jacobian determinant which relies purely on the deformation field. This is the reason voxel based morphometry is perhaps the most popular method in population neuroimaging analysis.

1.5. Multivariate morphometry: Machine learning in computational neuro-anatomy

A common feature of all the methods presented in the previous section is that they perform statistical analyses on a voxel by voxel basis. Thus, these methods cannot highlight multi-voxel patterns of imaging differences that may drive population stratification. Further this type of statistical analysis offers almost no value in terms of predictive analyses. That is, finding differences between two populations using any one of the above approaches does not rigorously quantify how well imaging may be used for diagnostic or prognostic purposes. Such predictive analytics are very important from a clinical perspective. Both these shortcomings lead to the development of machine learning based multivariate analysis techniques. In this section we present some background on how machine learning based analysis is typically applied in population neuroimaging. We elaborate in the following sections.

1.5.1. Supervised models

In machine learning theory, supervised learning constitutes a set of algorithms that reason from externally supplied instances (training data) to produce general hypotheses, which then make predictions about future instances (test data). Supervised learning makes the assumption that the probability distributions associated with training and testing data are identical. Supervised learning methods are perhaps the most widely applied machine learning algorithms in biological data analysis. In the context of medical imaging several authors have applied supervised machine learning techniques to make disease diagnosis from imaging data. While the exact analysis pipelines used for generating MRI-based diagnostic models using machine learning are quite varied they generally include the following steps:

1. Collecting a training data set: The objective of this step is to acquire a sufficiently large number of brain image scans from clinically well characterized subjects. Clinical properties used in defining subject categories may include diagnoses, pathological measures or even test scores, which can be used as the gold standard for the classification problem. However, it should be remembered that these definitions form the 'gold standard' which will be used by the classification algorithm for learning a model of the disease in the subsequent stages.

2. Feature extraction from raw imaging data: Just like the statistical morphometric techniques described in the previous section of this chapter, learning using standard machine learning tools requires that imaging data be pre-processed to be comparable across populations. As described earlier, several different registration based measures may be used for this purpose. In the case of structural MRI imaging data we use tissue density maps for machine learning algorithms throughout this thesis. The relevant tissue density map is obtained by pre-processing image informations from the required tissue type (Davatzikos et al., 2001). The map is then vectorized to obtain a feature vector that is used by standard machine learning techniques. Thus, in this thesis, features refer to individual voxel intensities in the RAVENS map images and each feature vector represents an entire image.

3. Dimensionality reduction and feature selection: Most of the data generated by neuroimaging studies contains at best a few hundred samples. However, the image associated with each sample is constituted of millions of voxels. Thus, one can view the dataset as a collection of a few hundred points in a space which has at least a million dimensions. The dimensionality can thwart any effort to estimate the distribution of such a point cloud using standard probablistic modeling techniques. Thus, several authors suggest reducing the dimensionality using one of several dimensionality reduction techniques. Such reduction inevitably causes some loss of signal. However, if a dimensionality reduction technique can effect a a greater reduction in the 'noise' of inter subject variability, then there is a potential payoff in increased classification accuracy.

4. Model training and optimization: The classifier uses the training data and the known

labels to learn a rule to separate the classes. Several classification algorithms may be used in this step. One of the most widely used method for diagnosis of neuropsychiatric disease using imaging data is the support vector machine (SVM). (Boser et al., 1992; Vapnik and Vapnik, 1998; Burges, 1998). The maximum margin formulation of the SVM algorithm generalizes well to high dimensional spaces. Hence, the SVM is capable of learning using high dimension low sample size (HDLSS) data of neuroimaging. Other algorithms that have been used for neuroimaging based classification include the relevance vector machine (RVM), neural networks and regularized linear discriminant analysis. Regardless of the choice of algorithm, there always exist a set of parameters than need to be tuned to fit the model to the data. The most common way of tuning these parameters is cross-validation. N-fold cross-validation involves randomly dividing the entire training data set into N subgroups and then training the algorithm with specific parameters on N-1 subgroups and testing on the left-out subgroup. This process is repeated by leaving each of the sub-groups out one at a time and estimating the average error over all the runs. The model (or the parameters) that gives the best accuracy is picked as the final model.

5. Application to test data: The final step in a typical supervised machine learnng pipeline is to apply the learned rule to a completely new data set that has been pre-processed exactly like the training data set. Effectively, this amounts to testing the generalizability of the model fitted in step 4 onto new data. Sometimes independent data may not be available due to limited sample sizes for testing and training. In such cases cross validation alone may be used to estimate model performance. However, it is important to distinguish between cross validation used for tuning the model and cross validation used for evaluating it. Model selection should be done using an inner cross validation loop while test accuracies are evaluated using an outer loop. Regardless of cross validation, the best approach is to test the trained model using an independent, hitherto unused test dataset.

Supervised classification technology has been used in for the diagnosis of several different diseases from brain images. Several authors have used supervised classification to diagnose

Alzheimer's disease from neuroimaging data. (Klöppel et al., 2008b; Davatzikos et al., 2011; Fan et al., 2007; Filipovych et al., 2012; Gaonkar and Davatzikos, 2013; Varol et al., 2012) A detailed review of the work on supervised classification of Alzheimer's disease is presented in (Cuingnet et al., 2011). Several authors have applied supervised classification to the diagnosis of schizophrenia and psychosis (Sun et al., 2009; Ho et al., 2011; Kawasaki et al., 2007; Koutsouleris et al., 2009). Other authors have applied this technology for developing imaging based biomarkers for autism (Ecker et al., 2010; Jiao et al., 2010), Huntingdon's disease (Klöppel et al., 2008a), Parkinson's disease.(Das, 2010) and fronto temporal dementia (Davatzikos et al., 2008).

1.5.2. Unsupervised and semi supervised models

The second major thrust of machine learning is the development of unsupervised learning methods. The key difference between supervised and unsupervised learning is the absence of excplicitly defined labels. Unsupervised learning focuses on uncovering latent structure in the data through the application of appropriate algorithms. The neuroimaging community has widely adopted unsupervised machine learning methods in the analysis of resting state functional imaging data. For instance independent components analysis and its myriad variants have been successfully used to tease out functional brain networks from data (Calhoun et al., 2001; McKeown et al., 1997; McKeown, 2000). Typical application of these types of learning methods includes:

1. Collecting the dataset. As in the supervised case, the objective of this step is to obtain a large set of brain images. However, unlike the supervised case these images may not correspond to groups of individuals with specific phenotypic characteristics.

2. Feature extraction and dimensionality reduction. As in supervised learning, feature extraction is an important aspect of unsupervised methods. Raw imaging data needs to be adequately pre-processed to ensure feature correspondences and a relatively noise free

dataset. Dimensionality reduction may be done if it is found to increase the signal to noise ratio.

3. Unsupervised analysis. Finally, one may apply the chosen method of unsupervised data analysis, to the processed data. In the case of independent components analysis applied to fMRI data, the analysis produces a set of 'independent component' images that are thought to correspond to resting state functional brain networks. Clustering of raw imaging data can yield insights into anatomical variation amongst individuals in a population.

In general, the objectives of unsupervised analysis tend to be more exploratory than discriminative. Perhaps this explains the limited application of unsupervised analysis to the exploration of disease effects or population wide differences. However, we believe that application of the unsupervised learning paradigm to explore population wide group differences represents a hitherto untapped opportunity that has the potential to generate several exciting insights. We explore this possibility in this dissertation.

1.6. Key challenges in using machine learning for computational neuroanatomy

While the application of machine learning techniques, both supervised and unsupervised presents an unprecedented opportunity for gaining insights from large neuroimaging data, it also presents substantial challenges. In this section we outline these challenges. In the following section we briefly discuss how this dissertation addresses some of these key challenges.

1.6.1. High dimensionality and low sample size

Brain images typically contain a large number of voxels. The number of non zero voxels in a typical T1- MR scan of the brain numbers in the hundreds of thousands, even when the image is downsampled and pre-processed. With improvements in scanner technology this number has been increasing continuously. On the other hand most population based neuroimaging studies of disease process will scan, at most a thousand individuals. A large majority of studies collect fewer than a hundred scans including patients and controls. Consequently, the majority of neuroimaging datasets that may be used for supervised or unsupervised machine learning analyses are beset by a situation where the number of measurements per sample (that is, the number of voxels) far outstrips the total number of samples in the study. In machine learning parlance this is known as the high dimension low sample size problem. Since the majority of theoretical developments in machine learning and statistics focus on data with high sample size and low dimensionality , the properties of high dimensionality are often poorly understood (Clarke et al., 2008).

The support vector machine algorithm is a specialized technique that has been found to outperform competing approaches in the high dimension low sample size setting (Statnikov et al., 2008). The support vector machine (SVM) is a powerful binary classifier rooted in statistical learning theory that can theoretically achieve a global optimum solution (convex optimization) and bypass the curse of dimensionality through the use of a maximum margin criterion for classification (Clarke et al., 2008; Statnikov et al., 2008; Jain et al., 2000). The SVM provides a way to control model complexity independent of dimensionality and offers the possibility to construct generalized, non-linear predictors in high-dimensional spaces using a small training set (Jain et al., 2000; Clarke et al., 2008).

Hence, it is ideal for use in the neuroimaging setting. Indeed several authors have used the SVM for creating diagnostic imaging based biomarkers (Klöppel et al., 2008b; Davatzikos et al., 2011; Fan et al., 2007; Filipovych et al., 2012; Gaonkar and Davatzikos, 2013; Varol

et al., 2012). Thus, this dissertation addresses the problem of interpreting of support vector machine models used in neuroimaging analysis.

1.6.2. Lack of interpretability of supervised models

While SVMs have been successfully used to create neuroimaging based biomarkers, they have typically been regarded as blackboxes. The question of interpretability of SVM models has received substantially lesser attention in neuroimaging literature, as compared to their application for biomarker development. Yet, it is extremely important to understand in a mathematically rigorous fashion : What voxels/regions does a particular supervised model of disease rely upon to make predictions/diagnoses? Answering this question is imperative from the point of view of understanding the physiological basis on which classifier predictions are based. It is only recently that this question is being looked at in neuroimaging literature. In this thesis we develop and validate a mathematically rigorous p-value based analytical approach to interpret support vector machine models used in neuroimaging. Unlike competing approaches which produce somewhat arbitrary weight values or supervoxels our approach produces a p-value map similar to that produced by VBM or TBM. That is we create a p-value map indicating which brain regions contribute significantly to the classification. The p-value is a well defined, well understood in multiple communities and a mathematically rigorous way to quantify statistical significance. Thus, it is naturally advantageous to interpret SVM models in terms of p-values as opposed to weights or supervoxels. Further, the p-value maps produced by our method can detect multivariate effects in the data and present an advantage over the regular VBM or TBM in terms of morphometric analysis. The next two chapters of this thesis describe in detail this framework for interpreting diagnostic SVM models used in neuroimaging analysis.

1.6.3. Insensititivity to heterogeneity

SVMs and other supervised learning approaches lead to powerful image based diagnostics However, the driving philosophy behind the majority of such analyses is that a single imaging pattern can distinguish between phenotypically distinct populations. However, diseases like schizophrenia and autism, are known to be clinically heterogeneous. For instance schizophrenia can be subdivided based on its symptomatic presentation into positive and negative subtypes (Andreasen and Olsen, 1982); autism spectrum disorders present no clear pattern of brain deficits due to heterogeneity (Amaral et al., 2008) and even Alzheimer's disease may be divided into distinct subtypes (Rabinovici et al., 2008). Given the complexity of the human brain, the subjectivity of clinical scoring, and the ample evidence for heterogeneity in behavioral symptomatology, it is likely that multiple sub-types/patterns of brain changes are associated with a particular population wide phenotypic/clinical differentiation, such as a disease or a dimension of cognitive impairment. Thus, image driven SVM classifiers, which make the single pattern assumption cannot match the diagnostic capabilities of clinicians. Humans combine a large number of observed facts to make a decision, wheareas the only data that the SVM has is an image. Thus, it is likely that the failure of SVM based imaging biomarkers is at least in part driven by the fact that; what is generally defined as a single disease by humans, is in fact the manifestation of several distinct pathological processes. Automated diagnostics aside, there do not even exist tools to explore such heterogeneity using imaging data. This is because population based image analysis tools like VBM are also based on the assumption that a single pattern of brain deficits drive pathology. While this assumption is convenient, it misses the tremendous opportunity that imaging data offer for objectively disentangling disease heterogeneity. Heterogeneity, as defined for the purposes of this dissertation, refers to the existence of subpopulations of patients that are clinically distinct. The aim of imaging based heterogeneity analysis is to identify the subpopulation structure by identifying distinct patterns of brain deficit associated with each subpopulation.

A data driven exploration of disease heterogeneity has the potential to improve not only biomarkers but also fundamentally advance our knowledge of the disease process itself. Thus, in this dissertation we address this we develop an algorithm for image driven exploration of disease heterogeneity. Our technique constitutes an application of unsupervised machine learning analysis and generates clinically relevant sub-types of disease using a data driven approach.

1.7. Addressing the challenges : Structure of the thesis

In this section we highlight how each chapter of this dissertation addresses the challenges presented above. In doing so we also present an overview of the work presented in each chapter.

1.7.1. Chapter 2

Material presented in the next chapter directly addresses the challenge of interpreting SVM models used in neuroimaging analysis using a statistical p-value based framework. As described earlier, SVM models have shown great promise for population based pattern analysis and classification of neuropsychiatric diseases. The interpretation of SVM models can allow us to 1) understand the basis of classifier decisions and 2) highlight the combination of brain regions used by the SVM to make a diagnostic/prognostic decision. The p-values produced by our model provide a multivariate alternative to the standard mass univariate methods used for making population wide statistical comparisons between using brain images. As such, voxel based analysis and related methods, cannot detect multivariate patterns associated with group differences while the SVM based morphometric formulation from chapter 2 can. We develop our theoretical framework as an asymptotic approximation of the permu-

tation testing procedure applied on SVM weight vector components. Ultimately, this yields an inference machinery that operates by comparing components of SVM weight components to normal distributions whose variances can be computed analytically from the data.

1.7.2. Chapter 3

The formulation developed in chapter 2 produces p-values that can detect univariate as well as multivariate effects. However, it ignores a fundamental facet of SVM theory; namely the SVM margin. Consequently, the use of this framework for interpreting SVM models or for morphometric analysis leads to extremely conservative inference. We address this deficiency in chapter 3 by further developing the theory presented in chapter 2 to with incorporation of the SVM margin. Specifically, we develop a statistic that explicitly accounts for the SVM margin. Further, we show that null distributions associated with this statistic are also asymptotically normal. We delineate the advantages of using this margin based statistic for interpreting SVM models. Our experiments show that this statistic is a lot less conservative as compared to weight based permutation tests and also less sensitive to variation in the extent of imaging differences between groups being studied. Ultimately, this new statistic enables us to better understand the multivariate patterns that the SVMs uses for neuroimaging based classification. Chapters 2 and 3 mainly address the challenge of interpreting SVM models used for neuroimaging based diagnostics.

1.7.3. Chapter 4

While chapters 2 and 3 focus on interpreting supervised machine learning models, chapter 4; explicitly deals with the challenge of imaging based exploration of disease heterogeneity. As opposed to chapters 2 and 3, which are interpretative, the analytic frameworks presented in chapter 4 are exploratory. While chapters 2/3 are based on the assumption that a

single pattern of imaging deficit drives the clinical difference between cases and controls, the analyses presented in this chapter allows for the existence of multiple such patterns of deficit. Thus, it explicitly accounts for the presence of heterogeneity in the definition of neuropsychiatric disease. Given the complexity of the human brain and the subjectivity of clinical scoring; it is likely that multiple different patterns of imaging deficit drive specific population wide phenotypic differences. Theoretical developments in this chapter utilize unsupervised machine learning analyses to define case sub-populations using imaging data. The proposed approach can highlight imaging patterns present only in sub populations of a group of cases rather than a 'common denominator' pattern of difference between cases and controls.

CHAPTER 2

Interpreting supervised learning models : An analytic inference framework that approximates permutation testing for SVM classifiers used in neuroimaging

2.1. Introduction

With the availability of cheap computational power and the execution of large population wide neuroimaging studies, it became feasible to attempt the development of imaging based diagnostics for neurologic and neuropsychiatric disorders. However, diagnoses of such disorders from brain images is distinct from traditional radiology. While traditional radiologic diagnostics rely on relatively localized intensity and spatial abnormalities in the brain (eg. a tumor or a stroke), the imaging signatures of neuropsychiatric disorders tend to be far more subtle. Typically, neuropsychiatric diseases are characterized by a combination of morphological and functional differences in several different regions of the brain. Machine

learning tools can 'learn' these complex multivariate patterns of imaging abnormalities from large imaging data-sets and then use them for imaging based diagnosis. The potential for creation of individualized imaging based diagnostic and prognostic biomarkers for neuropsychiatric disorders has led to a tremendous amount of attention being directed to supervised machine learning analysis in the neuroimaging literature over the past decade. However, development of imaging based diagnostics has been challenging because, neuroimaging studies collect a relatively small number of samples, typically in the hundreds and rarely above 1000 and yet each image is constituted of millions of voxels. Thus, neuroimaging data are 'high dimensional with low sample size'. Most learning techniques are designed to operate in the setting where number of samples far outstrips the number of measurements. However, some algorithms like the support vector machine (SVM) have been shown to operate effectively in the high dimension low sample size domain (Clarke et al., 2008). This is perhaps why, this is the most widely used algorithm for developing neuroimaging based diagnostics. However, interpretation of support vector machine models in terms of well understood statistical paradigms such as p-values has remained a key challenge. Addressing this challenge is the thrust of this chapter. With this objective, we describe the SVM algorithm in the next section.

2.2. The key challenge: Interpreting support vector machines (SVM) classification models in classification

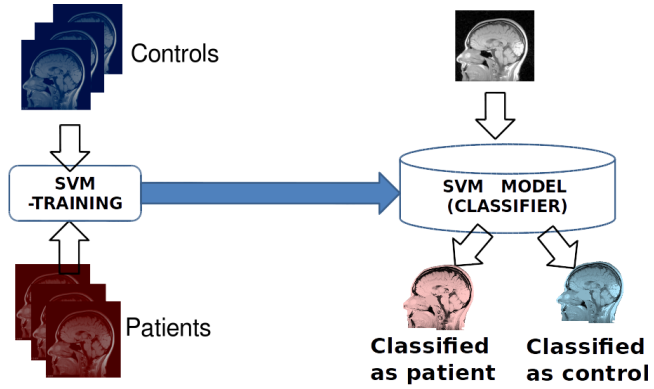
The support vector machine (SVM) is a powerful binary classifier rooted in statistical learning theory that can theoretically achieve a global optimum solution (convex optimization) and bypass the curse of dimensionality (Clarke et al., 2008). The support vector machine attempts to learn a model from data by finding the largest margin hyperplane that separates data from different conditions (e.g. baseline/activation) or groups (e.g. pa-

tients/controls)(see figure 1a). The process of finding this hyperplane using data with known labels(condition, group, etc.) is known as training. Now if data with an unknown label (test data) is presented, the hyperplane found by the SVM is used to estimate whether it belongs to a patient or to a control. To apply SVMs, individual data are treated as points located in a high dimensional space(see figure 1b). Figure 1c illustrates the concept of the algorithm in an imaginary 2D space: dots and crosses represent imaging scans taken from two groups or conditions. Even though the two groups cannot be separated on the basis of values along any one dimension the combination of two dimensions gives perfect separation. This corresponds to the situation where a single anatomical region may not provide the necessary discriminative power between groups, whereas the multivariate SVM can still find the relevant hyperplane.

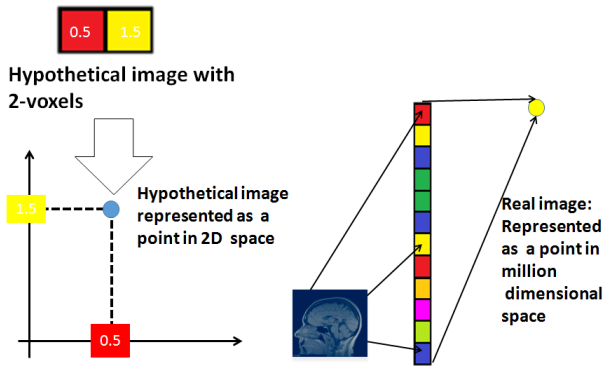
To apply SVMs in neuroimaging data, we convert an image with d voxels into a vector whose j^{th} component is equal to the measurement at the jth voxel in the pre-processed image. Thus we re-organize the ith image into a d-dimensional point that lives in $\mathbb{R}^d$. Let us denote the ith point by $\mathbf{x}_i$ where $i \in 1, ..., m$ indexes all subjects in the study. In most imaging studies we also have a label associated with each image which tells us whether the image belongs to a patient or a control subject. We denote these labels by $y_{(i)} \in \{+1, -1\}$. Then the support vector machine finds 'hyperplane coefficients' denoted by $\mathbf{w}^*$ and b^* such that:

$$\begin{aligned} \{\mathbf{w}^*, b^*\} = argmin_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m} \xi_i \\ suchthat.\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \ \ \forall i = 1, ..., m \\ \xi_i \geq 0 \ \ \forall i = 1, ..., m \end{aligned} \tag{2.1}$$

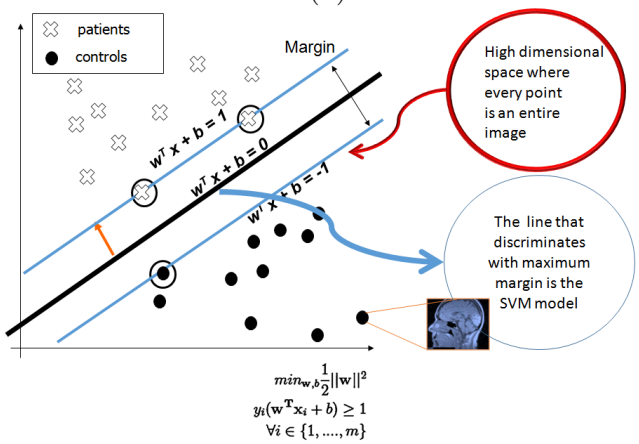
The ξ_i's are slack variables that allows the learnt SVM model to be robust by allowing for a small number of mis-classifications. The weight vector $\mathbf{w}^*$ represents the SVM model. Since the data $\mathbf{x}_{(i)}$ are in the voxel space (one voxel per dimension), the weight vector $\mathbf{w}^* \in \mathbb{R}^d$ is an image. Because the classifier is multivariate by nature, the best combination of all dis-



(a)



(b)



(c)

Figure 1: (a) The concept of imaging based diagnoses using SVMs (b) Images as points in high dimensional space (c) The maximum margin principle of classification used in SVMs

criminating voxels as a whole is identified by this weight vector based discriminative image. These discriminative maps are multivariate. Understanding what regions are most relevant to the classification/diagnosis rendered by the SVM, requires a statistical methodology to interpret these maps. Since the maps are based on a multivariate classifier, popular univariate methods (Ashburner and Friston, 2000) do not offer a direct solution. These maps have several other desirable properties which typical univariate analysis does not provide (Davatzikos, 2004). Until now the use of SVM based discriminative maps in neuroscience has been limited because these maps do not directly answer a critical question "What is the probability that a particular image voxel would have a weight vector component at least as large as the one observed in an experiment due to pure chance alone?" To answer such a question one needs to establish a null distribution on the weight vector components at each image voxel. An empirical approach for obtaining such a null distribution is through the use of permutation tests. We describe this permutation testing approach in the next section.

2.3. Addressing the challenge: Permutation testing to interpret SVM models

Permutation testing can be used to establish a null distribution on the weight vector components at each image voxel. The permutation testing procedure is illustrated in figure 2. In figure 2, the dots denote controls and the crosses denote patients. The first step involves generation of a large number of shuffled instances of data labels by random permutations. Each shuffled instance is used for training one SVM. This generates one hyperplane parameterized by the corresponding vector $\mathbf{w}$, for each instance of shuffled labels. Thus, for a particular component of $\mathbf{w}$, we have a set of possible values. Each value in this set corresponds to a specific shuffling of the labels. These values represent the empiricaly obtained

null distribution of that component of $\mathbf{w}$. Since each component of $\mathbf{w}$ corresponds to a specific voxel location in the original image space we now have a null distribution associated with every voxel in the image space. Comparing each component of $\mathbf{w}^*$ with the corresponding null distribution allows us to estimate statistical significance. It is obvious that running 1000 permutation tests requires training 1000 support vector machine classifiers. This can consume a considerable amount of computing time as well as memory and hard disk space. Further 1000 permutations can allow for a p-value resolution of 0.001. If a lower resolution is needed, for multiple comparisons correction analysis or for feature ranking with p-values, a higher number of permutations are required. Even if one considers 10000 or a 100000 permutations the computational time and memory requirements can quickly become unfeasible. To add to this complexity of population based neuroimaging studies often implies that analysis be repeated, for different population subsets, different pre-processing settings for with different data labeling and even with local 3D windows in the same image (Etzel et al., 2013). Resource requirements for performing permutation testing for each of these possible parameter setting in a neuroimaging study are very large and usually beyond the capacity of even large imaging laboratories. In the future, ever increasing scanner resolutions and sample sizes are likely to exacerbate these problems even further. In contrast to SVM permutation analysis traditional univariate methods (Ashburner and Friston, 2000) can run in a few minutes. This has lead to a much wider adoption of VBM as compared to multivariate analysis despite known advantages of the latter. In order to address this issue, in the following section, we present an analytic approximation of SVM permutation testing. The analytic approximation we present can effectively generate p-value maps identical to those produced by actual permutation tests without the computational overhead as long as the data used are high dimensional with low sample sizes. The following sections detail our approach.

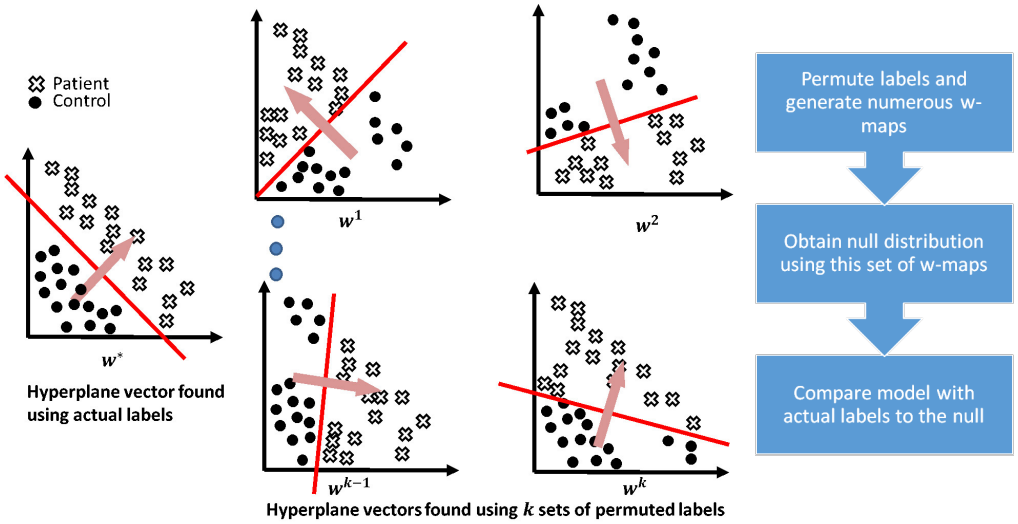


Figure 2: Illustration of the permutation testing procedure

2.4. Analytical framework approximating permutation tests

The primary thrust of this chapter is to lay down an analytic framework for interpreting SVM models. We do this by showing that the permutation testing procedure described above can be replaced by an analytic framework based on the Gaussian distribution that can be used to interpret SVM models in a small fraction of the time it takes for performing the actual permutation tests. Our approximation is based on core support vector machine theory and certain facts that empirically apply to high dimensional medical imaging data.

2.4.1. The case of balanced data

We begin with SVM models trained on datasets with equal numbers of positively and negatively labeled samples.

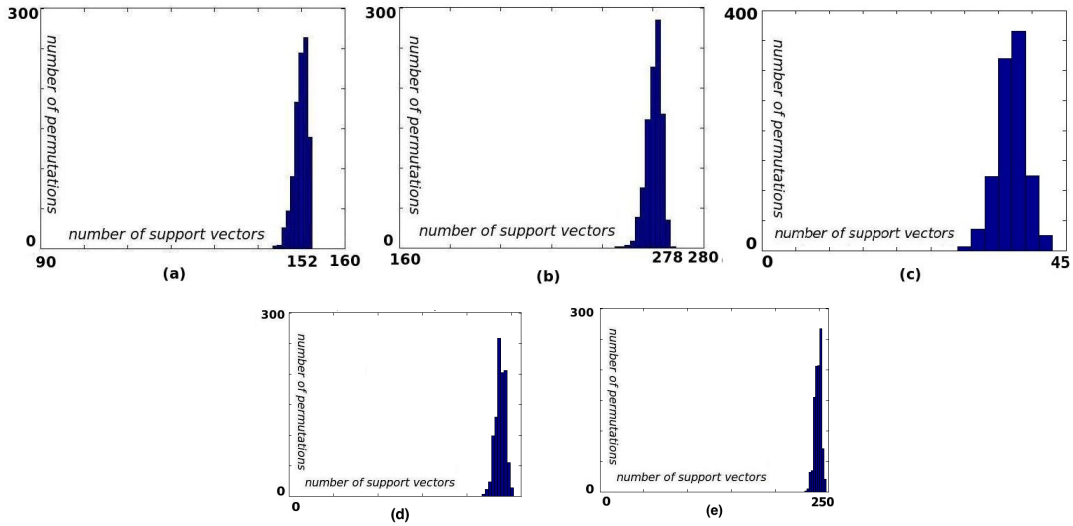


Figure 3: For most permutations the number of support vectors in the learnt model is almost equal to the total number of samples (a) simulated dataset(b) real dataset with Alzheimer's patients and controls (c) real dataset with liars and truth tellers (d) the cancer genome atlas gene expression data (e) the cancer genome atlas methylation data

We start by noting that VC-theory (Vapnik and Vapnik, 1998) dictates that linear classifiers shatter high dimension low sample size data. For example, less than 3 non-collinear points labeled using any combination of positive and negative labels can always be separated by a line in 2D space. When the dimensionality is in the millions and the sample sizes are in the hundreds one can always find 'hyperplanes' (the high dimensional analogue to lines) that can separate any possible labelling of points. Thus, when using linear SVMs, for any permutation of $\mathbf{y}$, one can always find a separating hyperplane that perfectly separates the training data. This allows us to use the hard margin support vector machine formulation from (Vapnik and Vapnik, 1998) instead of (2.1) for further analysis in this chapter. The hard margin support vector machine (see Vapnik and Vapnik (1998)) is written as:

$$\begin{aligned} min_{\mathbf{w},b}\frac{1}{2}||\mathbf{w}||^2 \\ subj.to.\ \ y_i(\mathbf{w^T x}_i + b) \geq 1 \\ \forall i \in \{1, ..., m\} \end{aligned} \tag{2.2}$$

It is required (see Bishop (2007)) that for the 'support vectors', indexed by $j \in \{1, 2, .., n_{SV}\}$, we have $\mathbf{w^T x}_j + b = y_j \quad \forall j$. Now, if all our data were support vectors this would allow us to write the constraints in optimization (2.2) as $\mathbf{Xw} + \mathbf{J}b = \mathbf{y}$ where $\mathbf{J}$ is a column matrix of ones and $\mathbf{X}$ is a super long matrix with each row representing one image. Since the labels generated through random permutations, typically do not correspond to a fundamental group difference that may be highlighted using imaging (or other types) of high dimensional data, SVM models trained using randomly permuted labels tend to overfit. To overfit a particular dataset, the model essentially stores all labels and data. That is, every sample is treated as a support vector. We observed this behavior in medical imaging datasets as well as genomic and epigenomic expression data that we investigated (figure 3). Accounting for this behavior allows us the latitude of approximating SVM solutions for a majority of the random permutations using a much simpler formulation. In fact for most permutations we can solve the following simpler optimization instead of (2.2):

$$\begin{aligned} min_{\mathbf{w},b} ||\mathbf{w}||^2 \\ subj.to. \;\; \mathbf{Xw} + \mathbf{J}b = \mathbf{y} \end{aligned} \tag{2.3}$$

The above formulation is exactly the same as an LS-SVM (Suykens and Vandewalle, 1999). This equivalence between the SVM and LS-SVM for high dimensional low sample size data, with random labeling, was also previously noted by (Ye and Xiong, 2007) where it was based on observations about the distribution of such data as presented in (Hall et al., 2005). This equivalence proves very useful because the LS-SVM, (2.3) can be solved in the closed form (Suykens and Vandewalle, 1999). We show next how to derive the dual problem and solve for dual variables '$\boldsymbol{\alpha}$'. One can use the method of Lagrange multipliers to solve (2.3). We introduce the dual variables $\boldsymbol{\alpha} \in \mathbb{R}^m$ as is standard procedure to yield the Lagrangian:

$$\mathcal{L}(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + \boldsymbol{\alpha}^T(\mathbf{Xw} + \mathbf{J}b - \mathbf{y}) \tag{2.4}$$

Setting $\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, b) = 0$, $\frac{\partial}{\partial \mathbf{b}}\mathcal{L}(\mathbf{w}, b) = 0$ and $\frac{\partial}{\partial \boldsymbol{\alpha}}\mathcal{L}(\mathbf{w}, b) = 0$, and solving for $\mathbf{w}$ yields the

following system of equations:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \mathbf{X}^T \boldsymbol{\alpha}, \tag{2.5}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \mathbf{J}^{\mathbf{T}} \boldsymbol{\alpha} = 0, \tag{2.6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}} = 0 \quad \rightarrow \quad \mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{J}b = \mathbf{X}\mathbf{X}^{\mathbf{T}} \boldsymbol{\alpha} + \mathbf{J}b \tag{2.7}$$

This yields a system of simultaneous equations:

$$\begin{bmatrix} 0 & \mathbf{J}^T \\ \mathbf{J} & \mathbf{X}\mathbf{X}^{\mathbf{T}} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{y} \end{bmatrix} \tag{2.8}$$

Now note that we can compute:

$$\begin{bmatrix} 0 & \mathbf{J}^T \\ \mathbf{J} & \mathbf{X}\mathbf{X}^{\mathbf{T}} \end{bmatrix}^{-1} \tag{2.9}$$

as being:

$$\begin{bmatrix} (-\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J})^{-1} & -\left(-\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1} \\ -(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J}\left(-\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J}\right)^{-1} & (\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1} + (\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J}\left(-\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1}\mathbf{J}\right)^{-1}\mathbf{J}^T(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-1} \end{bmatrix}$$

The inversion relies on the assumption that the kernel matrix $\mathbf{X}\mathbf{X}^{\mathbf{T}}$ is invertible. We can use this inverted form in order to solve for both $\boldsymbol{\alpha}$ and b. Specifically, we can write the dual variables $\boldsymbol{\alpha}$ as:

$$\boldsymbol{\alpha} = [(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-\mathbf{1}} + (\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{J}(-\mathbf{J}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-\mathbf{1}}\mathbf{J})^{-\mathbf{1}}\mathbf{J}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}})^{-\mathbf{1}}]\mathbf{y} \tag{2.10}$$

These are the dual variables in our formulation. The expression for $\mathbf{w}$ then follows from

the dual variables as:

$$\mathbf{w} = \mathbf{X^T}[(\mathbf{XX^T})^{-1} + (\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}]\mathbf{y} = \mathbf{Cy} \tag{2.11}$$

where we have defined:

$$\mathbf{C} = \mathbf{X^T}[(\mathbf{XX^T})^{-1} + (\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}] \tag{2.12}$$

Note that this expresses each component w_j of $\mathbf{w}$ as a linear combination of y_j's. Thus, we can hypothesize about the probability distribution of the components of $\mathbf{w}$, given the distributions of y_j. If we let y_j attain any of the labels (either $+1$ or -1) with equal probability, we have a Bernoulli like distribution on y_j with $E(y_j) = 0$ and $Var(y_j) = 1$ (the theory can be readily extended in the case of unequal priors). We first present the theory from the point of view of equal priors and then present it for unequal priors. Since, (2.11) expresses $\mathbf{w}$ as a linear combination of these y_j we have:

$$E(w_j) = 0 \qquad Var(w_j) = \sum_{i=1}^{m} C_{ij}^2 \tag{2.13}$$

where C_{ij} are the components of the matrix $\mathbf{C}$. Further, the variance of each component of $\mathbf{w}$, is controlled by the rows of the matrix $\mathbf{C}$.Thus:

$$Var(w_j) = \sum_{i=1}^{m} C_{ij}^2 \tag{2.14}$$

These predicted variances agree well with variance estimates obtained from the actual permutation testing (see figure 5). At this point, we have an analytical method to approximate the mean and the variance of the null distributions of components w_j of $\mathbf{w}$ (that would otherwise be obtained using permutation testing). To uncover the probability density function (pdf) of w_j, we use the Lyapunov central limit theorem. We show that when the number of subjects is large, the p.d.f of w_j may be approximated by a normal distribution. To this

end, from (2.13) and (2.12), we have:

$$w_j = \sum_{i=1}^{m} C_{ij} y_i = \sum_{i=1}^{m} z_i^j \tag{2.15}$$

where we have defined a new random variable $z_i^j = C_{ij} y_i$ which is linearly dependent on y_i. We can infer the expectation and variance of z_i^j from y_j as:

$$\mathrm{E}(z_i^j) = 0 = \mu_i \quad Var(z_i^j) = C_{ij}^2 \tag{2.16}$$

Thus, z_i^j are independent but not identically distributed and w_j are linear combinations of z_j^i. Then according to the Lyapunov central limit theorem(CLT) w_j is distributed normally if:

$$\lim_{m\to\infty} \frac{1}{\left[\sqrt{\sum_{i=1}^{m} Var(z_i^j)}\right]^{2+\delta}} \sum_{k=1}^{m} \mathrm{E}\left[\, |z_k^j - \mu_k|^{2+\delta} \,\right] = 0 \;\; for\ some\ \ \delta > 0 \tag{2.17}$$

As is standard practice we check for $\delta = 1$.

$$E\left[\, |z_k^j - \mu_k|^{2+\delta} \,\right] = (1/2)| + C_{kj} - 0|^{2+\delta} + (1/2)| - C_{kj} - 0|^{2+\delta} = C_{kj}^3 \tag{2.18}$$

Thus, we can write the limit in (2.17) as:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} C_{kj}^3}{\left[\sqrt{\sum_{i=1}^{m} C_{ij}^2}\right]^3} \tag{2.19}$$

Now note that the binomial expansion of the denominator would contain cross terms that are not a part of the numerator. Thus, in practice, this limit almost always goes to zero and the the Lyapunov CLT applies. We present a more detailed analysis of a related limit in the latter sections. When this limit is zero it allows us to approximate the distribution of individual components of **w** using the normal distribution as:

$$w_j \xrightarrow{D} \mathcal{N}(0, \sum_{i=1}^{m} C_{ij}^2). \tag{2.20}$$

These predicted distributions fit actual distributions obtained using permutation testing very well (figure 5). Thus w_j's computed by an SVM model using true labels can now simply be compared to the distribution given by (2.20) and statistical inference can be made. Thus, (2.20) ultimately gives us a fast and efficient analytical approach to interpret SVM models using p-values based on the theory of normal distributions. Next we extend the above theory to cases with unequal priors.

2.4.2. The case of unbalanced data

In the case of unbalanced data we let p denote the fraction of data with label $+1$ and hypothesize the distribution of y_i to be distributed as:

$$Pr(y_i = +1) = p \qquad Pr(y_i = -1) = 1 - p \tag{2.21}$$

This leads to an expected value and variance of y_i:

$$E(y_i) = 2p - 1 \qquad Var(y_i) = 4p - 4p^2 \tag{2.22}$$

We can use this to compute the expectation and variance of the components of $\mathbf{w}$ using (2.11) as:

$$E(w_j) = (2p-1)\sum_{i=1}^{m} C_{ij} \qquad Var(w_j) = (4p - 4p^2)\sum_{i=1}^{m} C_{ij}^2 \tag{2.23}$$

Note that we may write equation (2.11) as:

$$w_j = \sum_{i=1}^{m} C_{ij} y_i = \sum_{i=1}^{m} z_i^j \tag{2.24}$$

where we have defined a new random variable $z_i^j = C_{ij} y_i$ which is linearly dependent on y_i. Since the y_i are subject specific labels we expect them to be independent of each other and we can use the Lyapunov central limit theorem again to claim the asymptotic normality on

the distributions of w_j. However, application of this theorem requires:

$$\lim_{m\to\infty, d>>m} \frac{1}{\left[\sqrt{\sum_{i=1}^{m} Var(z_i^j)}\right]^{2+\delta}} \sum_{k=1}^{m} \mathrm{E}\left[\, |z_k^j - \mu_k^j|^{2+\delta} \,\right] \;\; for\ some\ \ \delta > 0 \tag{2.25}$$

where $\mu_k^j = (2p-1)C_{kj}$ To apply the Lyapunov CLT, we need the limit to vanish for some $\delta > 0$: We write down the limit for $\delta = 2$ here, as opposed to $\delta = 1$ in the last section. We do this because the limits are easier to write down and intuit with $\delta = 2$.

$$\lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} p|C_{kj} - (2p-1)C_{kj}|^4 + (1-p)|C_{kj} + (2p-1)C_{kj}|^4}{\left[\sqrt{(4p-4p^2)\sum_{i=1}^{m} C_{ij}^2}\right]^4} \tag{2.26}$$

The expectation is that this limit vanishes because the denominator contains cross terms not included in the numerator implying Gaussianity.

To see this first note that:

$$\lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} p(2-2p)^4|C_{kj}|^4 + (1-p)(2p)^4|C_{kj}|^4}{\left[\sqrt{(4p-4p^2)\sum_{i=1}^{m} C_{ij}^2}\right]^4} \tag{2.27}$$

may be broken down to

$$\lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} p(2-2p)^4|C_{kj}|^4}{\left[\sqrt{(4p-4p^2)\sum_{i=1}^{m} C_{ij}^2}\right]^4} + \lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} (1-p)(2p)^4|C_{kj}|^4}{\left[\sqrt{(4p-4p^2)\sum_{i=1}^{m} C_{ij}^2}\right]^4} \tag{2.28}$$

Since p, $1-p$ and $4p-4p^2$ are constants with respect to m, and the index i is interchangeable with k this boils down to the limit:

$$\lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} |C_{kj}|^4}{(\sum_{k=1}^{m} C_{kj}^2)^{4/2}} = \lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} (C_{kj}^2)^2}{\left(\sum_{k=1}^{m} C_{kj}^2\right)^2} = 0^* \tag{2.29}$$

For all experiments we performed using several different datasets this limit goes to zero. The

intuitive explanation for this is that The binomial expansion of the denominator contains cross terms in addition to the terms in the numerator (a total of m^2 terms), whereas the numerator contains only m terms. This intuition obviously relies on the assumption that the coefficients C_{kj}^2 are of comparable magnitude for different values of k and grow consistently with each other as m grows.

The question of what exact mathematical conditions are required for the limit in in (2.29) to go to zero belongs to the realm of real analysis. But one may gain some insight into the nature of this limit by further investigation of the matrix $\mathbf{C}$. Recall that we already know that:

$$\mathbf{C} = \mathbf{X^T}[(\mathbf{XX^T})^{-1} + (\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}] \tag{2.30}$$

Note that the Cayley Hamilton theorem gives us for $(\mathbf{XX^T})^{-1}$

$$(\mathbf{XX^T})^{-1} = \frac{1}{det(\mathbf{XX^T})}\sum_{s=0}^{m-1} c_s(\mathbf{XX^T})^s \tag{2.31}$$

where c_s are the appropriate constants.

Since the terms of $\mathbf{XX^T}$ are ultimately quadratic in the X_{uv} (the elements of the data matrix), the Cayley Hamilton theorem tells us that each term in the inverse can be expressed as a ratio of polynomials whose degree depends on m. This combined with the fact that $-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J}$ is essentially the sum of all terms in $(\mathbf{XX^T})^{-1}$ allows us to express the elements of $\mathbf{C}$ as:

$$C_{kj} = \frac{P_{kj}(X_{uv})}{Q(X_{uv})} \tag{2.32}$$

Here $P_{kj} \neq 0$ as long as the all entries in the jth column of $\mathbf{X}$ are not identical (that is we do not deal with the degenerate case where w_j will be 0). The limit, then boils down to:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m}(P_{kj}^2(X_{uv},m))^2}{\left(\sum_{k=1}^{m} P_{kj}^2(X_{uv},m)\right)^2} \tag{2.33}$$

Thus, given a specific value of m, the application of Cayley Hamilton theorem yields a

common denominator polynomial Q for all elements of $\mathbf{C}$. For a given m, the degrees of the polynomials P_{kj}, are also identical for all values of k and j. This can be verified with the use of the matlab symbolic math toolbox or Mathematica.

Since, the P_{kj} are polynomial functions of the elements of $\mathbf{X}$ of an identical degree, we expect them to grow identically with m. That is, $P_{kj} \in \Theta(g(m, X_{uv}))$. As long as a newly picked sample does not look too different from historical samples (that is assuming exchangeability in X_{uv}) we may safely assume $P_{kj} \in \Theta(g(m))$

Then, setting $a_k = P_{kj} > 0$ with $q = 2$, under the assumption that $a_k \in \Theta(g(m))$ for some function $g(m)$, we can show that:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} a_k^q}{(\sum_{k=1}^{m} a_k)^q} = 0 \tag{2.34}$$

To see this, note that the big-Θ notation implies that there exist constants M_1 and M_2 such that $M_1 g(m) \leq a_k \leq M_2 g(m)$.

Then for $q > 1$:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} a_k^q}{(\sum_{k=1}^{m} a_k)^q} \leq \lim_{m\to\infty} \frac{m M_2^q [g(m)]^q}{m^q M_1^q [g(m)]^q} = \lim_{m\to\infty} \frac{1}{m^{q-1}} \frac{M_2^q}{M_1^q} = 0 \tag{2.35}$$

Also:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} a_k^q}{(\sum_{k=1}^{m} a_k)^q} \geq \lim_{m\to\infty} \frac{m M_1^q [g(m)]^q}{m^q M_2^q [g(m)]^q} = \lim_{m\to\infty} \frac{1}{m^{q-1}} \frac{M_1^q}{M_2^q} = 0 \tag{2.36}$$

Then by the squeeze theorem on limits:

$$\lim_{m\to\infty} \frac{\sum_{k=1}^{m} a_k^q}{(\sum_{k=1}^{m} a_k)^q} = 0 \tag{2.37}$$

In summary, if one makes an assumption of exchangeability of the C_{kj} with respect to the sampling of $\mathbf{X}$, then $C_{kj}^2 \in \Theta(g(m))$ and we have

$$\lim_{m\to\infty, d>>m} \frac{\sum_{k=1}^{m} (C_{kj}^2)^2}{\left(\sum_{k=1}^{m} C_{kj}^2\right)^2} = 0 \tag{2.38}$$

As such this assumption allows for a broad range of values of C_{kj} and seems to be met in our experiments. A more formal treatment surrounding the behavior of this ratio is a topic of research in mathematical statistics, is beyond the scope of this work. We refer the interested reader to references (Fuchs et al., 2002; Ladoucette and Teugels, 2007; McLeish and O'Brien, 1982) where the behavior above ratio has been treated with far more rigor.

2.4.3. The case of the soft margin SVM

The above permutation testing approximation procedure has been developed for hard margin SVMs. Suppose we were to use soft margin classification instead, how would it change the approximation? First recall that typically for large values of the parameter 'C' in (2.2), the SVM penalizes errors in classification heavily. Hence, the slack $\xi_i = 0 \quad \forall i$. When this happens the solution to the soft margin and hard margin cases are the same. When perfect separability exists (such as the case of high dimensional low sample size data) the advantage of setting $\xi_i \neq 0$ can be realized only at extremely small values of 'C' where the optimizer typically forces $||\mathbf{w}||^2$, the first term in (2.2) to go to zero. When this happens the approximation described above will break down. However, it is important to note that typically the generalization performance of the classifier is also poor in when 'C' is very small (see figure 4). This has been previously noted for neuroimaging data in (Rasmussen et al., 2011). We found this to be true in our experiments as well (see figure 4).

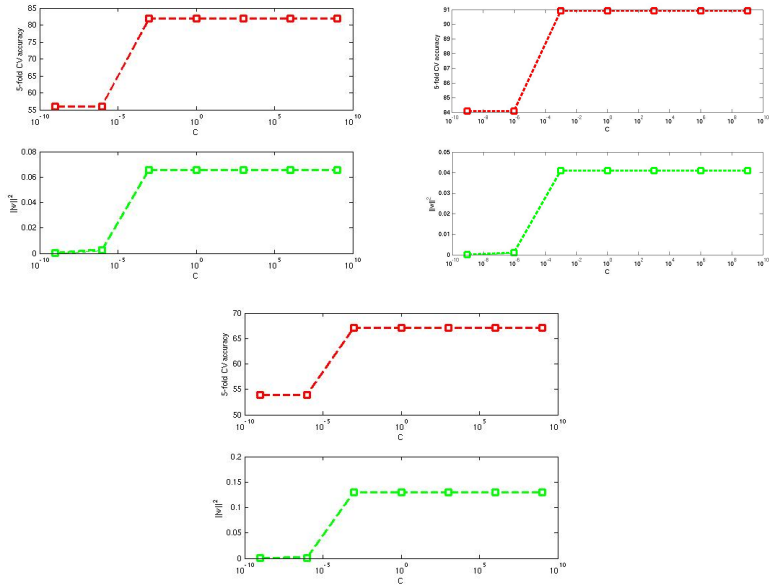


Figure 4: Top left: The effect of $10^{-10} < C < 10^{10}$ on the 5-fold cross validation accuracy(red) and the objective of classification $||\mathbf{w}||^2$ (green) for classification based on ADNI data. Top right: Effect for fMRI based lie detection Bottom: Effect for simulated data

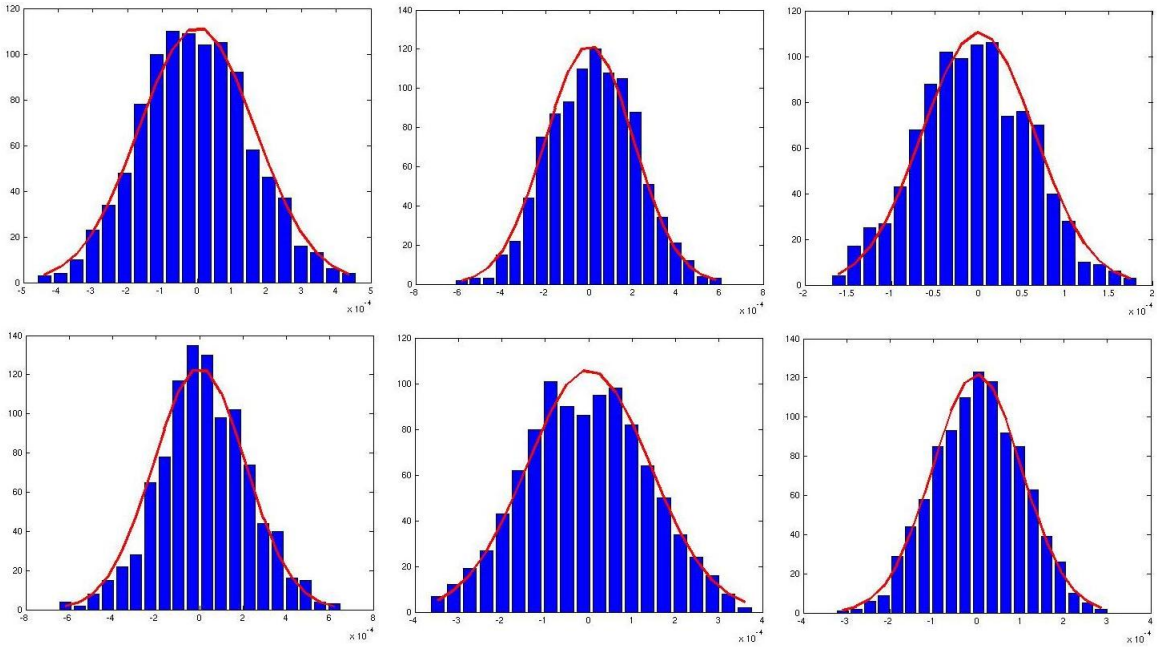


Figure 5: Comparison of weight vector component distributions and predicted normal distributions using permutation tests performed on real imaging data.

2.5. Generating p-value maps to interpret SVM models using the analytic framework

Let us denote the SVM model generated using the unpermuted labels by $\mathbf{w}^*$. We can compare $\mathbf{w}^*$ with null distributions obtained using analytical (or actual) permutation tests to get one p-value per voxel. These p-values can be collected into a p-value image. The mechanism of permutation testing implies that voxel locations where the p-values are small are also voxel locations where components of $\mathbf{w}^*$ differ significantly from the mean of the corresponding null distribution This implies that the model $\mathbf{w}^*$ significantly differs from a 'null model' at these locations. Thus, we expect these locations to be important in distinguishing controls from subjects. Our simulated experiments were in line with these expectations. Regions where differences between patients and controls were simulated turned up with lower p-values than the rest of the image. Thus, p-values generated by our analytic framework may be used for morphometry in addition to interpret SVM models.

2.6. Experiments and results

In this section we present three sets of experiments. The first set of experiments validates our statistical analysis framework with respect to simulated data. It also validates the analysis framework proposed above. The second experiment uses quasi-simulated imaging data. The third set of experiments demonstrates the application of this analysis framework in real imaging data.

2.6.1. Experiments on simulated data

Experiment comparing the proposed analytic framework with univariate analysis

An important feature of the proposed analysis is that it can detect multivariate patterns that univariate analysis will miss. This is despite the fact that we are performing hypothesis testing on individual hyperplane coefficients. We demonstrate this behavior with a simple experiment on simulated data. To simulate a multivariate effect we constructed labels and data that could only be separated using two variables combined. Thus, we simulated a bivariate pattern. Figure 6 (left) shows the simulated bivariate effect. These bivariate variables are represented by the red and green columns of the data, that are repeated column wise over and over again to simulate a differential effect between positively and negatively labeled samples. To generate these bivariate data we 1) sample 100 points (z_i) from a standard uniform distribution 2) sample points u_i from the standard normal distribution. 3) choose a factor $f < 0.1$ and generate point pairs $(z_i, z_i + fu_i)$. 4) generate labels using the criterion $label = sign(fu_i)$. A plot of a specific set of these pairs is shown in figure 6.

Using this process we generated a hundred relevant features. Further, we added 400 noise variables that had no relation with the labels to obtain the final dataset. Figure 6 (right) illustrates the scheme of simulation. Figure 7 (top-left) shows p-values obtained by running feature by feature univariate t-tests. Figure 7 (top-right) shows p-values obtained using our analysis framework. Figure 7 (bottom) shows SVM weights obtained by running a linear SVM using the data and the original labels. The figures show that the p-values generated by univariate tests are in the range of 0.25 to 0.65 for all simulated bivariate features. The p-values assigned by univariate testing to the remaining noise features also lie in the same range. As opposed to this, p-values generated by permutation testing are in the range 0 to 0.05 for relevant features. This range is much lower than the p-values the method assigns to the irrelevant features. The weight values associated by the linear SVM with the relevant features are not necessarily higher (or lower) than the weight values associated

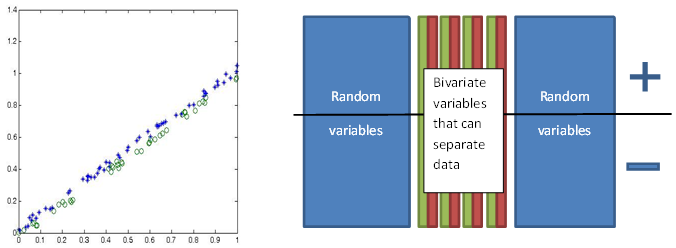


Figure 6: (Left) Bivariate pattern simulated using two features, (right) illustration of simulation procedure

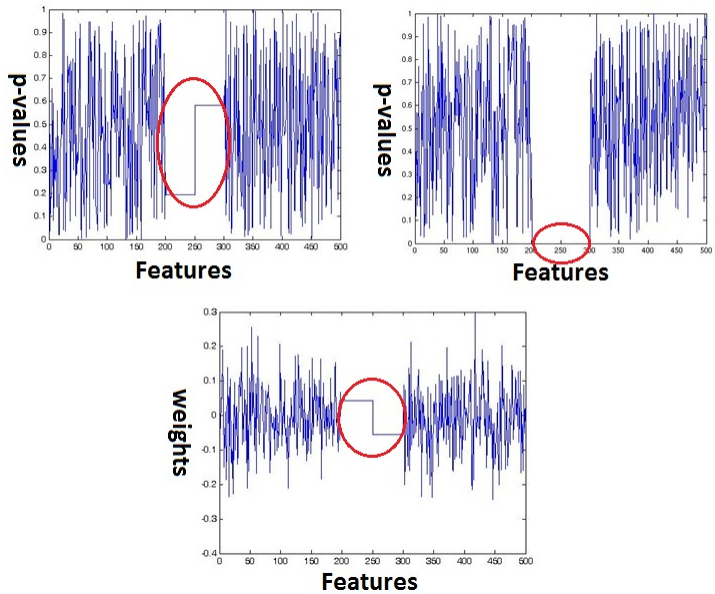


Figure 7: (Top-left) p-values from univariate testing on bivariate data (Top-right) p-values from proposed analytic permutation tests on bivariate data (Bottom) Weight values learnt by the SVM

with irrelevant features. Thus, the proposed permutation testing can detect multivariate patterns that univariate testing (or SVM weight vector thresholding) might completely miss.

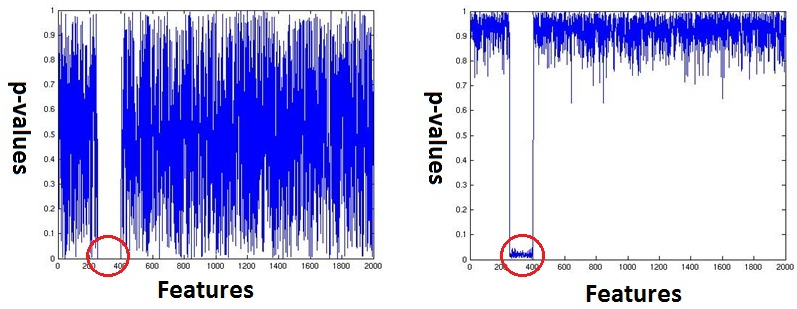


Figure 8: (Left) p-values generated using univariate tests which detect the effect, (right) p-values generated using SVM based permutation tests.

Experiment demonstrating the effectiveness of SVM based p-value maps in the presence of purely univariate effects

For this experiment, we constructed a simulated dataset as follows. We constructed labels and data that could be separated using only one variable. We repeated the univariate effect variables over and over to obtain sufficient dimensionality. This constitutes a multivariate pattern identifiable using univariate analysis. This pattern of relevant features spanned over 150 features. As before we added a large number of irrelevant noise variables that had no relationship with the labels. The simulated dataset contained a total of 100 feature vectors (50 labeled + 1 and 50 labeled 1) of dimensionality 2000. We introduced the simulated univariate effect in 151 features. We performed univariate t-tests feature by feature to obtain one p-value per feature. We then plotted these p-values in Figure 8 (left). We also performed SVM permutation tests using the procedure described in the paper and plotted the resulting p-values in Figure 8 (right). Figure 8 shows that univariate, as well as the proposed multivariate analyses, recover the features of interest.

2.6.2. Experiments on simulated imaging data

Simulated imaging data were generated by modifying a subset of grey matter RAVENS maps generated from raw T1-data of control images taken from the ADNI dataset. Specif-

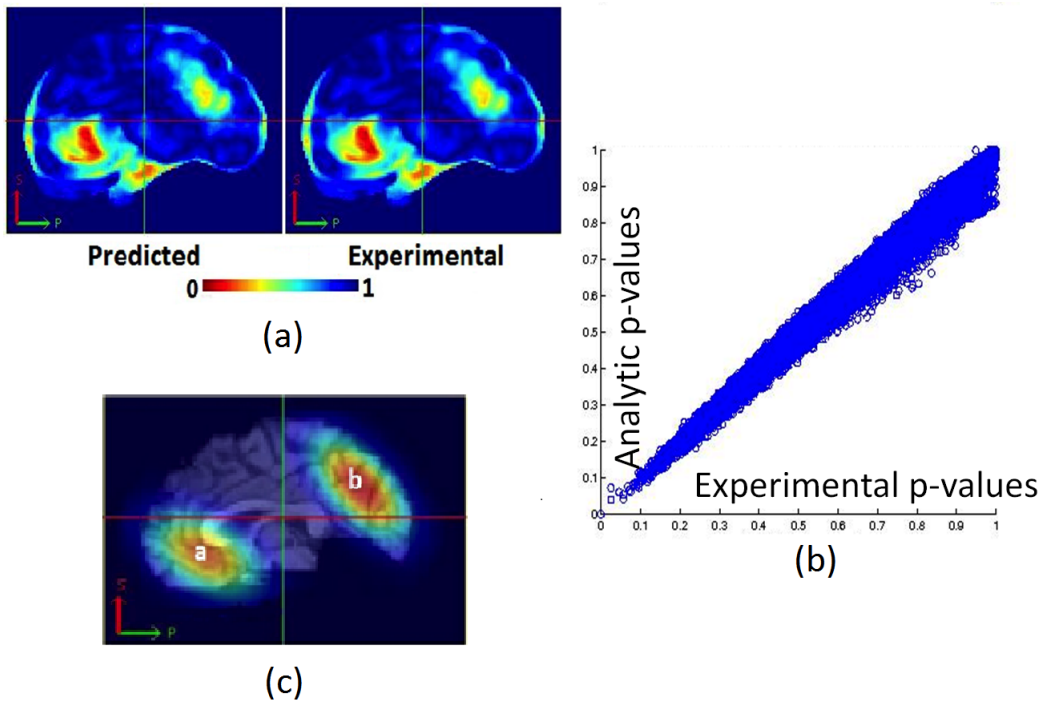


Figure 9: Results of experiments with simulated data. (a) A sagittal section through p-maps obtained from experimental and analytical permutation tests. (b) A scatter plot of p-values from experimental and analytical p-value maps. (c) Regions where simulated atrophy was introduced.

ically, we used 152 GM-RAVENS maps. We divided these RAVENS maps into two equal groups. In one of the two groups, (simulated patients) we reduced the intensity values of GM-TDMs over two large regions of the brain. We did this to simulate the effect of gray matter atrophy. We constructed these artificial regions of atrophy using 3D Gaussians. The maximal atrophy introduced at the center of each Gaussian was 33%. The reduction in the regions surrounding the center of this Gaussian was much lesser than 33%. We show the regions where we introduced artificial atrophy in Figure 9. We trained an SVM model to separate simulated patients from controls. We also performed permutation tests to obtain empirical approximations to null distributions of the weight vector components. We compared the components of the trained SVM models to the associated empirical null distributions for obtaining 'empirical p-maps. A similar comparison of SVM model com-

ponents with theoretically predicted null distributions yielded analytic p-maps. Figure 9 presents a 2D section of these p-maps as well as a scatter plot (using the full 3D image) of p-values obtained experimentally vs those obtained analytically. Figure 10 presents a visual comparison of the p-value maps in 3D by thresholding p-maps at several arbitrarily chosen thresholds. Figure 10 shows that analytically obtained p-maps are visually indistinguishable from experimentally obtained ones and thus validates our analytic framework.

2.6.3. Experiments with Alzheimer's disease data

In this section we discuss the application of our approach to interpreting SVM models trained in real data. A total of 278 GM, WM and ventricular tissue density maps were available for our experiment. The processed dataset contained images corresponding to 152 controls and 126 Alzheimer's patients. All three tissue density maps of a particular subject were downsampled and concatenated into a long vector and used as a feature vector for the analysis. Actual permutation tests were then performed to experimentally generate the null distributions and analytic approximations were also computed. We also trained an SVM model using the original labels and compared its components to the pre-computed experimental and analytic null distributions to obtain analytic and experimental p-value maps. Figure 11 presents these p-value maps as well as a scatter plot of p-values obtained experimentally vs those obtained analytically. Figure 12 presents a visual comparison of the p-maps in 3D by thresholding p-maps at several at thresholds of 0.01 and 0.05. Figure 12 shows that the SVM model finds information from the hippocampal regions to be most relevant to classification. Majority of literature implicates this region in the pathogenesis of Alzheimer's disease as well. Since the SVM model is based on this region we expect it to have a high generalization accuracy. Indeed, this turns out to be true. Generalization accuracy measured using the leave one out cross validation (LOOCV) accuracy for a linear SVM classifier trained on this dataset is 86%.

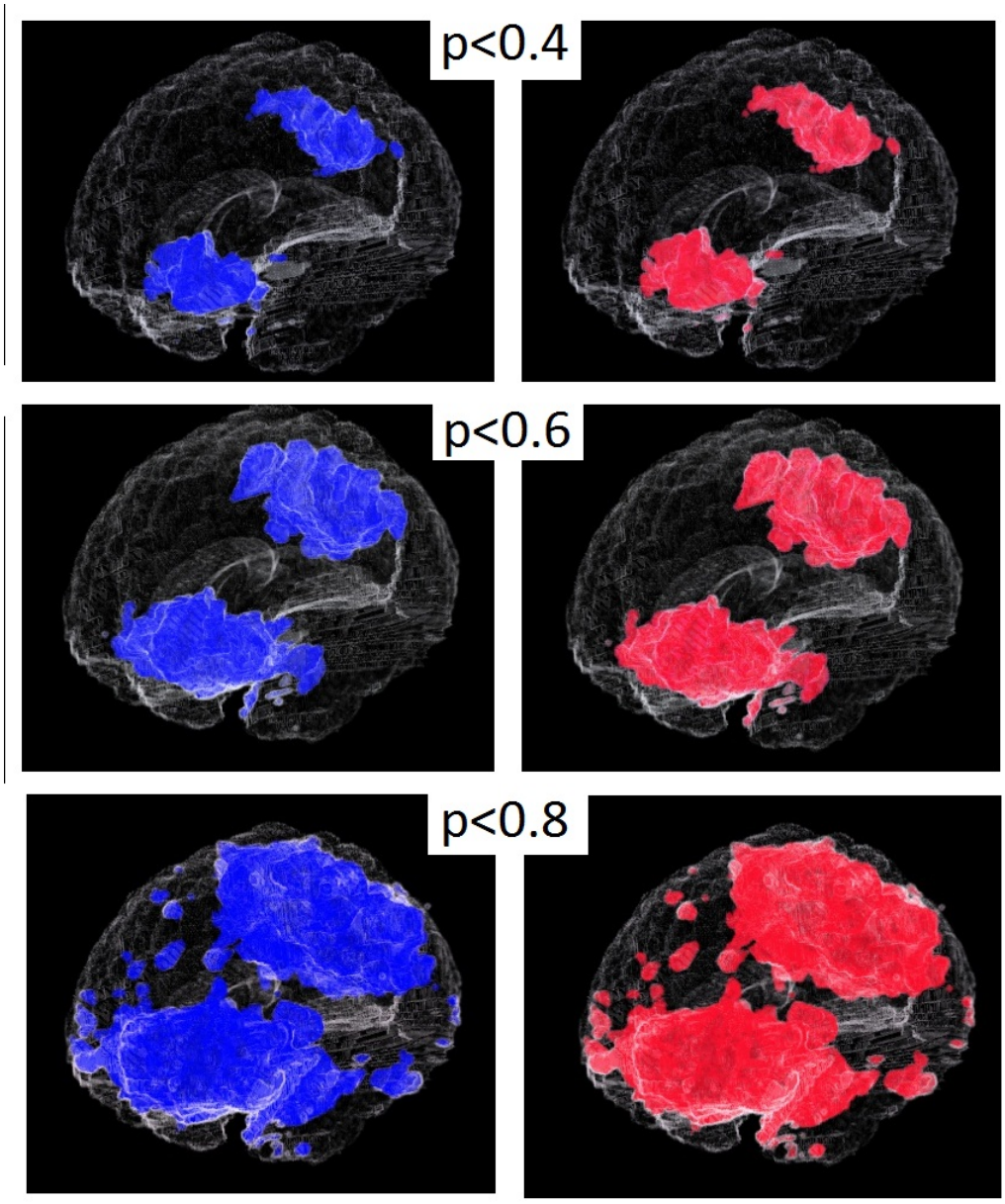


Figure 10: Simulated data: Experimental and analytical p-value maps thresholded at arbitrary p-values (3D).

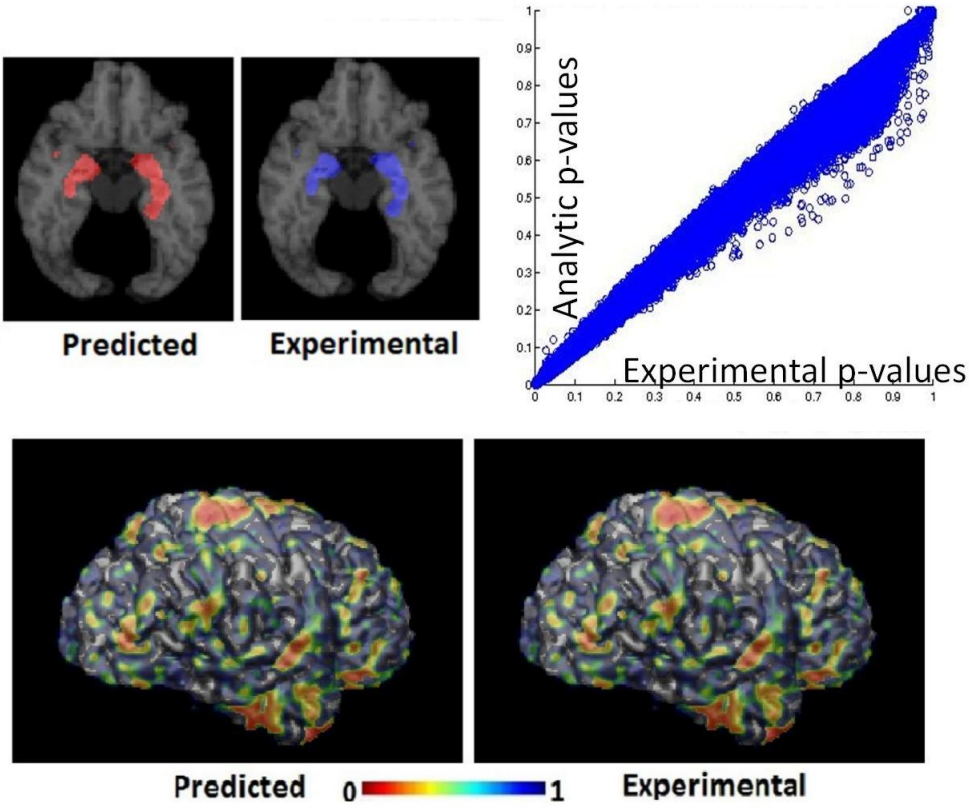


Figure 11: (Top left) Analytic and experimental p-value maps thresholded at 0.01 overlaid on the template brain (Top right) A scatter plot of p-values comparing experimental and analytical p-values.(Bottom) A 3D rendering representing predicted and experimental p-value maps (Right) A scatter plot of p-values

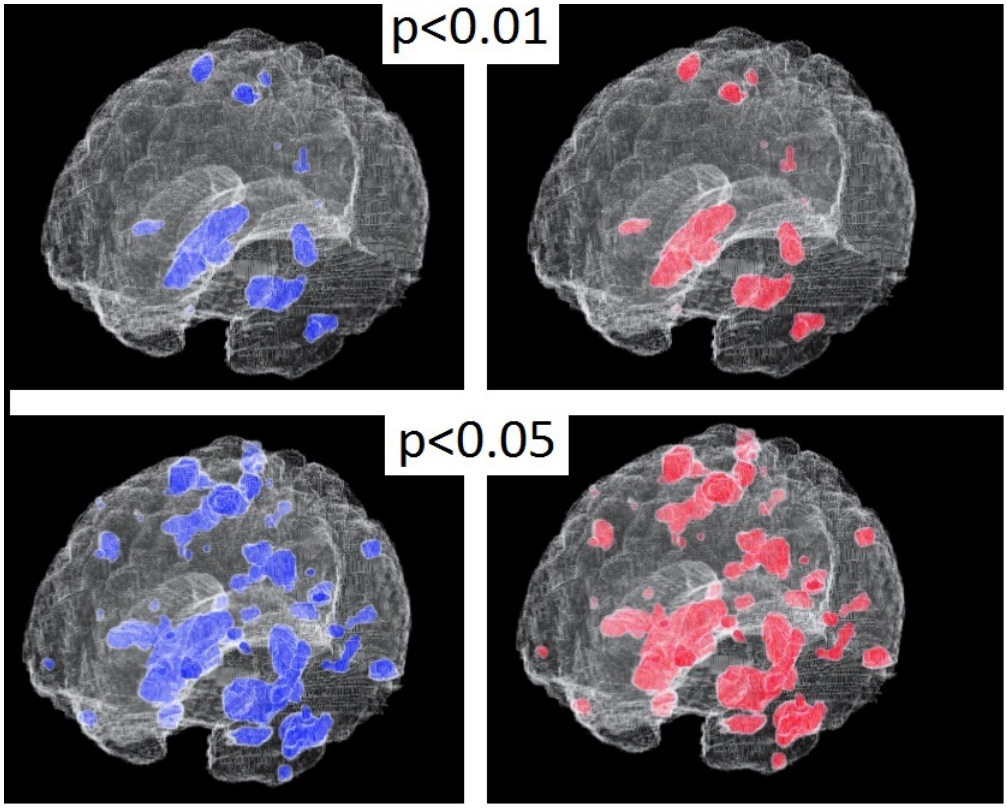


Figure 12: Experimental and analytical p-value maps thresholded at $\alpha = 0.01$ and $\alpha = 0.05$

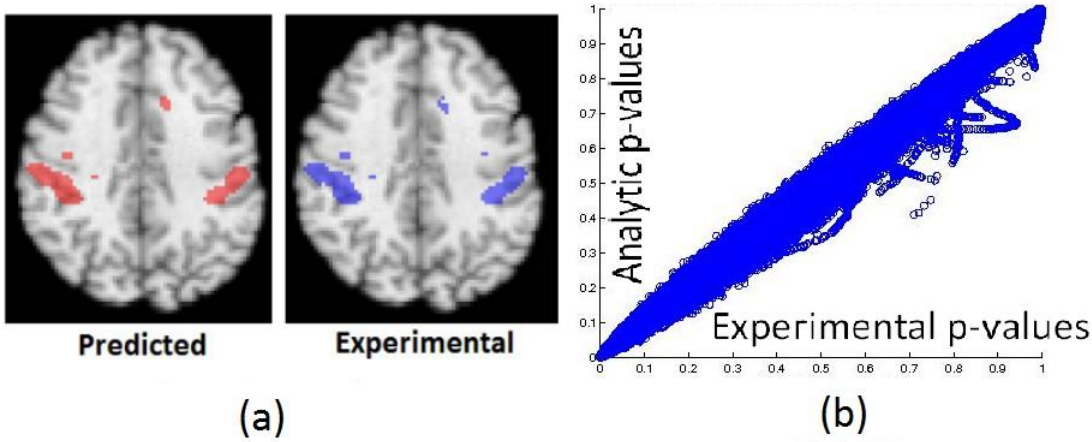


Figure 13: (Left) Experimental and analytical p-value maps thresholded at $\alpha = 0.05$ overlaid on a template brain (Right) scatterplot comparing analytically obtained and experimentally generated p-values

2.6.4. Experiments with functional data

Functional data were preprocessed to obtain parameter estimate images (PEIs) as described in (Davatzikos et al., 2005). A total of 44 PEIs, half of which consisted of lying responses and half of which consisted of truth-telling responses were used for the analysis. These data were obtained directly from the authors of (Davatzikos et al., 2005). Null distributions were obtained using analytic and experimental permutation testing as before. An SVM was trained using the actual labels as well. A section through the analytic and experimental p-value maps thresholded at 0.05 is presented in figure 13. The scatter plot of analytic vs experimental p-values generated using the entire 3D volume is also shown in figure 13. This plot shows that the approximation is less accurate here as compared to the simulated data or the Alzheimer's disease data. This is possibly due to the relatively smaller sample size. However, despite the relatively small sample size of 44 in this experiment it is still visually difficult to tell the difference between the regions obtained by thresholding theoretically predicted and experimentally obtained p-maps.

2.6.5. Experiments to study the accuracy of approximation with changes in sample size and dimensionality

An important question that is left unanswered by the qualitative analysis presented so far is that of how the performance of the approximation deteriorates. Specifically the effect of sample size and dimensionality on the approximation is not outlined by the experiments described above. Another interesting aspect is the study of how the number of permutations done in the experimental permutation tests affects the convergence of the approximation. In this section we present experiments to gain some insight into these questions. These experiments required the performance of empirical permutation tests with images of different dimensionalities and datasets of different sizes all of which had to be generated, stored and loaded from memory on a large parallel cluster. As such an enormous amount of computational time has gone into producing Figures 14, 15 and 16.

For all experiments presented here we have computed p-maps using the analytic approximation as well as empirical permutation testing. We use the average per voxel error between the two p-maps as a measure of deviation of the approximate from the empirical permutation testing result. Note that such a normalized measure of difference between images is especially useful while studying the effect of dimensionality on the convergence of the approximation. All three datasets described in the previous section have been used for experiments performed in this section. In case of the Alzheimer's disease dataset we randomly chose 100 patients and 100 subjects instead of using the entire data. This was done because it made it simpler to set up the experiment studying the effect of sample size. We describe each set of the experiments in more detail next.

Effect of number of permutations

We ran empirical permutation tests with 1500 permutations using all three datasets. We stored the models corresponding to each permutation to disk. To obtain the approximation accuracy for (randomly picked) one thousand permutations all we had to do was load 1000 results of the stored models, compute the empirical p-map and compare it with its analytic counterpart. We used this approach to generate figure 14. Figure 14 shows the average per voxel error in p-values obtained using actual permutation tests and the analytical approximation for all three datasets. Figure 14 indicates that the error reduces exponentially as the number of permutations increase.

Effect of reducing dimensionality

In this section, we address the impact of data dimensionality on the accuracy of the proposed approximation. To generate data of varying dimensionality we subsampled the imaging data at several different subsampling rates. Each subsampling rate yielded a new dataset whose dimensionality was much smaller than the original data. Then we ran empirical permutation tests (1000 random permutations) with SVMs on this subsampled data. We also computed the analytical approximation for each of the subsampled datasets. We plot the per voxel error rate between the analytic approximation and the experimental permutation testing in Figure 15. It can be seen that reduced dimensionality leads to a higher error rate. This indicates that the approximation works better when the data dimensionality is higher. We expected this intuitively, given that the approximation of an SVM by an LS-SVM is better when the dimensionality is higher. The experimental result simply confirms this intuition. From figure 15 one may speculate that increased dimensionality leads to an exponential decay in the approximation error.

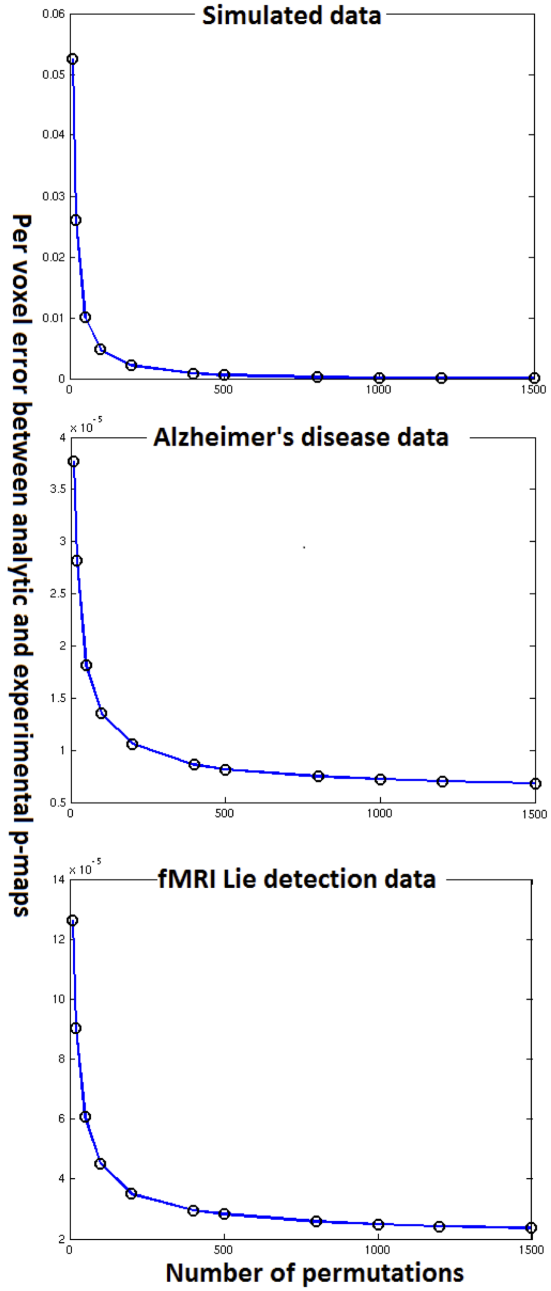


Figure 14: Approximation accuracy and number of permutations

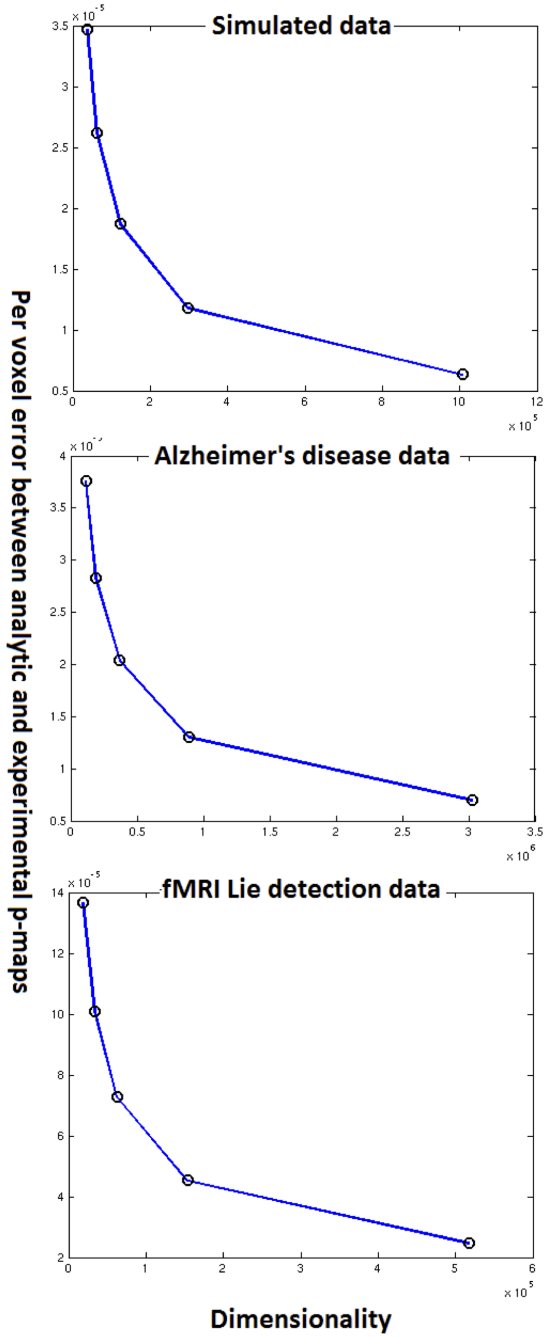


Figure 15: Effect of data dimensionality on approximation accuracy

LASSO			SVM permutations		Truth
Lambda	Number of true positives	Number of false positives	True positives at $p \leq 0.05$	False positives at $p \leq 0.05$	
1	0	0	151	0	151
0.9	2	0	151	0	151
0.5	17	0	151	0	151
0.05	69	0	151	0	151
0.004	99	0	151	0	151
0.002	95	9	151	0	151

Table 1: Comparison between LASSO and SVM permutation test based approach

Effect of reducing sample size

We have based the proposed analytic approximation on the central limit theorem. Hence we expect that an increased sample size would improve approximation accuracy. To understand the effect of sample size we consecutively halve the sample size and re-run both the empirical permutation tests (1000 random permutations) and the analytic approximations. For instance, if we had 100 patients and 100 controls, we ran experiments with the whole dataset, a dataset with 50 patients and 50 controls and 25 patients and 25 controls. In case of the Alzheimer's and fMRI datasets we added an extra point (75% of the full sample size) to better map the effect of sample size in this range. Figure 16 shows the variation of approximation accuracy with sample size for all three datasets that we ran experiments on. As expected, a larger sample size leads to higher accuracy. Note that even for sample sizes close to 20 the error in p-values is small (order of 10^5) for the fMRI data. In the Alzheimer's disease data where the dimensionality was substantially higher, the error is always in the order of 10^6. Just like dimensionality, increased sample size also seems to produce an exponential reduction in the error of approximation.

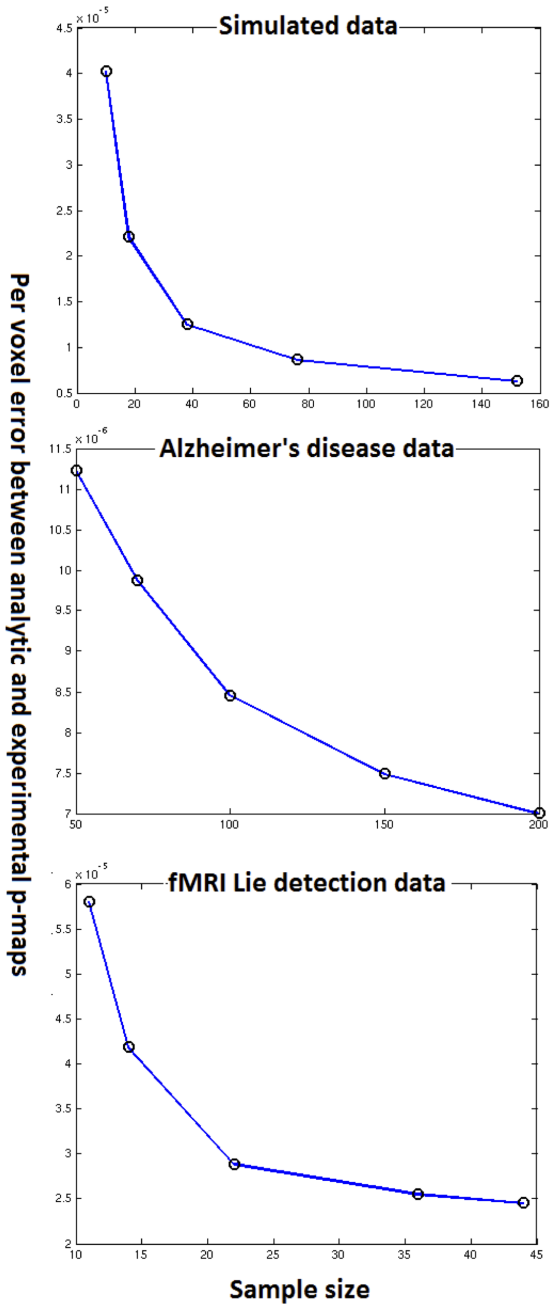


Figure 16: Effect of data dimensionality on approximation accuracy

Elastic net				SVM permutations		Truth
Lambda	Alpha	Number of true positives	Number of false positives	True positives at $p \leq 0.05$	False positives at $p \leq 0.05$	
1	0.25	138	0	151	0	151
1	0.75	15	0	151	0	151
0.1	0.25	138	0	151	0	151
0.1	0.75	82	0	151	0	151
0.01	0.25	106	7	151	0	151
0.01	0.75	90	0	151	0	151
0.001	0.25	72	81	151	0	151
0.001	0.75	93	53	151	0	151
1	0.05	151	0	151	0	151

Table 2: Comparison between elastic nets and SVM based permutation tests

2.6.6. Experiments comparing the proposed approach to sparse methods

A substantial body of literature has been developing around so called sparse methods for multivariate image analysis. For instance, methods described in (Ryali et al., 2012; Sabuncu and Van Leemput, 2011; Batmanghelich et al., 2012) attempt to apply sparsity to make interpretations/inferences. This is a growing body of literature and comparing the proposed method with every possible method out there is beyond the scope of this thesis. A large fraction of sparse methods use L-1 norm minimization. Hence we compared the proposed method with two methods that we believe to be representative of this literature. We compared SVM based permutation tests with the LASSO and the elastic net. We used the dataset used for generating figure 8. We ran LASSO for variable selection repeatedly with decreasing parameter values until the procedure started picking up false positives (started behaving like ordinary unregularized regression). We also ran elastic nets with several parameter settings, recorded the results and compared with the SVM permutations based

method. We tabulated results for the LASSO in table 1 and the results for the elastic net in table 2. From the tables, we can see that the LASSO can never find more features than the number of samples. This is a known limitation of the LASSO and one of the primary reasons for using Elastic nets. Selection of the minimum number of features using cross validation will yield an answer of 1. This alone is reason enough to avoid using the LASSO in neuroimaging analysis. In general we wish to use and visualize all regions used for prediction (and not eliminate them from the analysis). The SVM based permutation approach does not suffer from this limitation of the LASSO. The elastic net remedies the limitation of the LASSO and can find all the features introduced for certain parameter values. However, it still suffers from the parameter selection problem. In the simulated case, where relevant features are highly correlated cross validation based parameter selection would give the same accuracies whether we select one relevant feature or 151 relevant features. In such a case cross validation based parameter selection fails for the elastic net as well. Again SVM based permutation tests do not suffer from these limitations on account of their simplicity. While doing so they still retain the capability to find multivariate patterns and allow for a rigorous statistical p-value based interpretation.

2.6.7. When not to use the analytical approximation to permutation testing

One of the base assumptions in our approximation is that support vector machines treat all data as support vectors when attempting to learn from high dimension low sample size data. On the other hand the central limit theorem is an asymptotic result. This naturally leads to the question how much bigger than m does d have to be to safely apply the approximation. In order to understand this further we present here a plot of the ratio m/d to the ratio of $nSVs/m$ for simulated univariate data from the paper for various values of dimensionality d. It can be seen from the plot that for $m/d > 0.2$, less than 95% of samples remain

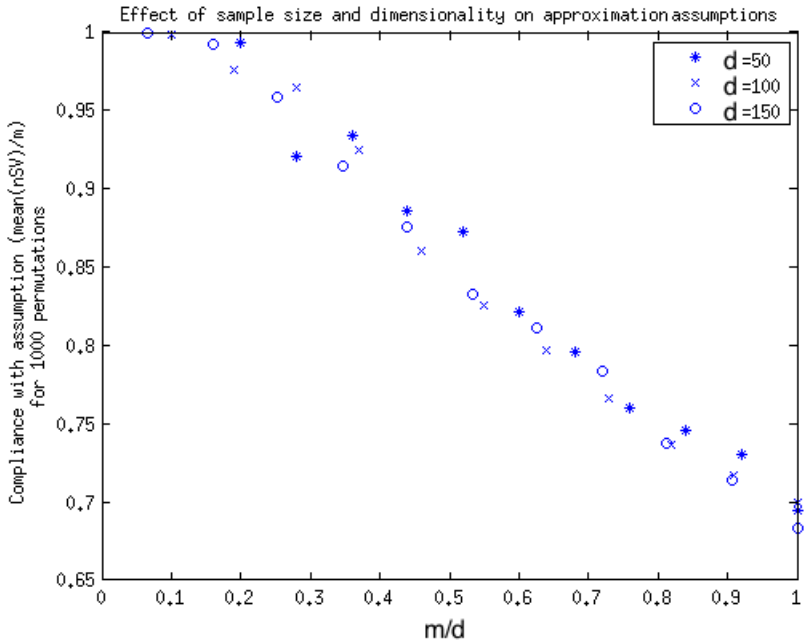


Figure 17: Plot of adherence to assumption that all samples are support vectors as m gets closer to d. Recall that m is the number of samples and d is the dimensionality of the data.

support vectors during permutation tests. This would constitute a substantial deviation from our assumption. Thus, it may not be wise to use the approximation in such a case. Fortunately, for image analysis, the number of voxels in an image (even a downsampled image) is in the range of millions while sample sizes barely touch a few hundreds. Thus, we expect $m/d << 0.2$ for neuroimaging studies for the most part. However, one must refrain from applying this approximation in case m/d gets too large.

2.7. Extending the proposed framework to Interpreting support vector machines (SVM) regression models

The theoretical and experimental developments presented so far concern support vector classification. Support vector regression analysis, on the other hand involves predicting a

continuous variable using imaging data. This variable can be a clinical score, age or even a protein or gene expression level. The support vector regression (SVR) algorithm may be used for addressing regression analysis in neuroimaging.

The SVR algorithm can predict continuous clinical variables from images. However, it provides no direct mechanism to assess what image regions were most significant in arriving at the predictions. This question is relevant in large clinical studies and is often asked by clinicians who lead such studies. Traditionally, regions associated with continuous clinical variables are found using mass univariate voxel based analysis (VBA). Such analysis associates a statistical significance test with each voxel in the image by regressing the voxel intensity directly with the target variable. Unlike MVPA, univariate analysis cannot predict target clinical variables and misses multivariate patterns in data. A multivariate alternative to VBA that is based on the interpreting a model learnt by an MVPA method such as an SVR is thus required and presented here.

Thus, in this section we describe a p-value based permutation testing based solution on the lines of what was presented in the previous section. We also extend the analytic approximation to cover the case of support vector regression. Analogous to classification we expect the proposed approach to highlight image regions used by the SVR to make predictions about the continuous output variable. We present the theory behind our approach along with the necessary background information on support vector regression next.

2.7.1. Support vector machines for regression

Regression analysis involves prediction of continuous clinical variables using medical images. The task at hand may be to predict clinical scores associated with disease stage or disease progression in patients. Regression models trained on normal data may be used to quantify mental age or infer gene/protein expression levels in at risk individuals. As in classification multivariate pattern analysis(MVPA) techniques such as support vector regression (SVR)

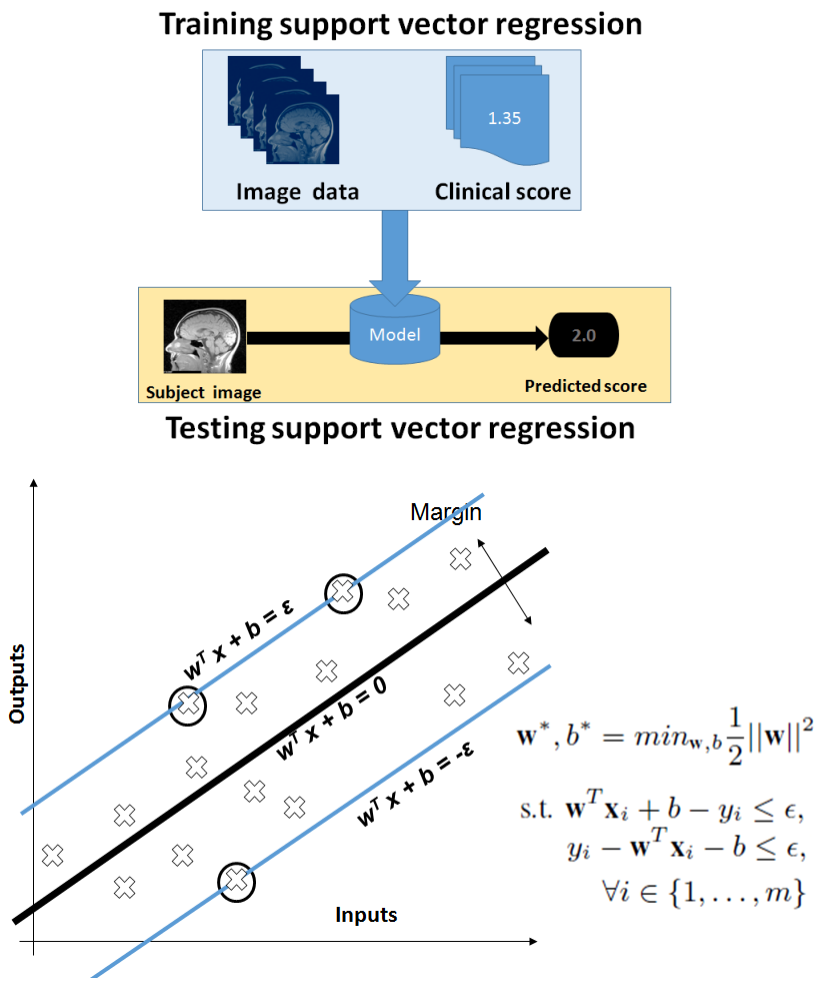


Figure 18: (Left) Support vector regression as applied in medical imaging. (Right)Concept of support vector regression in high dimensional space

directly address the image based regression paradigm. Most MVPA algorithms including SVR learn a model of disease by training on image data with known target variables. Target variable associated with a hitherto unseen test image can be computed using the 'learnt model' (see figure 18).

For training an SVR we stack preprocessed training image data into a large rectangular matrix $X \in R^{m \times d}$ whose rows $\mathbf{x}_i$ index individuals in the population, and columns index image voxels. A continuous target variable $y_i \in R$ is associated with every individual $\mathbf{x}_i$ for the training dataset. Further Note that the vectorized images $\mathbf{x}_i$ live in a Euclidean space

of dimension d. Then the $\epsilon-$SVR solves the following optimization problem:

$$\begin{aligned}
\mathbf{w}^*, b^* = min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\
subj.to.\ \ \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon + \xi_i \\
y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon + \xi_i^* \\
\xi_i, \xi_i^* \geq 0\ \ \forall i \in \{1, ..., m\}
\end{aligned} \tag{2.39}$$

The solution fits a tube of width ϵ -tube to the data (Schölkopf and Smola, 2002). When the number of samples is higher than the number of samples ($m > d$) then finding a tube of width ϵ that always contains all the data is not always feasible. The slack parameters ξ_i and ξ_i^* then allow a few datapoints to lie outside the ϵ -tube. In the medical image analysis setting we have $p > n$, that is the dimensionality is always much greater than the sample size. In this setting it is always possible to fit a d-hyperplane or a d-dimensional ϵ-tube through all the data points. Hence practically for all values of C, the solution to (2.39) is the same as the solution to:

$$\begin{aligned}
\mathbf{w}^*, b^* = min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\
subj.to.\ \ \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon \\
y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon \\
\forall i \in \{1, ..., m\}
\end{aligned} \tag{2.40}$$

Since the medical image analysis setting is almost exclusively high dimension low sample size we focus on the solutions to (2.40) instead of solutions to (2.39) throughout the rest of this article.

The SVR model is represented by the pair $\{\mathbf{w}^*, b^*\}$. For a new test subject whose vectorized image is represented by $\mathbf{x}_{test}$ the prediction y_{test} made by the SVR algorithm is $y_{test} = \mathbf{w}^{*T}\mathbf{x}_{test} + b^*$.

2.7.2. Permutation testing for support vector regression

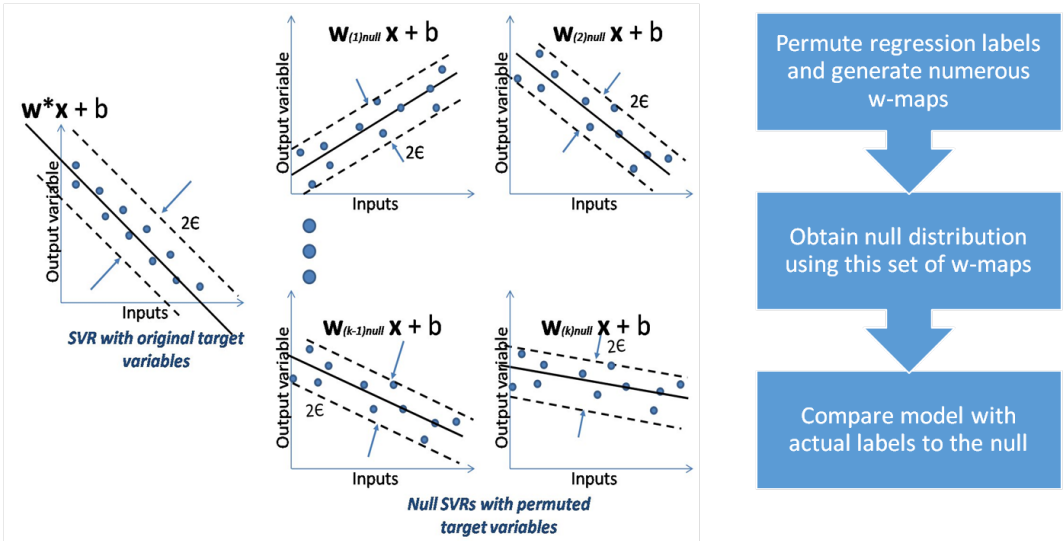


Figure 19: Concept of permutation testing in support vector regression. Comparison of $\mathbf{w}^*$ to the null distribution generated by $\{\mathbf{w}_{(1)null}, ..., \mathbf{w}_{(k)null}\}$ is used for inference

Note that the dimensionality of the model vector 'learnt' by the SVR $\mathbf{w}^*$ is the number of voxels in the image, that is $\mathbf{w}^* \in R^d$. Thus every component of the vector $\mathbf{w}^*$ can be mapped to a specific voxel in the image domain. This mapping associates an image with the SVR model. Henceforth, we call this image a $\mathbf{w}$-map. It is tempting to use to directly use this image for making inferences about which regions are most significantly involved in making predictions. However, these weights 1)can be biased to be large by the a simple scaling/translation operations on the underlying voxel intensities 2) provide no measure of statistical significance of a specific feature/voxel in the image. Thus a more appropriate method for interpreting the SVR model is needed. Permutation testing is one such method. The concept of permutation testing for SVRs in 2 dimensional space is illustrated by figure 19. In permutation testing, the target variables y_i are permuted randomly. For each random permutation an SVR is used to compute $\mathbf{w}^*_{randperm}$. After many thousands of permutations we can generate an approximation to the null distribution of every component of $w_j \rightarrow N^j_{null}$ where $j \in \{1, ..., d\}$. Finally, the original labels are used to train $\mathbf{w}^*$. Comparing the components w^*_j with N^j_{null} gives us a p-value associated with

every voxel. It is important to note that the null distribution at any voxel depends on the null distribution at all other voxels. This dependence is also true for the components of $\mathbf{w}$ themselves. Hence, each component wise test is based on data from all image voxels and is not univariate in the VBA sense.

2.7.3. The analytical approximation of permutation testing

In this section we present an analytical approximation to SVR based permutation testing. This approximation connects SVR permutation testing to standard statistical theory based on the normal distribution. Further it makes it possible to run multivariate analysis using computational resources comparable to what is required for VBA analysis. The fundamental assumption behind the analytical approximation is that in high dimension low sample size data for most random permutations majority of the samples lie on the support vector hyperplanes. This is analogous to the assumption made with respect to support vector classification. This assumption does not typically hold for the model trained with the actual targets since usually there is enough structure in the data to learn from. However, since most permutations are random the only way the algorithm can fit a model compatible with the entire dataset is by storing all of the data points and their labels as support vectors. Under this assumption, for most permutations the solution to (2.40) can be approximated by the solution to:

$$\begin{aligned} \mathbf{w}^*, b^* &= min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ subj.to.\ \ \mathbf{w}^T\mathbf{x}_i + b - y_i &= \epsilon \\ OR \\ y_i - \mathbf{w}^T\mathbf{x}_i - b &= \epsilon \\ \forall i \in \{1, ..., m\} \end{aligned} \tag{2.41}$$

Now note that one of the two constraints has to hold for every sample for every permutation. For a particular permutation a specific sample can either adhere to one constrain or another. Thus, for a particular permutation the optimization given by (2.41) can be solved using Lagrange multiplier theory as before to yield:

$$w_j = \sum_i C_{ij}(y_i + J_i\epsilon) \tag{2.42}$$

where $J_i = \pm 1$ depending on the constrain which holds for the corresponding sample and C_ij are elements the matrix $\mathbf{C} \in R^{d\times m}$ given by:

$$\mathbf{C} \doteq \mathbf{X^T}[(\mathbf{XX^T})^{-1} + (\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}] \tag{2.43}$$

Now over a large number of permutations we can expect either constraint in (2.41) to hold with a probability of 1/2 for each sample. Thus we may write the generic solution to (2.41) as:

$$\mathbf{w} = \mathbf{C}(\mathbf{y} + \mathbf{L}) \tag{2.44}$$

where the vector $\mathbf{L} \in R^m$ with components $L_i = \pm\epsilon$ and:

$$P(L_i = +\epsilon) = 1/2 \quad P(L_i = -\epsilon) = 1/2 \tag{2.45}$$

Note that (2.44) can also be written in its component form as:

$$w_j = \sum_{i=1}^{m} C_{ij}(y_i + L_i) \tag{2.46}$$

Since the C_{ij} are completely determined by the data matrix $\mathbf{X}$ we can treat these as constants and take expectations on both sides of (2.46) to obtain:

$$E(w_j)\sum_{i=1}^{m} C_{ij}E(y_i + L_i) = E(y_i)\sum_{i=1}^{m} C_{ij} \tag{2.47}$$

Note that $E(L_i) = 0$ and that $E(y_i)$ does not change with i allowing us to pull it outside

the summation sign. To explicitly acknowledge this invariance, henceforth we denote $E(y_i)$ simply as $E(y)$. Similarly the variance of w_j can be predicted by taking variances on both sides:

$$Var(w_j) = \sum_{i=1}^{m} C_{ij}^2 (Var(y_i) + Var(L_i)) = (Var(y_i) + \epsilon^2) \sum_{i=1}^{m} C_{ij}^2 \tag{2.48}$$

Note again that the term $Var(y_i) + \epsilon^2$ is invariant with respect to i. Henceforth we denote this term as $Var(y) + \epsilon^2$. Thus, we write:

$$E(w_j) = E(y) \sum_{i=1}^{m} C_{ij} \quad Var(w_j) = (Var(y) + \epsilon^2) \sum_{i=1}^{m} C_{ij}^2 \tag{2.49}$$

In regards to the distribution of w_j it can be shown to be normal using the Lyapunov central limit theorem as before. To see this define $z_i^j = C_{ij}(y_i + L_i)$ which is linearly dependent on $y_i + L_i$. We can infer the expectation and variance of z_i^j from y_j as:

$$E(z_i^j) = C_{ij} E(y) \quad Var(z_i^j) = C_{ij}^2 (Var(y) + \epsilon^2) \tag{2.50}$$

Note that z_i^j may be regarded as independent but not identically distributed and w_j are linear combinations of z_j^i. Thus, according to the Lyapunov central limit theorem(CLT) w_j is distributed normally if:

$$\lim_{m \to \infty} \frac{1}{\left[\sqrt{\sum_{i=1}^{m} Var(z_i^j)}\right]^{2+\delta}} \sum_{k=1}^{m} E\left[\, |z_k^j - \mu_k|^{2+\delta} \,\right] = 0 \;\; for\ some \;\; \delta > 0 \tag{2.51}$$

For $\delta = 1$.

$$E\left[\, |z_k^j - \mu_k|^{2+\delta} \,\right] = E\left[|C_{kj} y_k - C_{kj} E(y_k)|^{2+\delta} \right] = C_{kj}^3 E\left[|y_k - E(y_k)|^3 \right] \tag{2.52}$$

Again we note that $E\left[|y_k - E(y_k)|^3 \right]$ is independent of k and henceforth denote it simply

as $E\left[|y-E(y)|^3\right]$. Then, we can write the limit in (2.51) as:

$$\lim_{m\to\infty} \frac{E\left[|y-E(y)|^3\right]\sum_{k=1}^{m} C_{kj}^3}{\left[\sqrt{(Var(y)+\epsilon^2)\sum_{i=1}^{m} C_{ij}^2}\right]^3} = K\left(\sqrt{\lim_{m\to\infty}\frac{\sum_{k=1}^{m} C_{kj}^2}{\sum_{i=1}^{m} C_{ij}^2}}\right)^3 = 0 \tag{2.53}$$

where K is a constant independent of the sample indices k and i, defined by:

$$K = \frac{E\left[|y-E(y)|^3\right]}{\left[\sqrt{(Var(y)+\epsilon^2)}\right]^3} \tag{2.54}$$

Note that we have already investigated the limit in (2.53). It tends to zero under the assumptions presented before. Thus, in the limit we have normality of w_j by the Lyapunov central limit theorem.

2.7.4. Experiments and results

In this experiment, we applied the proposed method to the problem of white matter maturation in mouse brains. We used ex vivo acquired Diffusion Tensor images of a population of 79 inbred mice of C57BL/6J strain. The imaged mouse correspond to different postnatal stages, ranging from day 2 to day 80 (Verma et al., 2005). Early developmental stages were sampled more densely because development is more emphasized during that period.

The images were deformably registered to a template image chosen from the age group of day 10 using DROID (Ingalhalikar et al., 2010). DTI-Studio (Jiang et al., 2006) was used to estimate tensors from which, the Fractional Anisotropy was calculated resulting in images with dimension $300 \times 300 \times 200$.

As in the previous sections, we compare the experimental p-map with the analytic one. By visually comparing correspond slices from the two maps, we note that the predicted values closely follow the actual ones Fig. 20. We also observe distinctively low p-values in

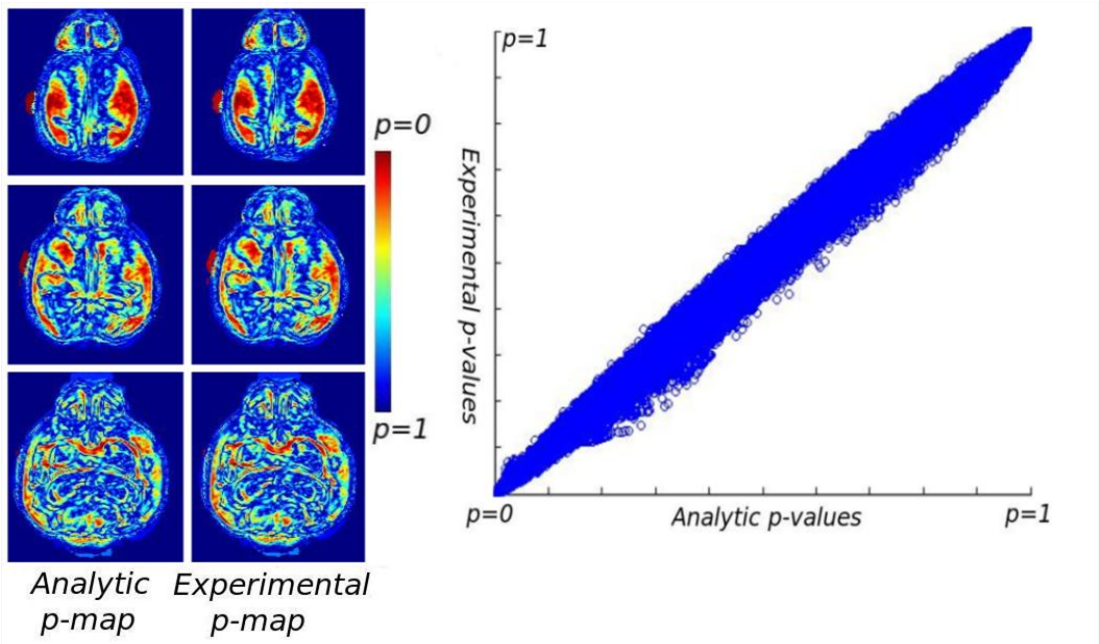


Figure 20: Representative slices of analytic and experimental p-maps (left) and scatter plot of corresponding analytic and experimental p-values for mouse brain data.

the cortex and the genu of corpus callossum. These areas have been previously reported exhibiting noteworthy maturation profiles (Verma et al., 2005). The scatter plot suggests that analytic and experimental p-values agree and highlight cortical connectivity changes in the developing mouse brain.

2.8. Conclusion

In this chapter we have described an analytic framework to interpret support vector models using statistical p-values. Our framework attaches a p-value based interpretation to diagnostic disease models learnt from imaging data using SVMs. We can use this machinery to quantitatively understand which regions/features contribute statistically significantly to the diagnosis made by a machine learning tool such as the SVM. Further, we can use it to compare two models generated using different data or in different populations in a

mathematically rigorous p-value based framework.

Our approximation to permutation testing makes multivariate analysis using SVMs possible using memory/time comparable to univariate VBM analysis. Thus, it provides a multivariate alternative to VBM type of univariate population based analyses. While we retain the advantage of multivariate analyses, our method scales well with increasing dimensionality and sample size. As imaging technology advances, we expect higher dimensionalities and higher sample sizes. Thus, the ability to perform multivariate population analyses inspite of ever increasing data sizes is of paramount significance. The methods we have presented above address this challenge appropriately.

2.8.1. A note on software

Throughout this work we use SVM implementation provided by the authors of LIBSVM (Chang and Lin, 2011). This is one of the most widely used and well tested libraries implementing SVMs in current practice.

CHAPTER 3

Improved interpretation of diagnostic SVM models: Enhancing inference using margin weighted statistics

3.1. Introduction

In this chapter we build upon the statistical analysis framework presented in the previous chapter. Consequently, the work presented in this chapter draws upon the work presented in the previous chapter quite heavily. In the previous chapter we developed a weight vector based framework for interpreting SVM models that applies in the high dimension low sample size settings found in neuroimaging. In this chapter we note that statistics based on $\mathbf{w}$, developed in the previous chapter, ignore a very important aspect of SVM theory, namely, the margin. This is an issue of critical importance since the SVM margin is closely related to the generalization error of the classifier and dictates the quality of the learnt model. Classification with a wider margin is inherently better than classification with a narrower one. This is in fact the basis of SVM theory (Vapnik and Vapnik, 1998). Using $\mathbf{w}$ alone, risks accepting the null hypothesis even when the margin associated with SVM classification is very high leading to very conservative inference. Thus, permutation testing must be done

with a statistic that is margin aware.

We introduce such a 'margin aware' statistic in this chapter. As expected, this statistic is closely tied with the $\mathbf{w}$ itself. On account of its relationship to $\mathbf{w}$ the null distribution associated with the proposed statistic is also Gaussian and in the high dimension low sample size case inference may be based on the normal distribution. Further, using simulated data we show that inference based on the margin aware statistic is superior in several aspects to analysis using the weights themselves. Wherever necessary, we repeat certain formulations and statements from the previous chapter for the sake of completeness and ease of reading.

3.2. The key challenge: SVM theory motivating the definition of margin based statistics

In this section we review SVM theory that drives the intuition behind the margin based statistic. While this is not meant to be a comprehensive review of SVM theory, we do present certain theoretical aspects that are relevant to the problem at hand.

First, the generalization performance of SVMs is usually measured in terms of the leave one out cross validation error (LOO error). Leave one out cross-validation involves leaving one sample in the training data set out and then training the algorithm on the rest of the data. Testing is done on the left-out subject. This process is repeated by leaving each possible sample out one at a time and estimating the average error over all the runs. Theoretical work presented by (Vapnik and Chapelle, 2000; Vapnik and Vapnik, 1998) presents several bounds on the expected leave one out error as a function of model and dataset characteristics. The simplest of these bounds is the radius margin bound. Specifically, if R is the radius of a

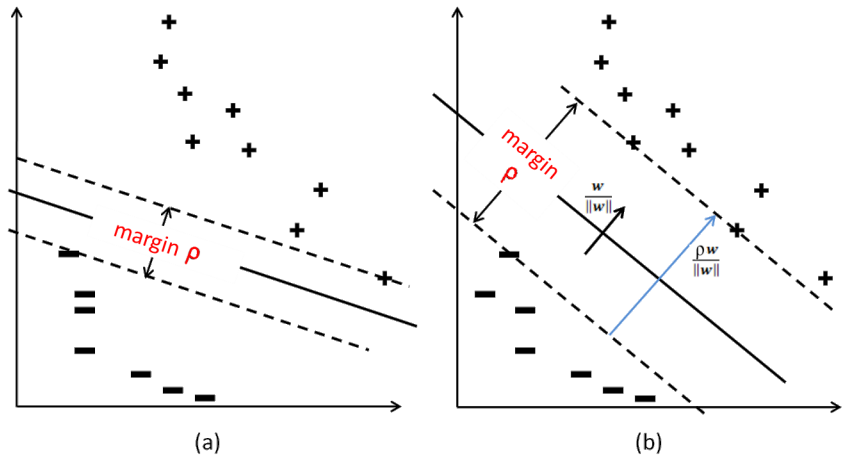


Figure 21: (a) Classification hyperplane with small margin (b) Classification hyperplane with larger margin preferred by SVM optimization. Also shown is the vector $\frac{\rho \mathbf{w}}{||\mathbf{w}||}$ which encodes margin information and is proportional the statistic used in this paper.

hypersphere containing all the data in high dimensional space, then:

$$E(LOOCVerror) \leq \frac{1}{m} E(4R^2 ||\mathbf{w}||^2) \tag{3.1}$$

Since the SVM margin is inversely proportional to $||\mathbf{w}||$, this is often known as the radius margin bound. Noting that the margin as measured between alternately labelled support vectors may be written as:

$$\rho = \frac{2}{||\mathbf{w}||} \tag{3.2}$$

the bound becomes:

$$E(LOOCVerror) \leq \frac{1}{m} E(\frac{R^2}{\rho^2}) \tag{3.3}$$

Thus, we expect models associated with higher margins to yield a lower generalization error in classification. For the purposes of inference using permutation testing this implies that we must give more consideration to weights of SVM models with larger margins as compared to weights of SVM models with smaller ones. This is precisely the intuition behind the margin based statistic.

3.3. Addressing the key challenge: The margin based statistic

Based on the intuition presented above we now define the margin based statistic. For each component w_j of $\mathbf{w}$ we define:

$$s_j = \frac{\rho}{2}\frac{w_j}{||\mathbf{w}||} \tag{3.4}$$

The statistic s_j represents the components of the vector $\frac{\rho \mathbf{w}}{||\mathbf{w}||^2}$ that is perpendicular to the separating hyperplane and has magnitude proportional to the margin associated with the classifier. This geometric interpretation is shown in two dimensions in figure 21. From figure 21 one can see that this new statistic incorporates not only the direction of SVM hyperplane but also the margin that it achieves. Using s_j instead of w_j for permutation testing incorporates the higher confidence associated with a higher margin directly into the statistical p-values generated by the permutation procedure. This is the primary theme of this chapter. We show using experiments that using s_j ultimately yields better interpretation as well. In the next section we extend the analytical framework of chapter 2 to permutation testing using s_j.

3.4. Analytic approximation for margin based permutation testing

In this section we show that inference using the statistic s_j can be done by comparing the value of this statistic at each voxel to a specific normal distribution. This is motivated by the fact that permutation testing with s_j yields a null distribution that is Gaussian in nature. We briefly review the relevant aspects and equations on the previous chapter and

proceed to develop the approximation based on these.

As in chapter 2 we consider imaging data are vectorized and $\mathbf{x}_i \in \mathbb{R}^d$ represents the ith image. Pathological (or functional) states are denoted by labels $y_i \in \{+1, -1\}$. and the SVM model is parameterized by $\mathbf{w} \in \mathbb{R}^d$, the solution to:

$$\begin{aligned} \{\mathbf{w}^*, b^*\} &= \min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ subj.to.\ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) &\geq 1 \ \ \forall i = 1, ..., m \end{aligned} \tag{3.5}$$

where m is the number of subjects in the training data. p-value based inference associated with this model may be achieved by comparing each component $\mathbf{w}_j^*$ to the null distribution given by:

$$\frac{w_j - \mu_j}{\sigma_j} \xrightarrow{D} \mathcal{N}(0, 1)$$

as $m \to \infty$ with

$$\mu_j = (2p - 1)\sum_{i=1}^{m} C_{ij} \quad \sigma_j^2 = (4p - 4p^2)\sum_{i=1}^{m} C_{ij}^2 \tag{3.6}$$

with $i \in \{1, ..., m\}$ indexing the samples, $j \in \{1, ..., d\}$ indexing voxels and p being the fraction of labels that are $+1$. C_{ij} are elements of the matrix $\mathbf{C}$ defined as:

$$\mathbf{C} = \mathbf{X^T}[(\mathbf{XX^T})^{-1} + (\mathbf{XX^T})^{-1}\mathbf{J}(-\mathbf{J^T}(\mathbf{XX^T})^{-1}\mathbf{J})^{-1}\mathbf{J^T}(\mathbf{XX^T})^{-1}] \tag{3.7}$$

with $\mathbf{J} \in \mathbb{R}^m$ being a vector with each component equal to 1 and $\mathbf{X} \in \mathbb{R}^{m \times d}$ is a matrix with $d >> m$ formed by stacking vector representations of imaging data as explained in the previous chapter. Our task in this section is then to derive a related distribution that allows for p-value based inference using s_j instead of w_j. To see this we begin by by noting:

$$\mathbf{CJ} = 0 \tag{3.8}$$

Thus:

$$\sum_{i=1}^{m} C_{ij} = 0 \implies \mu_j = 0 \implies E(w_j) = 0 \tag{3.9}$$

We then proceed using Taylor asymptotic approximations to estimate the mean and the variance of s_j (Casella and Berger, 2002):

$$E(s_j) = E\left(\frac{w_j}{\mathbf{w}^T\mathbf{w}}\right) \approx \frac{E(w_j)}{E(\mathbf{w}^T\mathbf{w})} = 0 \tag{3.10}$$

And similarly we can approximate the variance as:

$$\begin{aligned} var\,(s_j) \approx & \frac{var(w_j)}{E(\mathbf{w^T w})^2} + \frac{E(w_j)^2}{E(\mathbf{w^T w})^4} var(\mathbf{w^T w}) \\ & - 2\frac{E(w_j)}{E(\mathbf{w^T w})^3} cov(w_j, \mathbf{w^T w}) \\ = & \frac{var(w_j)}{E(\mathbf{w^T w})^2} \end{aligned} \tag{3.11}$$

We estimate $E(\mathbf{w}^T\mathbf{w})$ using the theory of quadratic forms (Searle, 2012):

$$E(\mathbf{w}^T\mathbf{w}) = tr(\mathbf{\Sigma_w}) + \mu^{\mathbf{T}}\mu = \sum_{k=1}^{d}({\sigma_k}^2 + \mu_k^2) = \sum_{k=1}^{d} {\sigma_k}^2 \tag{3.12}$$

Thus, we can write (3.11) as:

$$var\left(\frac{w_j}{\mathbf{w^T w}}\right) \approx \frac{var(w_j)}{E(\mathbf{w^T w})^2} = \frac{\sigma_j^2}{\left[\sum_{k=1}^{d} \sigma_k^2\right]^2} \tag{3.13}$$

Further, since s_j may be written as a continuous and smooth function of the components of $\mathbf{w}$ which are themselves normally distributed with a positive definite covariance matrix, we have that s_j is approximately normally distributed by the multivariate delta method (Casella and Berger, 2002):

$$\left(\frac{\sigma_j}{\left[\sum_{k=1}^{d}\sigma_k^2\right]}\right)^{-1} s_j \xrightarrow{D} \mathcal{N}(0,1) \tag{3.14}$$

An alternate explanation of normality may be uncovered by writing down the cumulant generating function (CGF) for the distribution of s_j. Recall that the CGF of a random variable t may be written as:

$$\begin{aligned} g(t) = log(E(e^{qt})) &= E(t)q + [E(t^2) - E(t)^2]\frac{q^2}{2!} \\ &\quad + [E(t^3) - 3E(t^2)E(t) + 2E(t)^3]\frac{q^3}{3!} + ... \end{aligned} \tag{3.15}$$

For the statistic s_i we note that the order of magnitude of the cumulants is in increasing powers of $\frac{w_i}{\mathbf{w^T w}}$:

$$\begin{aligned} E(s_j) &\sim (w_j/\mathbf{w^T w}) \\ E(s_j^2) - E(s_j)^2 &\sim (w_j/\mathbf{w^T w})^2 \\ E(s_j^3) - 3E(s_j^2)E(s_j) + 2E(s_j)^3 &\sim (w_j/\mathbf{w^T w})^3 \end{aligned} \tag{3.16}$$

Recall that we are working under the assumption that:w

$$(w_j/\mathbf{w^T w}) >> (w_j/\mathbf{w^T w})^2 >> (w_j/\mathbf{w^T w})^3 >> ... \tag{3.17}$$

Ignoring terms with a cubic or higher order in (3.15) the yields:

$$log(E(e^{q(\frac{w_j}{\mathbf{w^T w}})}) \approx E\left(\frac{w_j}{\mathbf{w^T w}}\right)q + \left[E\left(\left[\frac{w_j}{\mathbf{w^T w}}\right]^2\right) - \left[E\left(\frac{w_j}{\mathbf{w^T w}}\right)\right]^2\right]\frac{q^2}{2!} \tag{3.18}$$

Now using (3.13) we can write:

$$log(E(e^{q(\frac{w_j}{\mathbf{w}^{\mathbf{T}}\mathbf{w}})}) \approx (0)q + \frac{\sigma_j^2}{\left[\sum_{j=1}^{d}(\sigma_j^2 + \mu_j)\right]^2}\frac{q^2}{2!} \tag{3.19}$$

which is exactly the CGF of the Gaussian distribution with mean and variance given by (3.10) and (3.13) respectively.

In the next section we present experiments and results that validate the proposed statistic and the analytic inference framework associated with it.

3.5. Experiments and Results

3.5.1. Experiments on simulated data 1

In this section we present inference on simulated datasets using the proposed framework. The primary aim of this section is to demonstrate the validity of the proposed machinery and to show that permutation testing with s_j is at least as effective as permutation testing with w_j presented in the last chapter. Towards this end we demonstrate how inference using s_j successfully identifies regions driving group differences in simulated datasets similar to the ones used in the previous chapter. We show that the proposed statistic can indeed identify features/regions we would expect the SVM model to utilize for making a diagnosis. We simulate high dimension low sample size data that contain univariate and multivariate effects of interest which differentiate two subsets of the data. When single features (voxels in neuroimaging; genes and measures of their expression in genomics) can be used to detect differences between two groups, we say that a univariate effect is present at that feature. When one feature has to be used in conjunction with another (or many other) features to distinguish between two groups we say that multivariate effects are present. We simulate

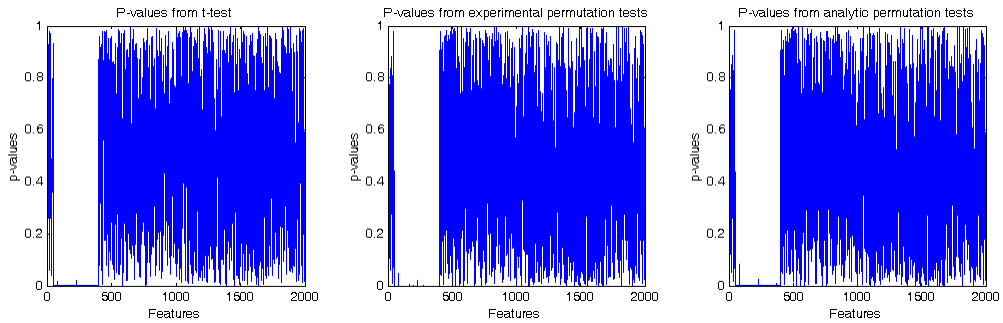


Figure 22: Inference for data where univariate effects may be used to distinguish labels (left) with p-values calculated by t-tests (middle) p-values calculated by permutation testing using the margin aware statistic and (right) p-values calculated by the analytical approximation to permutation testing using the margin aware statistic

data with both univariate and multivariate effects and show that our framework can be used to identify these effects. We also contrast our method with the widely used univariate analyses. Results presented in this section establish that inference using s_j is at least comparable to inference using w_j presented in the previous chapter. The next section demonstrates how it is actually better.

3.5.1.1. Detection of simulated univariate effect

The aim of simulation was to show that the proposed statistic can detect regions that differ between groups in a univariate sense. The data simulation scheme is similar to the one used in chapter 1 (in the section named 'Experiment showing the value of permutation testing analysis with respect to univariate analysis'). Briefly it involves 1) generation of random noise data $\mathbf{X} \in \mathbb{R}^{100\times 2000}$ by sampling a standard uniform distribution 2) random assignment of labels $+1$ and -1 to the 100 samples 3) subtracting a fixed value of 0.3 from 350 features in all samples labeled $+1$. The results are presented in figure 22.

The intention here was to simulate an effect which could easily be detected using a t-test. This can be seen from the low p-values assigned by the t-tests to the subtraction region in figure 22. The SVM based permutation test using the margin aware statistic can find this region as well. The analytic approximation to the margin based statistic is equally effective

in identifying the required region. Thus, the performance of our proposed statistic and the associated inference framework is at least comparable to the t-test when pure univariate effects differentiate between high dimensional data.

3.5.1.2. Detection of simulated multivariate effect

While univariate analysis using t-tests can detect effects simulated above, they cannot detect multivariate effects like the ones simulated in figure 6. The weight map based statistics w_j presented in the previous chapter are capable of detecting these effects. In this subsection we show that this capacity is inherited by s_j as well. Again we present a simulation on the lines of that shown in figure 6. Bivariate features are generated as before and represented in figure 23 where the green circles and blue crosses indicate distinct labels. The x and the y co-ordinates of each point represent values of features. Note that either x or y , used alone cannot differentiate blue from green. However, when used together one can easily draw a line that separates blue from green on the plot. As in chapter 2 we generate these bivariate features by 1) sampling 100 points (z_i) from a standard uniform distribution 2) sampling points u_i from the standard normal distribution. 3) choosing a factor $f < 0.1$ and generate point pairs $(z_i, z_i + fu_i)$. 4) generating labels using the criterion $label = sign(fu_i)$

However, this time we increase the total number of signal features and noise features. We use a total of 400 signal features (as opposed to 100 in the previous experiment) and 1600 noise features (as opposed to 400 in the previous experiment). The alignment of the signal and noise features is also illustrated in figure 23.

For inference purposes we run t-tests, actual permutation tests using the margin aware statistic and the analytic permutation tests proposed. The results are presented in figure 23. We also show SVM weights corresponding to the simulated features in figure 23 for comparison purposes. The figure shows that SVM weights or t-tests alone may not be sufficient to identify regions that the SVM model uses for classification. Using permutation testing to model the variance of the weights provides a certain edge over using the weights

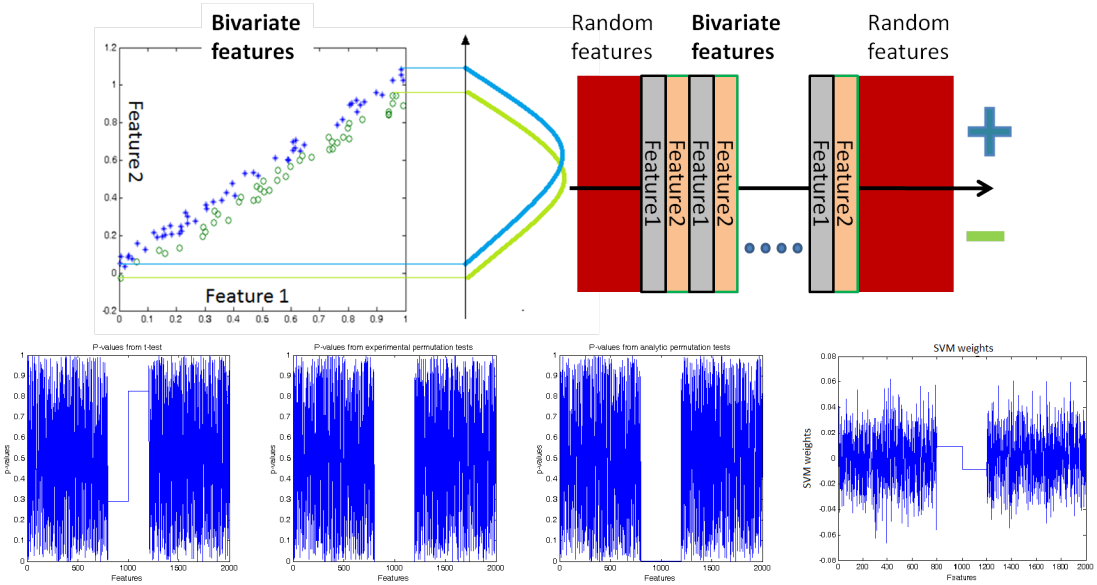


Figure 23: (Top-left)Features which can be used in a combined way but cannot be used individually to separate categories (Top-right) Illustration depicting simulation procedure for generation of multivariate toy data (Bottom) Inference on multivariate toy data p-values generated using standard t-tests (left) Inference using experimental permutation tests (middle-left) Inference using analytic permutation tests (middle-right) Inference using SVM weights (right)

directly or over univariate testing in this sense.

Both experiments presented above essentially focus on comparing permutation testing using s_j to permutation testing using w_j. We show that these two are comparable in both the univariate and the multivariate setting. Next, we present experiments contrasting the two. We build a case for why s_j should be used for permutation testing instead of w_j

3.5.2. Experiments on simulated data 2

The main problem with permutation testing using w_j is its sensitivity to abnormality size. SVM weights associated with each voxel/feature get smaller if the number of voxels/features driving the group difference increase. This raises the p-values associated with these features and reduces the sensitivity of the associated p-value map. In this sense using 'w' alone

seems to work more like the LASSO or the elastic net. We show here through simulated experiments that the margin based statistic alleviates this shortcoming.

3.5.2.1. Effect of simulated abnormality size on the analysis

In order to demonstrate the increased sensitivity of the weight based statistic to the size of the abnormality we create 3 separate datasets. Each of these datasets have their own simulated abnormality. The procedure used to create the data includes 1)generation of a random noise matrix of size $\mathbb{R}^{100\times1000}$2) 2) Randomly labeling 50 of these samples as $+1$ and the other 50 as -1. 3) Subtracting 0.3 from a pre-chosen subset of the 1000 features. Depending of the effect desired we chose either 50, 150 or 250 features from which the subtraction was made. For each of these datasets we show a plot of p-values generated using a) univariate testing b) the analytic approximation to the SVM weight vectors presented in chapter 1 c) The analytic testing framework proposed in this paper. These results are shown in the figure 24. From figure 24 we see that 1) using permutation tests based on SVM weights can detect small abnormalities 2) as the dimensionality increases, p-values based SVM weights can get as high as 0.8-0.9 which makes the abnormality undetectable at the standard threshold 0.05 3) The margin based p-value is relatively robust in this respect. In the following subsection we present similar results in neuroimaging data with simulated abnormalities.

3.5.2.2. Neuroimaging data with simulated abnormalities of different sizes

All the previously presented simulations use simulated effects as well as simulated noise. In order to bring the simulations closer to actual neuroimaging data we present a few experiments in this subsection using actual neuroimaging data. For this experiment we used grey matter tissue density (RAVENS) maps generated using ADNI data that was used in the prior experiments as well.

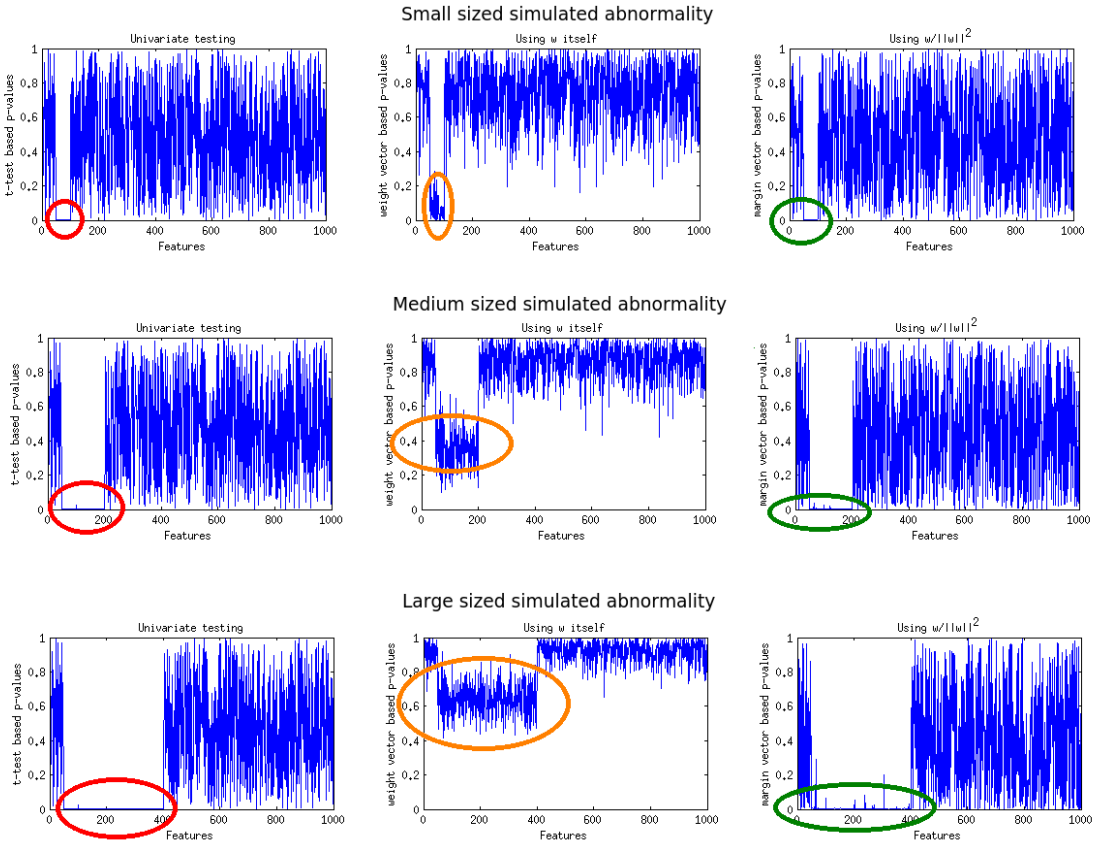


Figure 24: (Left column) p-values generated using t-tests with red circles indicating the location of the ground truth simulated effects (Middle column) p-values generated using SVM weights alone as described in chapter 1. Note that as the size of the simulated abnormality is increased the p-values increase to as high as p=0.8. Orange circles indicate the approximate location of ground truth (Right column) p-values generated using the margin based statistic. Green circles indicate approximate location of ground truth.

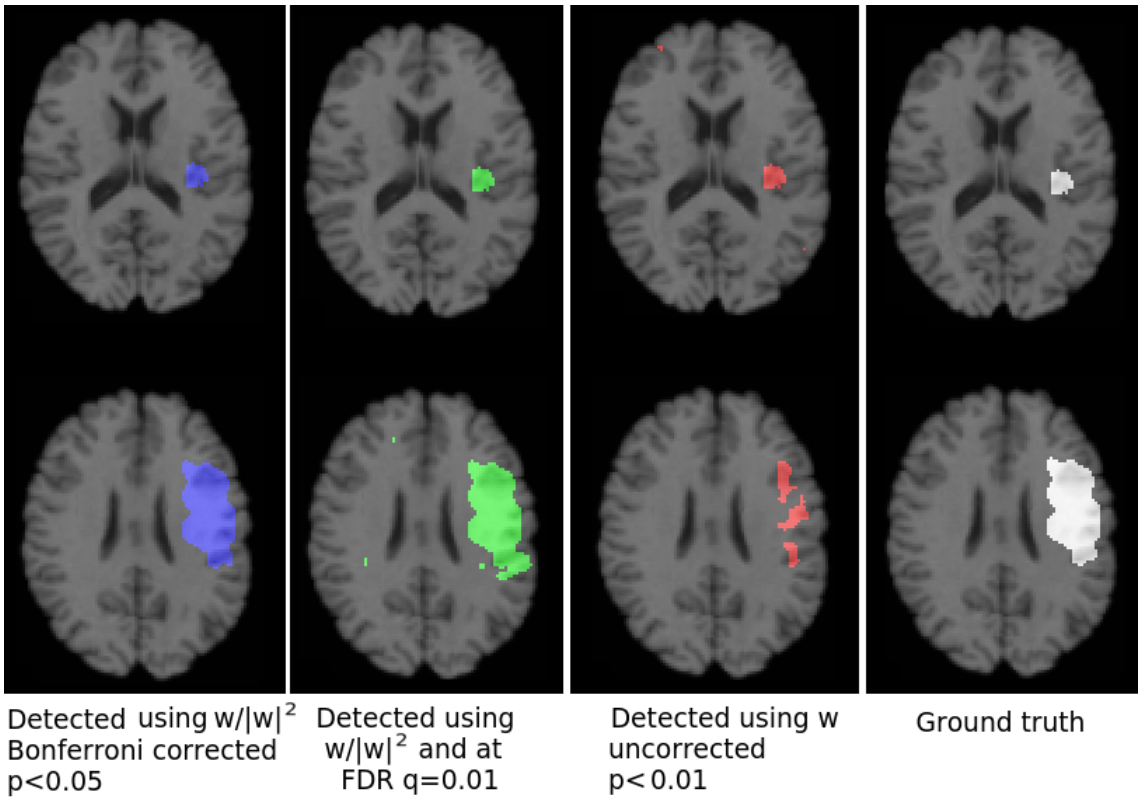


Figure 25: Detecting focal and non-focal simulated effects in neuroimaging data. First two columns from the left show detection using proposed statistic. Third column shows regions detected using permutations based on SVM weights only. The last column shows ground truth

We chose 152 grey matter RAVENS maps , from selected normal controls and introduced simulated an abnormality in exactly 76 of them. The region in which the abnormality was to be introduced was painted in using ITK-SNAP (Yushkevich et al., 2006). The abnormality is simulated by reducing the map intensity by 30%.

We used both SVM weight based permutation tests of chapter 2 and the margin based statistic described in this chapter to analyze the data. We also performed the above simulation with both, a small and a large simulated abnormality. The results are shown in figure 25. In general the following conclusions may be drawn from figure 25: First even in the presence of the noise profile associated with actual neuroimaging data, the proposed method of inference performs well. Second, this experiment re-iterates the finding that

inference based on permutation tests using SVM weights is highly sensitive to the size of the simulated abnormality. The margin based statistic does not seem to suffer from this problem.

3.5.3. Experiments with ADNI data: Qualitative analysis

The aim of the experiment presented here was to present a use case of our framework in real data. We also show qualitative comparisons between p-values produced by the analytical and experimental analysis frameworks. We use 100 controls and 100 patients from ADNI for this experiment. Grey matter, white matter and ventricular RAVENS maps were computed for each subject. All 3 RAVENS maps were downsampled and concatenated into a single long vector. Thus, we obtained one feature vector per subject. One thousand SVMs were trained using random permutations of the labellings (controls and patients) to obtain the null distributions of the proposed statistics for these subject specific feature vectors. These null distributions were used to compute experimental p-values. The analytical p-values were obtained using the distributions described by the framework presented in this chapter.

Figure 26 shows volumetric renderings of the negative logarithm of p-values corresponding to the grey matter tissue density maps overlaid on a brain volume. We include renderings for both analytically and experimentally obtained p-values overlaid on the T1-brain image. It can be seen from these images that the experimental and analytic p-value maps are at least visually indistinguishable. Based on our simulated experiments we interpret regions with lower p-values (higher -log(p-values)) to be more important to the classification function. Regions identified by thresholding the p-value map using the Benjamini-Yekutieli procedure at a q-value (Benjamini and Hochberg, 1995) of 0.1 are shown in figure 27. Regions discovered with q-value thresholds of 0.01,0.05 and 0.1 are shown in figure 28. Note that for FDR $\leq$0.1 the hippocampus, temporal cortex, the precuneus and the orbito-frontal cortex are all detected as regions relevant to SVM based classification. This is consistent with

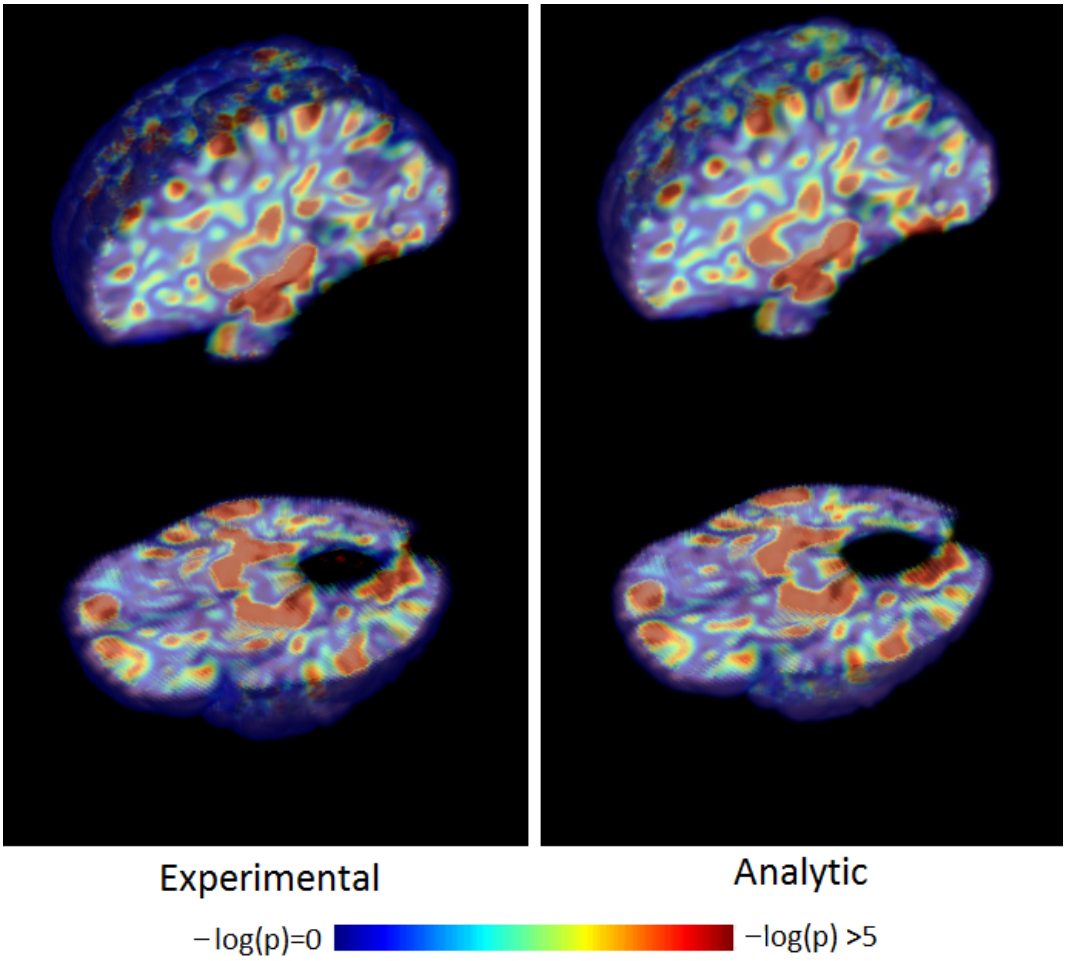


Figure 26: Visual comparison of experimentally (left) vs analytically (right) generated -log(p-value) maps using RAVENS maps data from ADNI

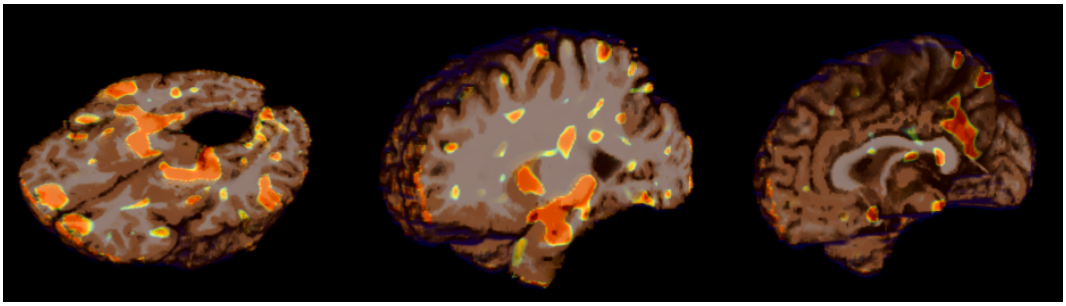


Figure 27: Regions detected after applying multiple comparisons corrections using the Benjamini-Yekutieli procedure at $q \leq 0.1$

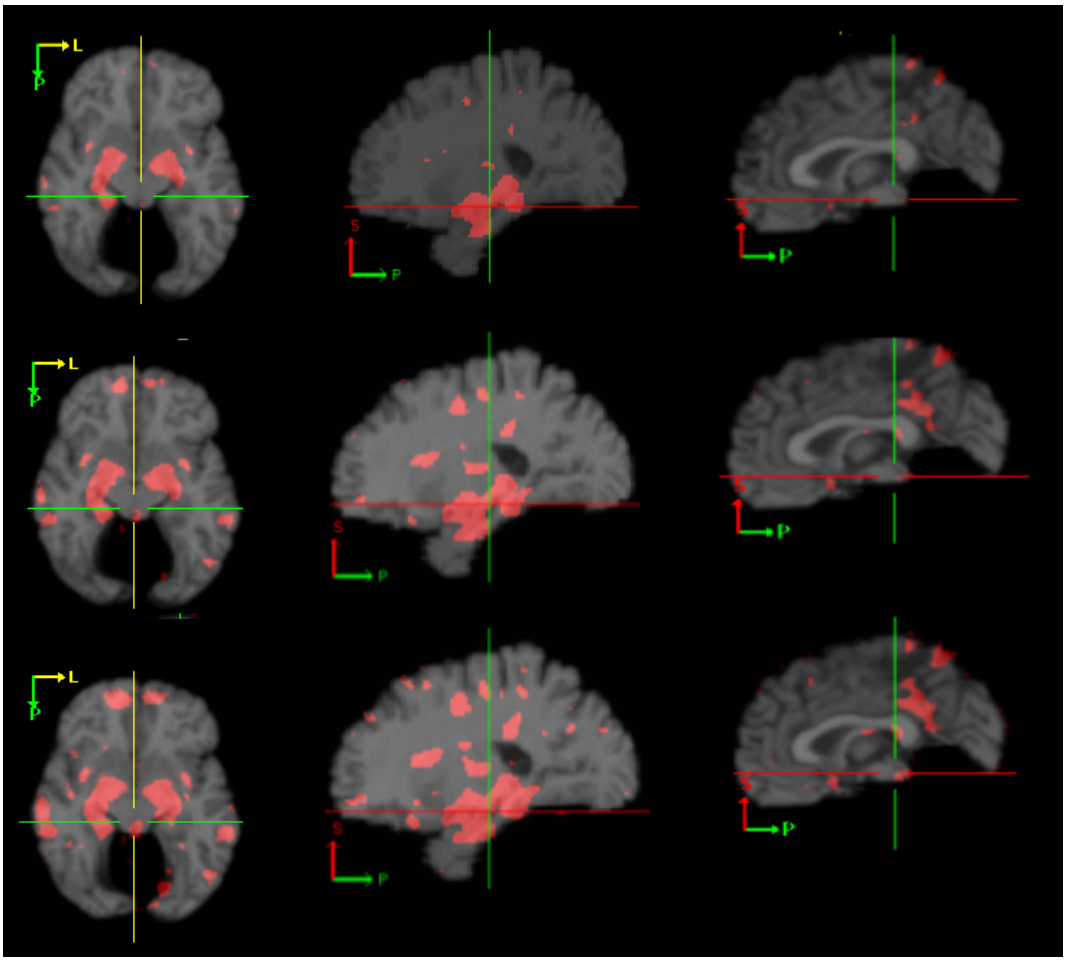


Figure 28: Regions detected after applying multiple comparisons corrections using the Benjamini-Yekutieli procedure at $q \leq 0.01$ (top row) ,$q \leq 0.05$ (middle row) and $q \leq 0.1$ (bottom row)

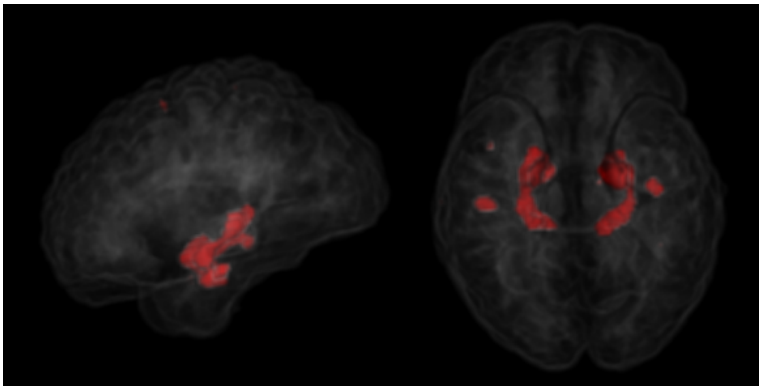


Figure 29: Regions detected after applying multiple comparisons corrections using the Bonferroni correction at $\alpha = 0.05$

neuroanatomical and neurofunctional literature relating to Alzheimer's disease pathology. At FDR$\leq$0.01 the procedure is extremely conservative but it still detects small sub regions of all of the above mentioned regions as significantly involved in the SVM prediction. Upon thresholding the p-map using the Bonferroni procedure at $\alpha = 0.05$ only the hippocampi are highlighted (See figure 29). Despite the usefulness of the two corrected p-value maps we surmise that the the negative log p-value maps of figure 26 provide for a better depiction of the relative importance of different brain regions in relation to the SVM model. While figures 26, 27 and 29 provide a visual description of the SVM p-value maps on ADNI data, for the sake of completeness we have included a more quantitative picture of the approximation in the scatter plot presented in figure 31.

3.5.4. Experiments with ADNI data: Comparison with local univariate analysis

As described in the introduction the primary focus of this work is to understand what regions of the brain are utilized by a support vector machine model to deliver diagnostic scores from imaging data. Thus, we are addressing the global multivariate paradigm and asking the question: What network of interacting regions does this SVM model use to make the diagnosis? This is slightly different from the local paradigm that is typically a subject of more traditional local univariate analyses which is : Which specific regions differ between two groups which are apriori known to be distinct?

The SVM models are thought to utilize global structural or functional imaging patterns to differentiate between groups. Thus, SVM based analyses may be suitable to identify a network of regions that acts synergistically to manifest a group difference. In contrast local univariate analyses may be more suitable to identify how a specific region differs between two groups. Thus, there is an inherent complementarity between these two types of analyses. We demonstrate this complementarity by comparing the result of a univariate analyses to

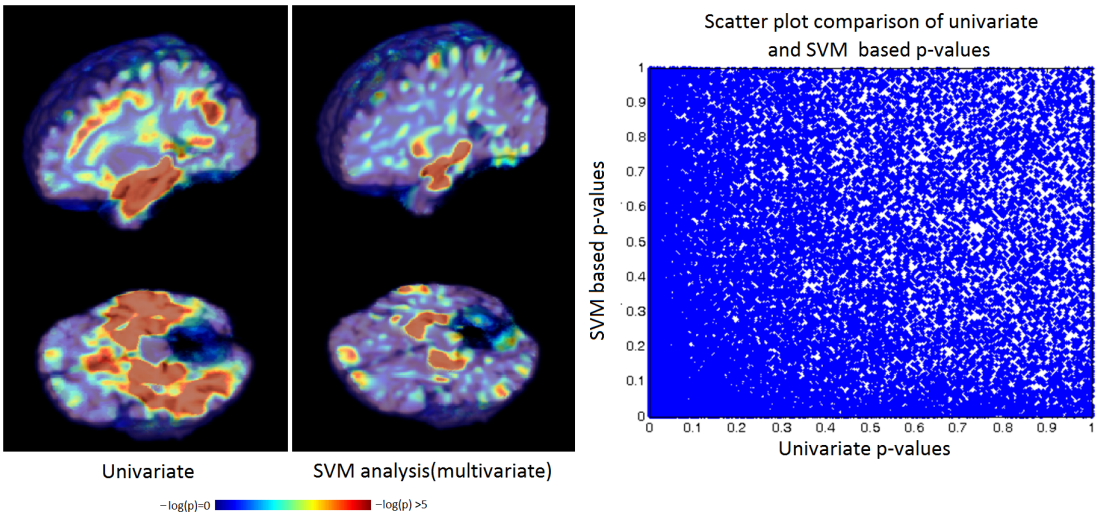


Figure 30: Complementary nature of SVM based analysis and univariate analysis (left) scatter plot of p-values obtained from univariate and multivariate analyses (right)

that of the proposed multivariate analyses on ADNI data in figure 30.

The univariate p-value map shown in figure 30 is generated by performing two sample t-tests between the grey matter tissue density values at individual voxels. The analytical approximation to permutation testing is used to generate the SVM based p-value map. A scatter plot comparing p-values obtained from the two analyses is also included in figure 30. The scatter plot shows that the two p-value maps are distinct. Some regions such as the hippocampi are significant according to both p-value maps. However, some other regions such as the orbito-frontal lobes are better highlighted in the SVM based map. Other regions in the temporal lobe are seen more clearly in univariate analysis. These results are not necessarily novel, but are simply presented here as a confirmation of the the view that local univariate analyses and global SVM based multivariate analyses offer complementary information for population based statistical analysis.

Further, it is also important to remember that the multivariate analysis presented here is focussed on interpreting the SVM model. Thus, we might say that the SVM uses a global pattern involving the hippocampus in combination with the highlighted regions of the orbito-frontal cortex to make its predictions. This, the SVM does despite the fact that

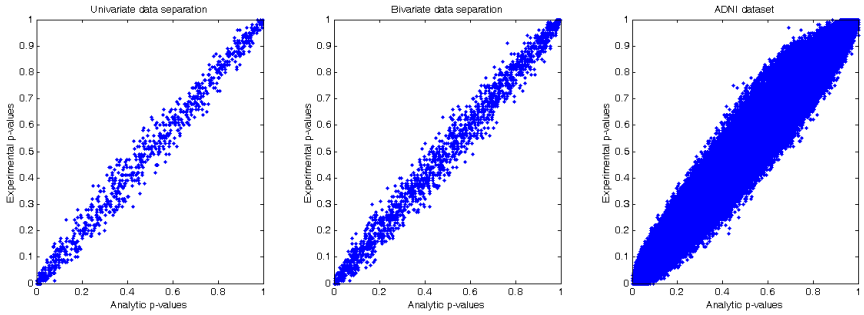


Figure 31: Scatter plots comparing actual permutation testing with analytical approximation

the univariate voxelwise differences between patients and controls in this region are not as strong as some other regions in the temporal lobe.

Thus, we may interpret, to successfully achieve better separation between controls and patients in a multivariate sense, the SVM model relied not only on the hippocampus and the temporal lobe but also on the orbito-frontal regions. Further, the SVM leave one out cross-validation accuracy of 87% gives us an idea of the predictive power of the highlighted multivariate pattern. As such there is no comparable measure to cross-validation accuracy in univariate analysis.

3.5.5. Quantitative analysis

We present scatter plots between experimental and analytical p-values in figure 31. The approximation accuracy seems to be higher at the low and high p-value ranges. These plots are based off the data used for generating figures 22, 23 and the ADNI data. The convergence between the analytic p-values and the experimental ones as measured by the average per voxel error is rapid. From figure 32 we can see that the average per voxel error in the p-values does not change substantially whether one uses 500 permutations or 1000 permutations. It does change substantially if one uses 100 permutations rather than 50. This was one of the factors behind choice of one thousand permutations for our experiments.

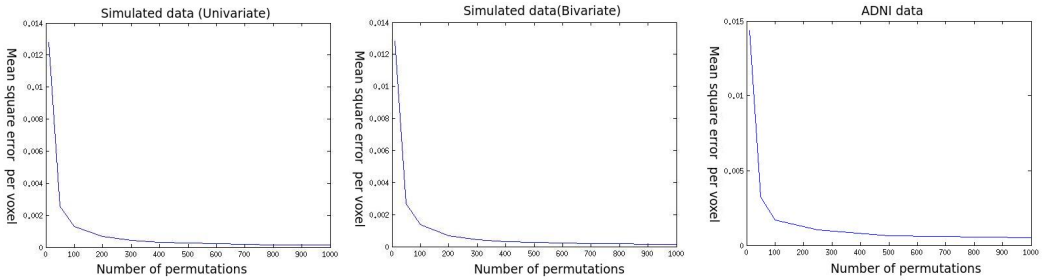


Figure 32: Convergence of the analytical approximation to experimental permutation tests

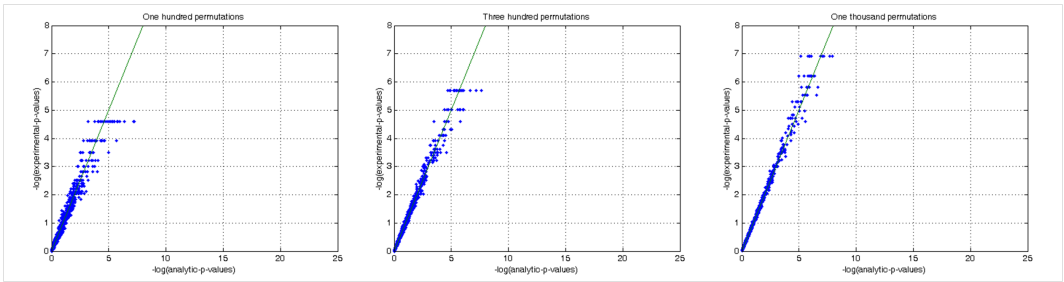


Figure 33: Variation of approximation error at low p-values with number of permutations

To investigate the behaviour of the approximation at ultra low p-values, we plot the negative logarithm of experimentally obtained p-values against their analytic counterparts in figure 33 using simulated data (that was also used for generating figure 22). At first glance it seems that the approximation is worse for low p-values and this seems like a limitation of the approximation itself. However, if one repeats the experiment with the use of successively larger number of permutations to obtain the experimental p-value maps a different picture emerges. The approximation error at low p-values is lower as the number of permutations used for generating experimental p-value maps is increased. Thus, the errors at low p-values are possibly a limitation of our inability to perform a large enough number of permutation tests as opposed to a limitation of the approximation itself.

To understand how the accuracy of this approximation decays with factors such as dimensionality, sample size and number of permutations we can use the fact that s_j can be expressed as a function of w_j and $\mathbf{w^T w}$.

$$s_j = f(w_j, \mathbf{w^T w}) = \frac{w_j}{\mathbf{w^T w}} \tag{3.20}$$

Then if Δw_j is the uncertainity in estimating w_j we can use error propagation theory to deduce:

$$(\Delta s_j)^2 \approx (\partial_{w_j} f)^2 (\Delta w_j)^2 + (\partial_{\mathbf{w^T w}} f)^2 (\Delta \mathbf{w^T w})^2 + \ cross \ \ terms \tag{3.21}$$

Ignoring, the higher order terms we get:

$$\left(\frac{\Delta s_j}{s_j}\right)^2 = \left[\left(\frac{\Delta w_j}{w_j}\right)^2 + \left(\frac{\Delta \mathbf{w^T w}}{\mathbf{w^T w}}\right)^2\right] > \left(\frac{\Delta w_j}{w_j}\right)^2 \tag{3.22}$$

Thus the relative error in approximating s_j is larger than that of approximating w_j alone. The above expression provides a relationship between the approximation error on s_j and w_j. The behavior of the approximation error on p-values generated using the w_j (as compared to actual permutations) has been documented in chapter 1. Any increase in error in approximating components of $\mathbf{w}$ will produce a monotonic increase in the approximation error of s_j. Thus, factors such as a lower dimensionality and lower sample size which can increase the approximation error of w_j, automatically lead to a corresponding increase in approximation error for s_j. Also, based on (3.22) the increase in relative error for s_j will be larger than the corresponding increase in relative error for w_j. Consequently, we would expect the error in approximation of the p-value maps to be larger for maps based on s_j as compared to those based on w_j.

3.6. Applications

In this section we present p-value maps generated by applying our SVM based inference framework to several real imaging datasets.

3.7. Application to major depressive disorder data

The dataset used for this experiment consisted of 23 patients with Major Depressive Disorder and 20 healthy controls matched for age, gender and IQ. Patients with MDD assessed with the Structured Clinical Interview for DSM-IV Axis I disorders (SCID-IV First et al. (2012)). met criteria for single or repeated episode MDD repeated episodes MDD without psychotic features as defined by Diagnostic Statistical Manual of Mental Disorders, Fourth edition, text revision (Association et al., 2000). Healthy controls were screened to ensure that they did not meet criteria for any mental disorders based on SCID-IV. MRI scans were acquired on a 3.0 T GE SIGNA HDx (Milwaukee, USA) at Kings College London. Image preprocessing was done to obtain grey matter and white matter RAVENS maps in a manner similar to the ADNI dataset. An SVM classification model trained using white matter RAVENS map data yielded a leave one out cross validation accuracy of 67.44% and an AUC of 0.71. Inference using p-value maps computed using the theory presented in this chapter are shown in figures 34 and 35. The regions used by the classifier have been previously implicated in literature relating to depression.

3.8. Application to schizophrenia data

We applied our method to analyze an SVM model trained on data from a study comparing controls to schizophrenia patients. The subjects of this study were recruited at the Department of Psychiatry and Psychotherapy at Ludwig-Maximilians University, Munich, Germany, and included 163 patients with an established DSM-IV diagnosis of schizophrenia and 163 matched normal controls (NC).

All participants provided their written informed consent prior to MRI and clinical examina-

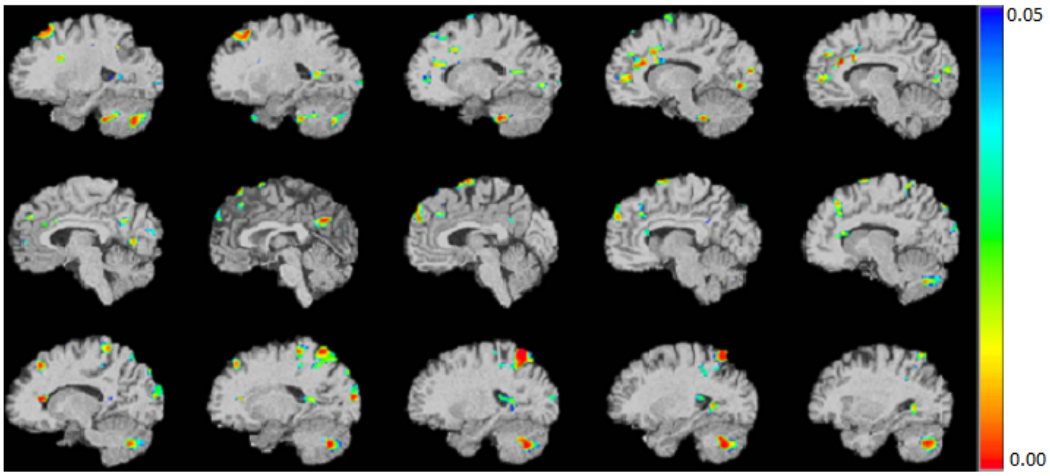


Figure 34: p map showing white matter regions after thresholding to $p < 0.05$ The white matter regions that showed highest contribution towards group difference were bilateral cerebellar and occipital regions, left parietal and right frontal lobes

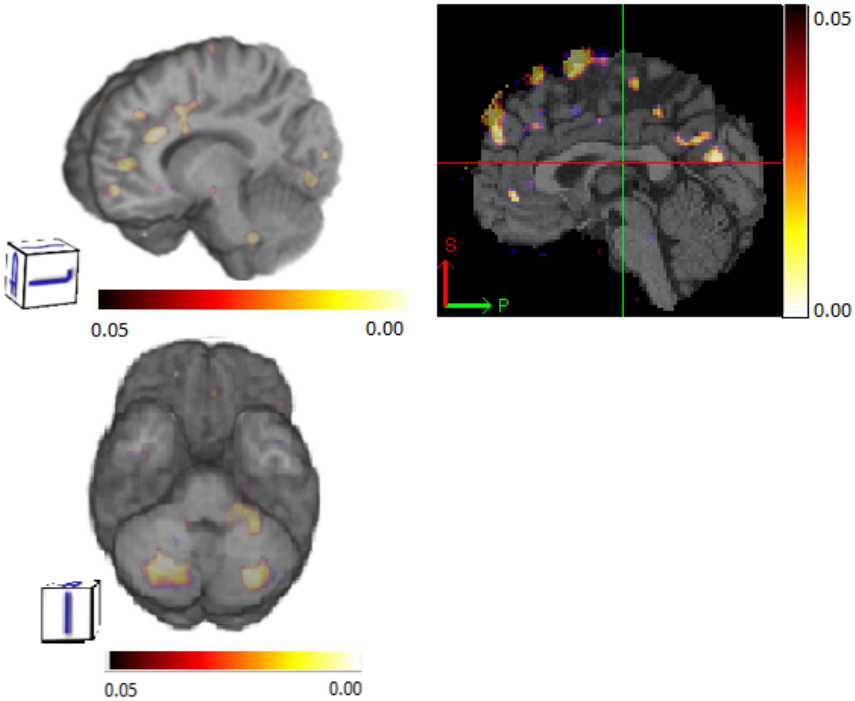


Figure 35: Bilateral cerebellar and occipital and right frontal regions (all $p < 0.05$)

tion. Patient recruitment was performed by trained clinical investigators and consisted of a structured clinical interview for DSM-IV-axis I disorders (SCID-I), a standardized clinical interview for the assessment of medical and psychiatric history. All subjects were diagnosed based on a consensus between 2 experienced psychiatrists who used the DSM-IV criteria and the SCID-I. Participants were excluded if they had other psychiatric and/or neurological diseases, past or present regular alcohol abuse, and/or consumption of illicit drugs, as well as past head trauma with loss of consciousness or electroconvulsive treatment.

T1-weighted 3D-magnetization-prepared rapid acquisition with gradient echo sequences (repetition time, 11.6ms; echo time, 4.9ms; field of view, 230mm; matrix, 512x512x126, contiguous axial slices of 1.5mm thickness; voxel size, 0.45x0.45x1.5mm) were acquired on a 1.5 T Magnetom Vision scanner (Siemens, Erlangen, Germany). The images were first preprocessed by means of the VBM8 toolbox (publiclly available at http://dbm.neuro.uni-jena.de/vbm8/)an extension of the Statistical Parametric Mapping software (SPM, publically available at http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) for skull stripping, bias correction, and segmentation (Ashburner and Friston, 2005). The skull-removed partial volume estimation images were then spatially registered to the respective partial volume image of the single-subject Monteal Neurological Institute (MNI) template through a robust method for elastic registration called deformable registration via attribute matching and mutual-saliency weighting (Ou et al., 2011). The deformation field resulting from this spatial registration was then applied to the segmented images in order to generate RAVENS maps of the gray matter (GM), white matter, and cerebrospinal fluid segments. In these RAVENS maps, the tissue density reflects the amount of tissue present in each subjects image at a given location after mapping to the standardized template space.

Subsequently an SVM model was trained on the grey matter RAVENS maps (5-fold cross validation accuracy 69.63%). The p-value map associated with this model has extremely low values bilaterally in regions of the orbito-frontal cortices and the cerebellum. Sections through this map are shown in figure 36.

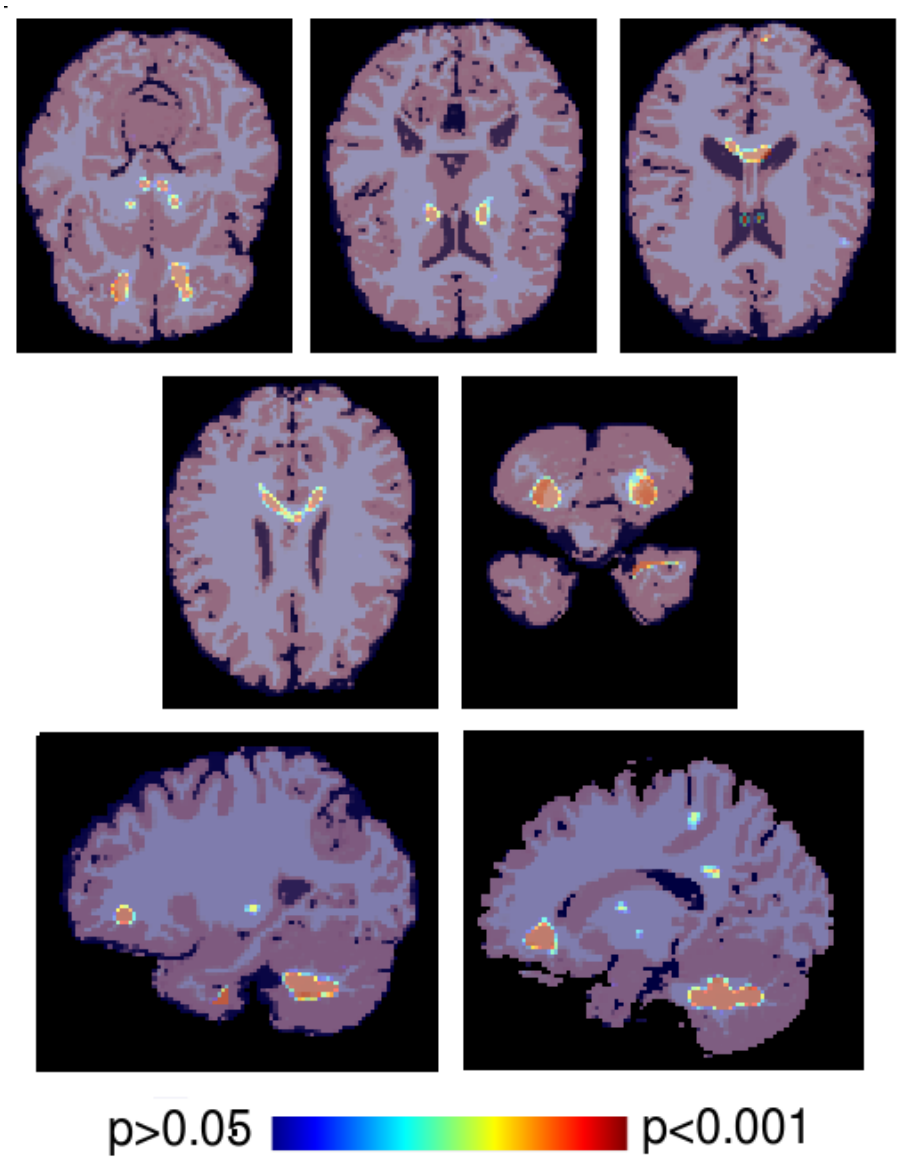


Figure 36: Regions with low p-values associated with SVM model trained on schizophrenia data

3.9. Application to the Baltimore Longitudinal study of aging

The Baltimore Longitudinal study of aging (BLSA) has been prospectively collecting multidisciplinary data related to physical and psychological aging since 1958. Its neuroimaging component, currently has followed approximately 160 individuals (aged 55 to 85 years at enrollment) with annual or semiannual imaging and clinical evaluations. The neuroimaging sub-study of the BLSA, which controls for consistency of imaging data over time, is described in detail in Resnick et al. (2003).

We used T1-weighted MR images to measure regional patterns of brain atrophy using RAVENS maps. The image acquisition parameters have been described in Resnick et al. (2000). The BLSA protocol included an axial T1-weighted volumetric spoiled gradient recalled (SPGR) series (axial acquisition, 1.5 T, repetition time (TR)=35 ms, echo time (TE) = 5 ms, flip angle = 45, voxel dimensions of 0.94x0.94x1.5 mm slice thickness). All scans were acquired on one of three GE (Schenectady, NY, USA) Signa 1.5 T scanners with similar operating systems.

Image preprocessing involved (1) alignment to the AC-PC (anterior commissure-posterior commissure) plane; (2) removal of extracranial material (skull-stripping) and cerebellum; (3) N3 bias correction (Sled et al., 1998); (4) tissue segmentation into gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), and ventricles (Pham and Prince, 1999); (5) high-dimensional image warping (Shen and Davatzikos, 2002) to a brain atlas (template) (MacDonald et al., 2000) in the Montreal Neurological Institute (MNI) standardized coordinate system; and (6) formation of regional volumetric RAVENS maps, generated to enable analyses of volume data rather than raw structural data.

Ultimately, we trained a support vector classifier on grey matter and white matter RAVENS

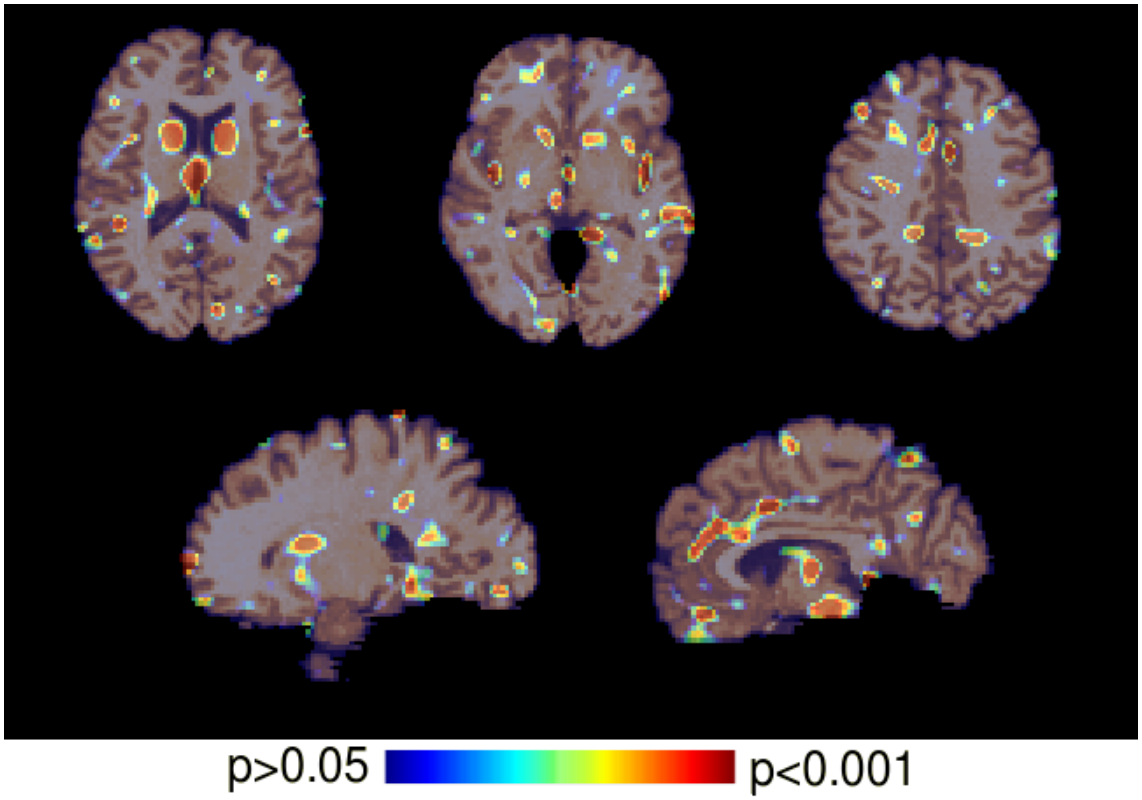


Figure 37: Regions with low p-values associated with SVM model trained on grey matter RAVENS map based classification of the sexes using BLSA data. Periventricular grey matter seems is prominently picked up by the model in addition to several cortical regions

maps to differentiate the gender between 53 females 70 males. The study is longitudinal, and acquired data from each subject over multiple time points. In experiments presented here we chose to use only RAVENS maps generated from data acquired from the first time point. The rationale behind this choice was that SVM theory assumes that data points are independent. Including scans of the same subject over multiple time points would violate this assumption. Grey matter RAVENS based classification yields a 5-fold cross validation accuracy of 81.3% for classifying males vs females. White matter RAVENS based classification yields a 5-fold cross validation accuracy of 80.5%. p-value maps associated with each SVM model are shown in figures 37 and 38.

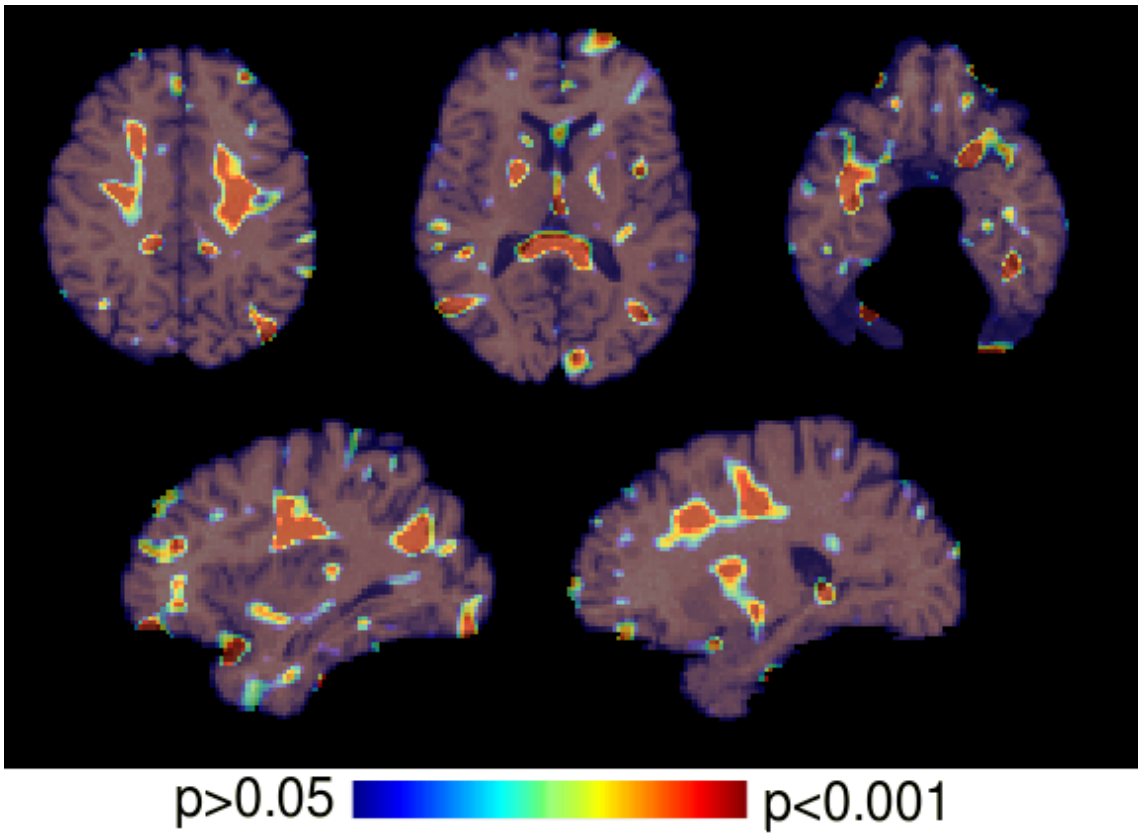


Figure 38: Regions with low p-values associated with SVM model trained on white matter RAVENS maps obtained from BLSA. White matter changes in the corpus callosum and adjacent to it are most prominently picked up by the model.

3.10. Conclusion

In conclusion, we have presented work that improves upon the statistical framework for interpretation of support vector machine models presented in the previous chapter. Our new framework explicitly incorporates the fact that SVM classification with a higher margin is superior to SVM classification with a lower margin. We have shown using simulated data that the margin based statistic alleviates certain shortcomings associated with the original weight based statistics. We have also shown that null distributions associated with this margin based statistic are normal. The proposed framework provides a statistical p-value maps for interpretation of SVM models in neuroimaging. These p-value maps make SVM based multivariate inference as accessible to use as VBM based univariate inference. They also provide a multivariate view of the phenomenon under investigation that is complementary to univariate analysis. While the speed up provided by the approximation is important , the fact that the null distributions associated with the statistics are asymptotically normal is also very important. The gaussianity of the proposed statistics, opens up a whole world of statistical properties, tests and analyses predicated on the gaussianity assumption. All of this theory can be brought to bear for understanding SVM models. To illustrate practical applications of our framework we have presented p-value maps associated with SVM models trained on data from several different imaging studies. Further discussion on potential extensions and applications of this framework is presented in the Conclusions chapter.

CHAPTER 4

Unsupervised machine learning for the analysis of heterogeneity in population neuroimaging : Clustering for heterogeneity analysis and mapping (CHAMP)

4.1. Introduction

The previous two chapters of this thesis have primarily focused on multivariate pattern analysis (MVPA) in neuroimaging using the SVM, a widely used supervised classification tool. We provided a mathematically rigorous interpretation of SVM models for neuroimaging using statistical p-values. In doing so, we showed that the SVM model uses a single brain wide pattern of differences to drive diagnosis/classification. In general, this search for a single pattern of imaging difference that differentiates phenotypically distinct populations is a common feature of other MVPA methods as well. Thus, all these analyses techniques implicitly make an assumption that 'single imaging pattern can distinguish between clini-

cally distinct populations'.

Incidentally, this is also the driving assumption behind the most widely used univariate methods for neuroimaging analyses. As explained in chapter 1, voxel-based morphometry (VBM Ashburner and Friston (2000)), and its extensions (Chung et al., 2001; Davatzikos et al., 2001), rely on univariate two sample t-tests comparing voxel-wise measurements between two populations. In doing so, they search for a single image-wide pattern that quantifies group difference between the two populations.

However, many neurological, neurodevelopmental, and neuropsychiatric disorders have a substantially heterogeneous clinical presentation (Kramer and Miller, 2000; Dickerson et al., 2011; Butters et al., 1996; Tsuang, 1975; Steen et al., 2006; Durston, 2003; Shiino et al., 2006). Given this heterogeneity, the complexity of the human brain and the subjectivity of clinical scoring it is unlikely that real diseases are driven by a homogeneous pattern of brain deficit. Mounting evidence from clinical studies points the other way. In fact, it is very much conceivable that multiple patterns of brain deficits drive the complex and heterogeneous symptomatology of most neuropsychiatric diseases. Thus, there is need for neuroimaging analysis tools that can search for multiple imaging difference patterns associated with a specific population wide clinical difference.

While traditional parametric statistical approaches do not readily yield themselves to the analysis of heterogeneity associated with population wide group differences, modern machine learning methods do. Towards this end, we propose, in this chapter, an alternate strategy for case-control analysis of neuroimaging data which allows for multiple patterns of brain changes to be associated with a population wide group difference (figure 39). This is a fundamental improvement over the work presented in previous chapters.

The framework we propose draws heavily upon concepts in existing econometrics literature on matching estimators (Todd, 2008). In order to apply unsupervised machine learning methods to the analysis of heterogeneity one needs to be able to first quantify individual

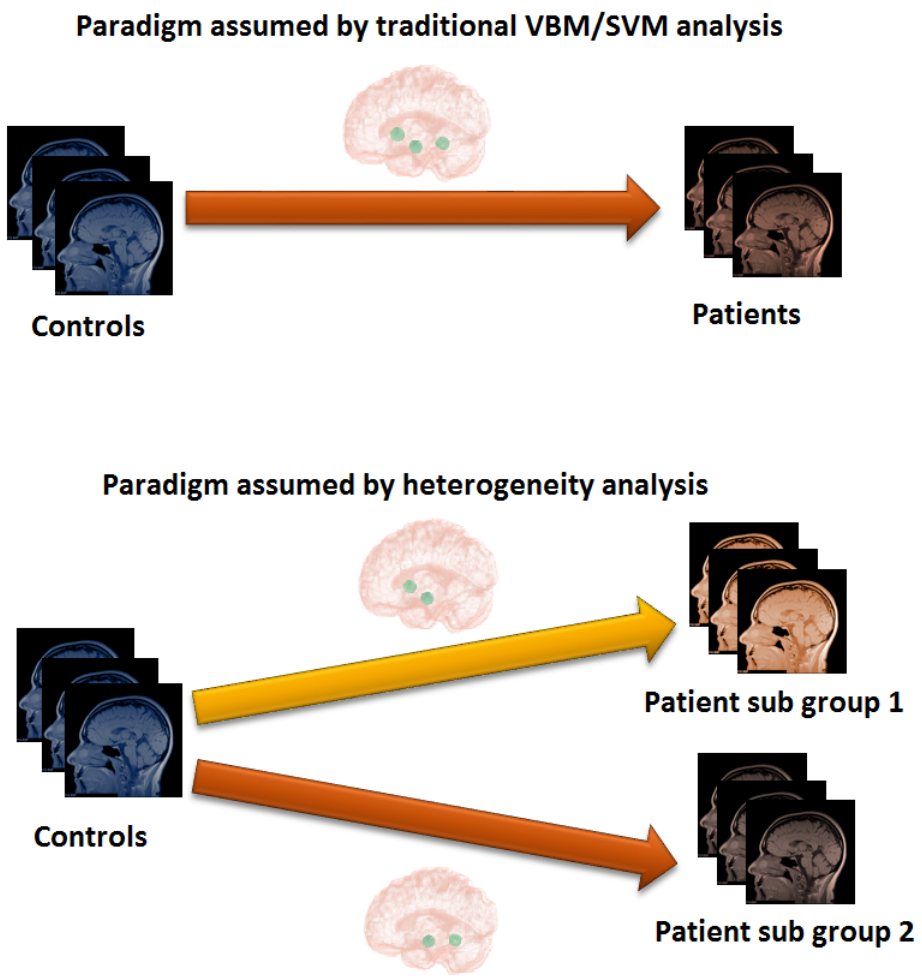


Figure 39: Difference between traditional neuroimaging analysis paradigm used by VBM/MVPA from heterogeneity analysis proposed in this chapter.

specific effects of a given disorder. In the theory of matching estimators this is done by assuming that every individual in the study can exist in two states, either as a 'case' or as a 'control'. In any dataset collected by a clinical neuroimaging study, only one of these outcomes can be actually observed. However, assessing the impact of a physiological process driving a population-wide phenotypic difference requires us to infer what 'case' images would look like had they been observed in the 'control' population. The theory of univariate matching estimators provides several specific procedures for achieving this end. We develop a high dimensional analog of the simplest of these matching procedures, namely the nearest neighbor matching estimator. We use the high dimensional nearest neighbor matching estimator to quantify individual specific effects of disease. Then, we apply unsupervised learning (or clustering) to summarize the effects of disease into a few representative patterns of deficit measured by imaging.

In summary, the primary aim of this chapter is to present a framework that urges the neuroimaging community to ask the question, "How many and what patterns of imaging differences can we identify between two groups?" instead of the question, "By what single pattern do two groups differ?". We present clustering for heterogeneity analysis and mapping (CHAMP) a methodological framework that systematically addresses the aforementioned question and characterizes the heterogeneity of difference between two groups of brain images. Although our experiments are based on identifying group differences in brain structure, CHAMP is a general method that can be used in a broader spectrum of imaging analyses.

4.2. Why to use unsupervised analysis for population neuroimaging: Heterogeneity in neurological disorders

The diagnosis of neuropsychiatric disorders is currently made based on standards set forth in the Diagnostic and Statistical Manual of Mental Disorders (DSM 5). The criteria set forth in this manual are mostly based on patient responses to clinical interviews rather than on a specific neurobiological or neuroimaging basis. For instance, a diagnosis of depression may be based on a constellation of symptoms such as loss of appetite, loss of pleasure in formerly pleasurable activity, sleep loss, and a general feeling of depression for more than two weeks. Clearly, this is a rather nebulous constellation of signs and symptoms. The population of individuals meeting these criteria will be quite heterogeneous (Nikolcheva et al., 2011). When neuroimaging studies recruit patients based on these complex criteria, the data collected must automatically reflect the inherent heterogeneity of the diseased population.

Thus, heterogeneous imaging phenotypes associated with neuropsychiatric disease have been reported in several studies. For instance (Noh et al., 2014) divided early stages of of AD dementia into, medial temporal-dominant atrophy, parietal-dominant subtype and diffuse atrophy subtype. Similarly, (Sauer, 2012) reviews brain imaging morphometry studies addressing the issue of heterogeneity within the diagnostic category of schizophrenia. They implicate three different patterns of deficit all of which show an overlap in frontal changes but diverge in terms of structural deficits in other areas such as the thalamus, hippocampus, or cerebellum. In a similar spirit (Lenroot and Yeung, 2013) discuss imaging heterogeneity in autism and (Dosenbach et al., 2013) present evidence for imaging based heterogeneity in attention deficit and hyperactivity disorder (ADHD). Altogether, these studies demonstrate that brain structure is not a uniform endophenotype for any neuropsychiatric disorder. In

fact several distinct combinations of regional deficits may be the primary characteristic of disease.

Thus, one of the key challenges to the development of neuroimaging analyses tools for neuropsychiatric disorders is allowing for this 'heterogeneity' of disease. Addressing this challenge in a principled manner using imaging can substantially further our understanding of these diseases and provide information that may better reflect the underlying biology as compared to symptoms defined by the DSM. This is the key challenge addressed in this chapter.

4.3. The key challenge: Heterogeneity analysis in the presence of confounding variation

The main purpose of this chapter is to present an approach to answering the question: "In how many different ways (patterns) does group 2 differ from group 1?". We propose to answer this question by clustering imaging-based measures of differences between the two groups under comparison: 'group 1' and 'group 2.' Figure 40 provides a visual illustration of the rationale behind our question and the challenges involved in addressing it. In the cartoon presented in Figure 40, size variation is present in both (green and blue) groups. However, color variation is exclusive to group 2, reflecting that there are two patterns of difference between groups 1 and 2 (blue vs light green, and blue vs. dark green). We wish to develop an analysis technique that highlights each of these patterns of difference individually. It is simple enough to see that comparing the images using voxel-based analysis or generic MVPA methods will not achieve this end. A possible alternative would be to use unsupervised analysis of the cartoons in group 2. However, this might not always achieve this end either. More complex approaches such as principal components analysis

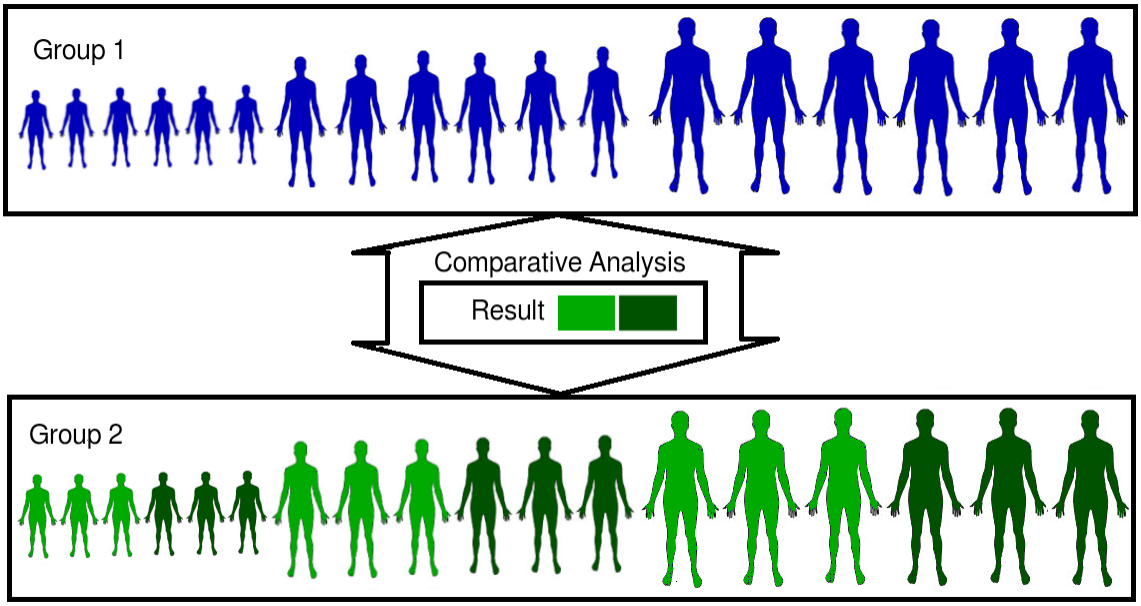


Figure 40: Rationale behind clustering differences between the groups instead of data from the groups themselves. We wish to design a simple method which can tell us that group 2 in this cartoon differs from group 1 in color, and that there are two such differences (dark green vs blue, and light green vs. blue). Standard analytic approaches that search for a single pattern of difference between these groups cannot achieve this. If we directly cluster the images in group 2 we might get clusters based largely on the size of the cartoon as opposed to the group difference. Thus, the need for difference-based heterogeneity analysis tools like CHAMP.

might yield directions of variation that capture the heterogeneity in color, but these too latch onto every possible source of variation in group 2. What we need, is an approach that highlights the variation in the differences between the two groups while suppressing variation that is common to both the groups. Undoubtedly, there exist several methods that may be used to approach this question. CHAMP as presented here, offers but one possible solution to this veritably complex problem. In the next section, we develop CHAMP in the context of neuroimaging analyses. It should be noted that the questions asked and the concepts presented are much more generic. As such, they may be applied to the analysis of group differences in other biological data as well.

4.4. Addressing the challenge: The approach

4.4.0.1. Overview and notations

For the remainder of this manuscript we use the following standard set of notations. The dataset contains m_1 appropriately processed brain images from group 1, and m_2 such images from group 2. We wish to analyze heterogeneity in group 2 with respect to group 1. Without loss of generality, we assume a vector representation of three dimensional images. Such a representation may be generated by concatenating voxel intensity values into a single long vector of dimensionality equal to the total number of image voxels (d). Keeping with this notation, we index the vectors associated with images of the first group with j as $\mathbf{x}_j^1 \in \mathbf{R}^{1\times d}$ where $j \in \{1, \cdots m_1\}$. Vectors associated with group 2 images are denoted as $\mathbf{x}_i^2 \in \mathbf{R}^{1\times d}$ and indexed by $i \in \{1, \cdots m_2\}$. One may stack all the vectors associated with the first group into a matrix denoted by $\mathbf{X}_1 \in \mathbf{R}^{m_1\times d}$ and vectors from group 2 into a matrix denoted by $\mathbf{X}_2 \in \mathbf{R}^{m_2\times d}$. In addition to imaging data, most clinical studies collect a large set of ancillary clinical variables as well. Examples of such variables may be age, sex, ethnicity, scanner types, and even the concomitant presence or absence of another disease. CHAMP uses a sub set of these ancillary variables that are a) distributed independently of the group labeling under study and b)have a substantial influence on brain anatomy and function. A prime example of such a variable might be 'age' in an 'age-matched' imaging study. Knowledge of a subject's age gives us no additional information about it's group label in such a study. Yet, it's influence on brain anatomy is undeniable. We call such a variable a 'meta-variable' and matching subjects between groups using these 'meta-variables' is the basis of the CHAMP approach. We denote the collection meta-variables associated with $\mathbf{x}_j^1$ in the vector $\mathbf{v}_j^1$ and meta-variables associated with $\mathbf{x}_i^2$ with $\mathbf{v}_i^2$. These meta-variable vectors may be assumed to be multi-dimensional, but their dimensionality is assumed to be much lower than the sample size of the study.

4.4.1. The difference representation matrix

As stated earlier, we aim to identify heterogeneous patterns in $\mathbf{X}_2$ that are not concomitantly present in $\mathbf{X}_1$, and are associated with the differences between groups 1 and 2. We operate under the assumption that the majority of such variation is driven by the group difference itself. Here, we propose to encapsulate this variation into a difference representation matrix, denoted by $\mathbf{D} \in \mathbf{R}^{m_2 \times d}$. The subsequent text presents a procedure to construct such a matrix alongside the intuition behind this procedure.

Assume $p(\mathbf{x}^1)$ and $p(\mathbf{x}^2)$ to be probability distributions associated with groups 1 and 2. Recall that we chose meta-variable distribution to be independent of group labeling. Thus, if we assume $p(\mathbf{v})$ to denote the distribution from which meta-variables are sampled, then we may write:

$$p(\mathbf{x}^1) = \int_{\mathbf{v} \in \mathcal{V}} p(\mathbf{x}^1|\mathbf{v})p(\mathbf{v})d\mathbf{v} \tag{4.1}$$

and

$$p(\mathbf{x}^2) = \int_{\mathbf{v} \in \mathcal{V}} p(\mathbf{x}^2|\mathbf{v})p(\mathbf{v})d\mathbf{v} \tag{4.2}$$

where $\mathcal{V}$ indicates the domain of $\mathbf{v}$. In the expressions above we may be able to model $p(\mathbf{v})$ using data. However, it is nearly impossible to estimate the distributions $p(\mathbf{x}^1|\mathbf{v})$ and $p(\mathbf{x}^2|\mathbf{v})$ from sample sizes typically available in neuroimaging studies. This is because imaging data are high dimensional in nature. Yet, we use these conditional distributions to elucidate certain concepts that are critical in the design of our method.

Consider, $\mathbf{x}_i^2 \sim p(\mathbf{x}^2)$ to be a subject that was observed in group 2. Define $\mathbf{x}_{j(i)}^1 \sim p(\mathbf{x}^1)$ to be the image that would have been observed for subject i if, contrary to fact, subject i had been observed in group 1. That is, we would have observed $\mathbf{x}_{j(i)}^1$ had this subject been spared of the pathological process that drove subjects from group 1 to group 2. In the statistics and causal inference literature, $\mathbf{x}_{j(i)}^1$ is referred to as a *counterfactual* or *potential*

outcome Rubin (1974) Holland (1986) Rubin (2005).

Under certain assumptions, it is possible to obtain a reasonable estimate of $\mathbf{x}^1_{j(i)}$ using the observed data from group 1. First, as previously mentioned, we must assume that $\mathbf{v}$ arises from common distribution $p(\mathbf{v})$ for both groups. Second, we must assume most or all of the variation in the images beyond that caused by the disease itself is captured by $\mathbf{v}$. When these assumptions hold, $\mathbf{x}^1_{j(i)}$ can be estimated using data from patient group 1. Next, we outline our proposed approach for estimating $\mathbf{x}^1_{j(i)}$ for each patient.

In this work, we propose to model $\mathbf{x}^1_{j(i)}$ as the mean of the conditional distribution $p(\mathbf{x}^1|\mathbf{v} = \mathbf{v}^2_i)$. The intuition behind using such an estimation is that meta-variables like age and sex provide a fairly low dimensional measure that captures a large proportion of variation in normal brain anatomy. Like the size of the cartoon in figure 1, the variation captured by meta-variables is assumed to be common across the populations being compared. Thus, the mean of $p(\mathbf{x}^1|\mathbf{v} = \mathbf{v}^2_i)$ provides a reasonable estimate of how the true $\mathbf{x}^1_{j(i)}$ might look like. Formally, one may write:

$$\mathbf{x}^1_{j(i)} = \int_{\mathbf{x}^1 \in \mathcal{X}_1} \mathbf{x}^1 p(\mathbf{x}^1|\mathbf{v} = \mathbf{v}^i_2) d\mathbf{x}^1, \tag{4.3}$$

where the integration only uses the subset of $\mathbf{x}^1$ that has meta-variable values which match $\mathbf{v}^i_2$. This definition of $\mathbf{x}^1_{j(i)}$ is based on a distribution that is conditioned on the meta-data rather than the full distribution of the imaging data themselves.

Now in real data, sample size constraints severely limit the total number of samples with $\mathbf{v} = \mathbf{v}^i_2$. So we estimate $\mathbf{x}^1_{j(i)}$ using the mean of r samples chosen from group 1 which have meta-variable values closest to $\mathbf{v}^i_2$ in a Euclidean sense. If we let $l \in S_i$ index this set of r samples for a particular $\mathbf{x}^2_i$, then we may approximate $\mathbf{x}^1_{j(i)}$ as:

$$\hat{\mathbf{x}}^1_{j(i)} = \frac{1}{r} \sum_{l \in S_i} \mathbf{x}^1_l. \tag{4.4}$$

The expression on the right of the above expression represents our best estimate of $\mathbf{x}^1_{j(i)}$. Using our estimate of $\mathbf{x}^1_{j(i)}$, we may finally compute:

$$\mathbf{d}_i = \mathbf{x}^2_i - \hat{\mathbf{x}}^1_{j(i)}. \tag{4.5}$$

This vector represents our best guess of the component of $\mathbf{x}^2_i$ that arises from phenomena driving the group difference. Thus, the distributional assumptions made above ultimately lead to a construction of $\mathbf{d}_i$ that is based on nearest 'group 1' neighbors of $\mathbf{x}^2_i$ where,'nearness' is defined using meta-data rather than the imaging itself. This contrasts the traditional machine learning perspective, where feature vectors themselves are used to define 'nearness'. The intuition behind this meta-variable centric approach may be exemplified by considering the following scenario. Suppose, we were analyzing imaging data from a study comparing autistics to normal controls. If a six year old male autistic child did not have autism, his brain should look like a six year old normal male child. It would be incorrect to assume that a 6 year old autistic male brain would look like to a 4 year old normal female brain if autism never manifested. The invalidity of such an assumption would hold in spite of any observed similarity in terms of brain structure/imaging.

This idea is presented visually in figures 41 and 46 and captured in the theory presented here. Thus, we use meta-variable based neighborhoods to compute the vectors $\mathbf{d}_i$ for all $i \in \{1, \cdots, m_2\}$ and stack them row-wise to construct the matrix $\mathbf{D}$. We theorize that this matrix exposes variation associated with group differences.

Heterogeneity visualization and mapping

The study of heterogeneity using the difference vectors $\mathbf{d}_i$ is the search for clusters of self similar difference vectors. A re-ordered cross-correlations matrix-based approach may be used to visualize similarity between the rows of $\mathbf{D}$. While it is the similarity between the rows of the matrix that we need to capitalize on, no additional information is conveyed by

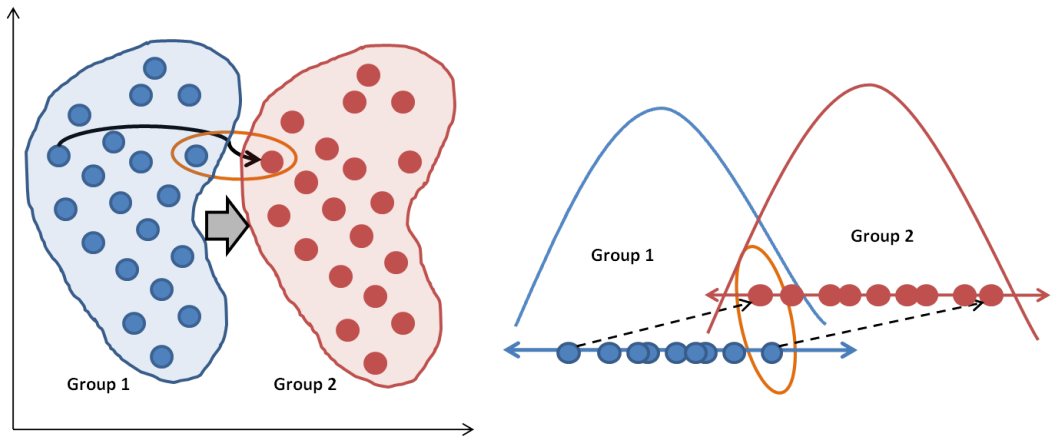


Figure 41: Group differences are most likely expressed as changes in voxel intensity in case of imaging data. Thus, if group 2 was essentially generated by subtracting or adding a fixed number to the intensities in group 1, neighborhoods based on Euclidean distances would not generate appropriate difference maps. In the illustrations above the orange ellipses indicate Euclidean distance based neighbors, whereas the black arrows indicate the neighbors we want to find.

the extremely high similarity of every row with itself. With this in mind, we compute the hollow matrix $\mathbf{F} \in \mathbf{R}^{m_2 \times m_2}$ with elements given by:

$$F_{pq} = (1 - \delta_{pq}) K_{pq}, \tag{4.6}$$

where the matrix $\mathbf{K}$ is defined as: $\mathbf{K} = \mathbf{D}\mathbf{D}^{\mathbf{T}}$ and δ_{pq} is the Kronecker delta symbol. Using the rows of F as features in a k-means algorithm we define cluster memberships that define sub groups of group 2. In order to understand how each of these subgroups differs from group 1, we run voxelwise univariate analysis between each subgroup and group 1. We visualize the resulting p-value maps by overlaying them on a template brain.

4.5. Experiments and results

In this section we present results generated by applying CHAMP on the ADNI dataset. We used controls as group 1 and patients as group 2 for our experiments. Thus, we have used CHAMP to investigate heterogeneity in patients with respect to controls. Age and sex are used as meta-variables. The dataset we use contains 100 controls and 100 Alzheimer's patients. We show results for various different parameter settings of the method. We discuss these results and contrast them with other plausible approaches to heterogeneity analysis in the discussion at the end of the chapter.

4.5.1. Data preprocessing for ADNI

Raw ADNI data from the ADNI-1 study was used for all experiments presented in this work. The data was bias corrected with N3 and skull stripped using multi-atlas skull stripping (MASS Doshi et al. (2013)). This was followed by segmentation into three tissue types: gray matter (GM), white matter (WM), and ventricles (VN) using MICO (Li et al., 2014). DRAMMS (Ou et al., 2011) registration was run to register each subject to a common template. Tissue density maps were generated from the resulting deformation fields using the RAVENS (Davatzikos et al., 2001) approach. We exclusively use GM tissue density maps (GM-RAVENS) in the present work.

4.5.2. Results using simulated data

Before applying the method to the actual ADNI data we demonstrate its validity by using it on simulated data with known ground truth. The data simulations were performed using imaging data from control subjects drawn from the ADNI study. We divided these data

into two equal groups: simulated controls and simulated patients. We introduced two separate patterns of atrophy in the simulated patient group by reducing the intensity of the tissue density maps in specific locations by fifty percent. These ground truth patterns of atrophy are shown in figure 42. The first pattern of atrophy was introduced in half of the simulated patients and the second pattern in the other half. These patterns were meant to mimic heterogeneous effects of disease. In addition to the two patterns of atrophy, we also introduced a global reduction of GM tissue density values linked with a randomly generated meta-variable. The association between the randomly generated meta-variable and percent reduction of atrophy was quadratic. This meta-variable, and the associated global simulated reduction of tissue density values was intended to mimic the highly nonlinear and multivariate effects of similar meta-variables (like age) in real neuroimaging data. These simulations are naive in comparison to the formidable complexity that is likely to underlie the effects of disease, age, sex, ethnicity, or any other such variables in real data. However, the simulations did allow us to 1) validate our methodology and 2) provide interesting insights into the workings of this method. CHAMP could successfully detect the simulated clusters in this data. VBM comparison of each cluster (produced by CHAMP) to simulated controls clearly delineates the two distinct patterns of simulated atrophy, present in the data. The results are summarized in Figure 42. Similar results can be obtained for either a 40% or a 30% reduction in gray matter tissue density.

4.5.3. Results from ADNI

We applied CHAMP to investigate heterogeneity of brain atrophy in AD, using data from the ADNI. Recall that the two parameters involved in applying CHAMP include a) r - the number of neighbors from group 1 used in the modeling, and b) k the number of clusters the patient group is divided into.

In what follows we highlight how clusters produced by CHAMP using different parameter

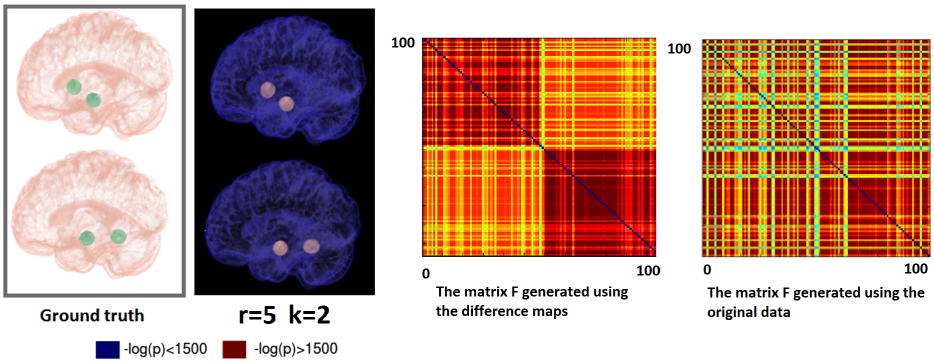


Figure 42: The leftmost column shows the regions where the atrophy was simulated by reducing tissue density maps by 50%. The next column shows that p-value maps generated by CHAMP can be used to identify regions of simulated deficit in the imaging data. The two rightmost columns show that a re-ordered F matrix generated using the difference maps is reflective of the latent structure of disease heterogeneity. This is not true if the original data are themselves used to generate this image.

values might be reflective of processes traditionally measured using clinical scores. Towards this end, we investigate CHAMP clusters with respect to scores generated using the mini mental state examination (MMSE). The MMSE (Folstein et al., 1975) standard verbal test is based on a 30 point questionnaire, and it is used extensively in clinical and research settings to measure cognitive impairment (Pangman et al., 2000) and to diagnose Alzheimer's dementia. Scores greater than or equal to 27 points indicate normal cognition. Scores below 27 indicate cognitive impairment. The lower this score, the worse the dementia.

Figures 43 and 44 summarize results produced by CHAMP using ADNI data and parameter values $r = 10$ and $k = 2$. The two clusters produced by CHAMP differ significantly in terms of MMSE. VBM based p-value maps quantifying the difference between each cluster and controls are shown in figure 43. The two VBM maps differ from each other substantially. The p-value map associated with the cluster with higher MMSE shows deficits of the hippocampal and para-hippocampal regions alone. On the other hand, the map associated with the cluster with lower MMSE scores shows difference in the hippocampus, the entorhinal cortex and the precuneus. The differences between the clusters themselves can be further elucidated by running VBM to explore imaging-based differences between

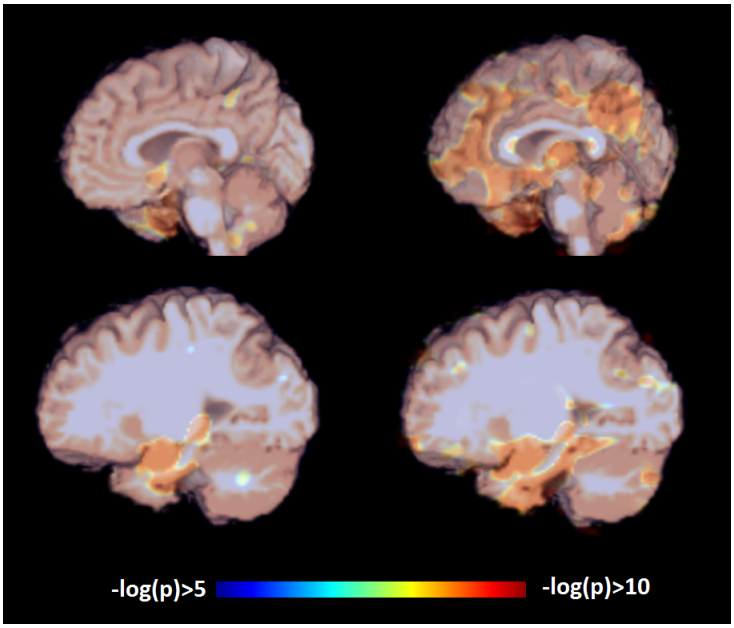


Figure 43: Cluster 1 is significantly different from the controls in the hippocampus and the parahippocampal GM regions. Cluster 2 shows strong deficits in the hippocampal regions, the precuneus and the entorhinal cortices.

these two clusters. The associated p-value maps are shown in figure 45. Note that the patient subgroups generated by CHAMP differ significantly in terms of the MMSE, but they do not significantly differ with respect to age, implying that the effects elucidated by the method are not reflective of aging itself, but rather of the disease process, and that this process seems to be relatively invariant to age. In short, we show that, CHAMP produces clinically distinct patient subgroups using imaging data, along with age and sex information.

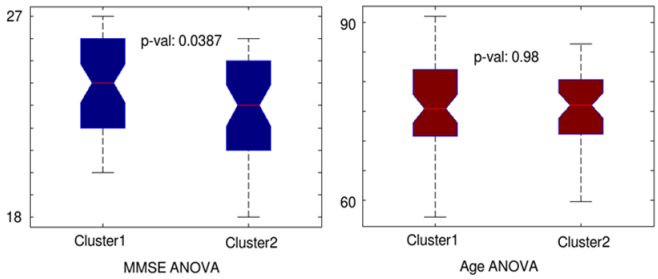


Figure 44: Cluster 1 has significantly higher MMSE values as compared to cluster 2. However, the two clusters do not differ in terms of age.

In order to highlight stability of CHAMP results with respect to perturbations in parameter values, we compute CHAMP clusters using several different parameter settings and show p-value maps and ANOVA plots corresponding each setting. Results generated using $k = 3$ and $r = 10$ are shown in figure 46. These show substantial difference between clusters in terms of MMSE, but not in age. In fact the cluster associated with the p-value map showing a deficits in a larger number of brain regions also corresponds to the cluster with worse MMSE scores. When we perform CHAMP analysis with $k = 4$ and $r = 10$ the clusters generated do not differ significantly in terms of age or MMSE (figure 47). This is despite the fact that relatively familiar patterns do show up in the associated p-value maps. Based on these results we chose not to proceed with values of $k > 4$. The results presented in figures 43 through 47 give us valuable insight into variation of CHAMP results with respect to number of clusters k. Next, we explore how the results vary with respect to variations in 'r'.

To do this we apply CHAMP by keeping k fixed at 2 and varying r. This yields figures 48 and 49. p-value map patterns similar to those shown in Figure 43 appear. ANOVA reveals differences between MMSE amongst the clusters generated without a corresponding difference in age.

While the results presented thus far provide interesting insights into the workings of CHAMP at various parameter settings, they do not provide a broad and objective evaluation of the behavior of the method over a large range of parameters. However, understanding such behavior is paramount in guiding parameter selection for heterogeneity analysis in a new dataset. Towards this purpose, we use the following procedure. We generate CHAMP clusters for values of r ranging from 2 to 30 and values of k ranging from 2 to 4. We perform ANOVA to evaluate the difference between MMSE (and age), between clusters generated at each setting. We plot the negative logarithm of the associated ANOVA p-values for every possible parameter value in figures 50 to 52. These plots are discussed in more detail in the subsequent 'Discussions' section of this chapter.

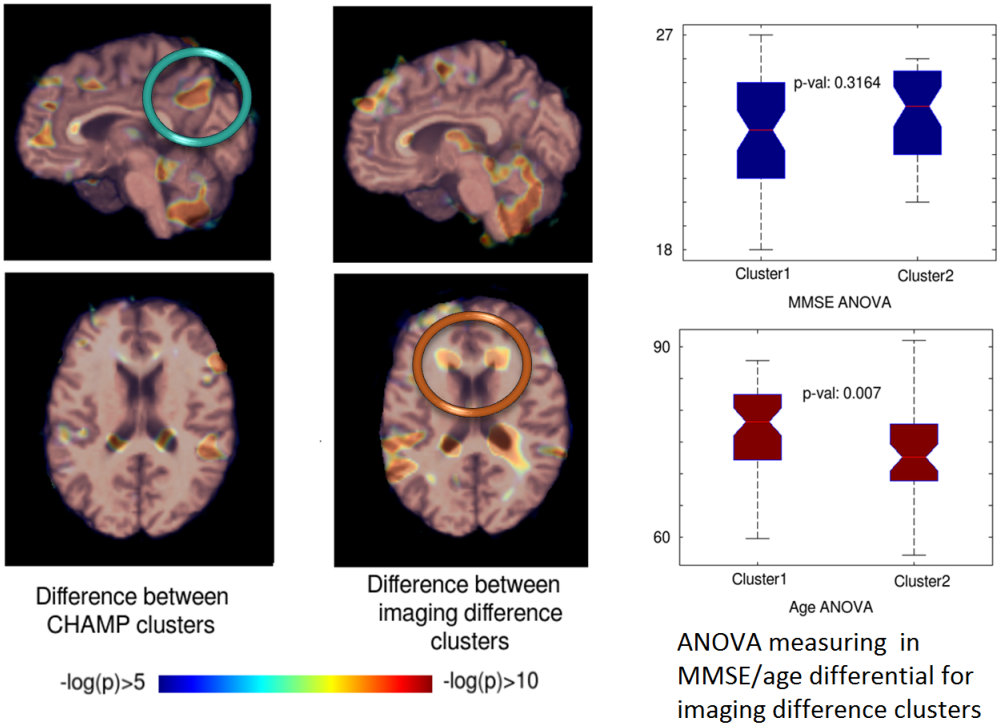


Figure 45: First column shows p-value maps highlighting differences between clusters generated using CHAMP which uses the proposed method. Second column shows similar differences generated using imaging-based difference maps. Third column shows ANOVA measuring how MMSE/age differ between clusters generated using imaging-based difference maps.

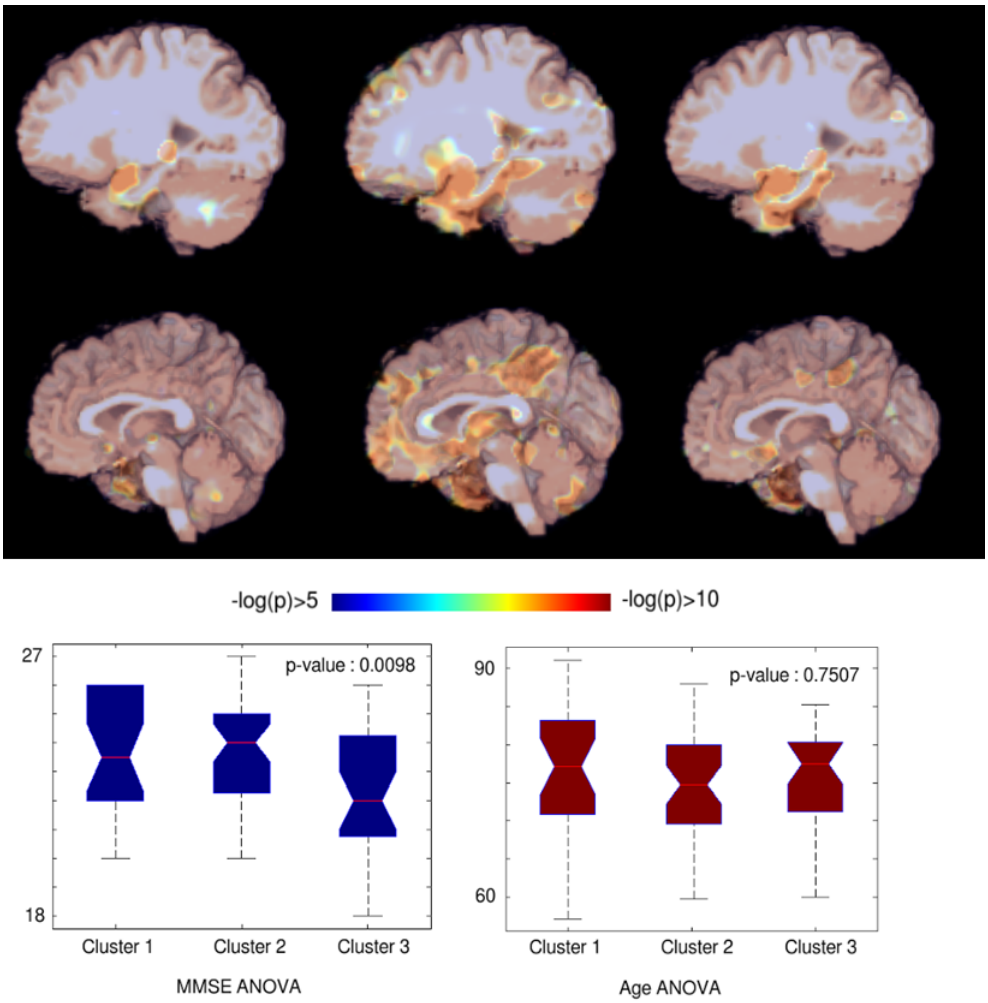


Figure 46: CHAMP results generated from ADNI data using k=3 and r=10. Cluster 1 presents deficits mainly in the hippocampus, cluster 2 in the hippocampus and small regions of the precuneus, cluster 3 and entorhinal cortex and hippocampus and precuneus. The MMSE differs substantially between the three clusters while the age does not.

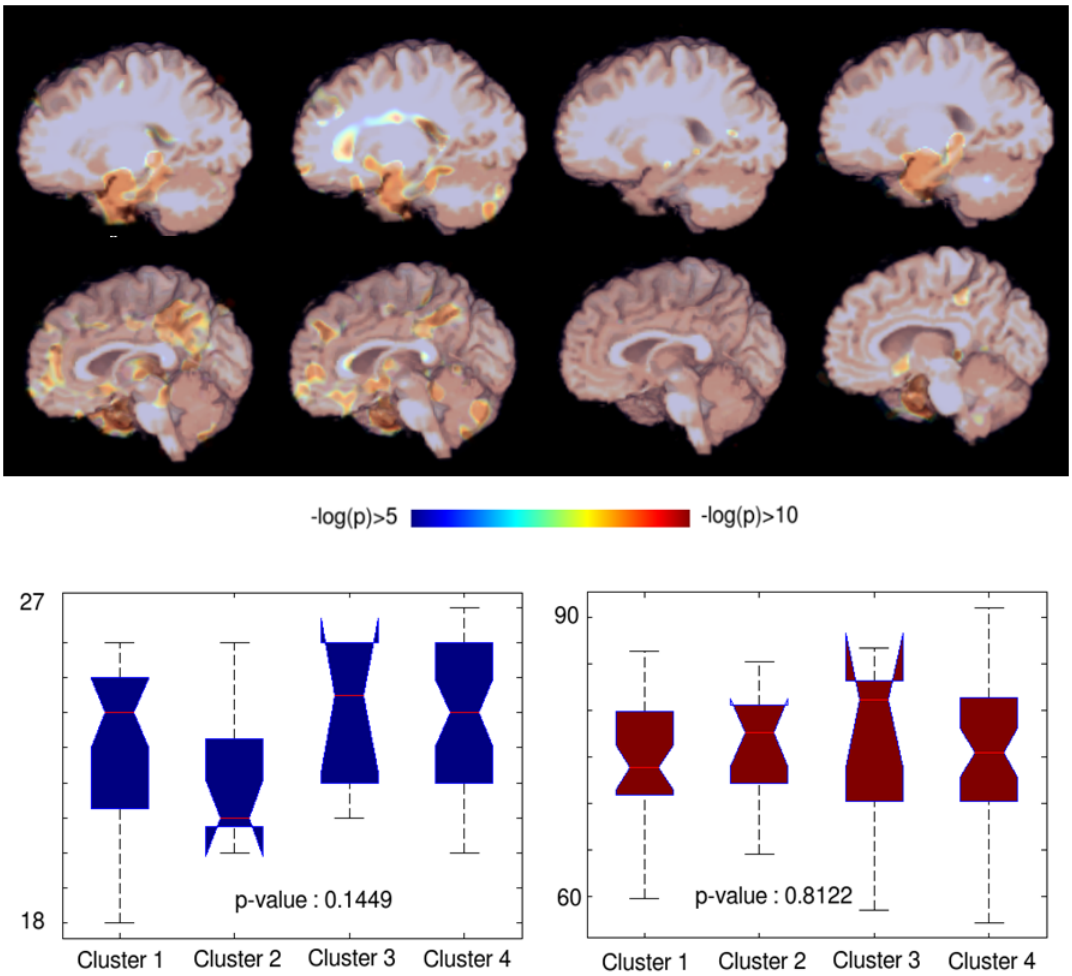


Figure 47: CHAMP results generated from ADNI data using k=4 and r=10. While some of the patterns are familiar the clusters do not differ amongst themselves in a statistically significant manner.

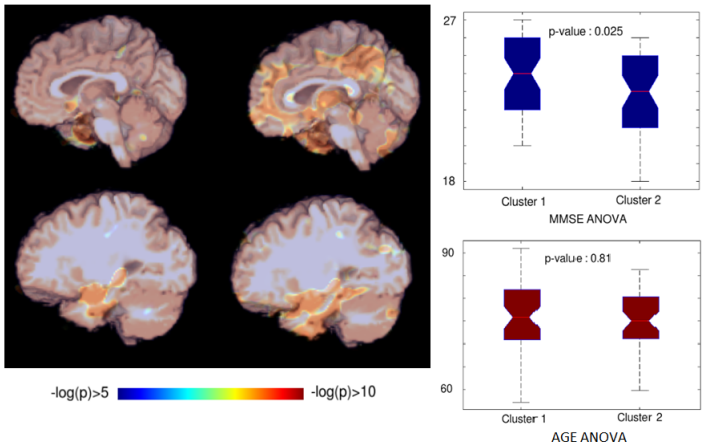


Figure 48: CHAMP results with number of nearest neighbors set (r) set to 8 and number of clusters set to 2.

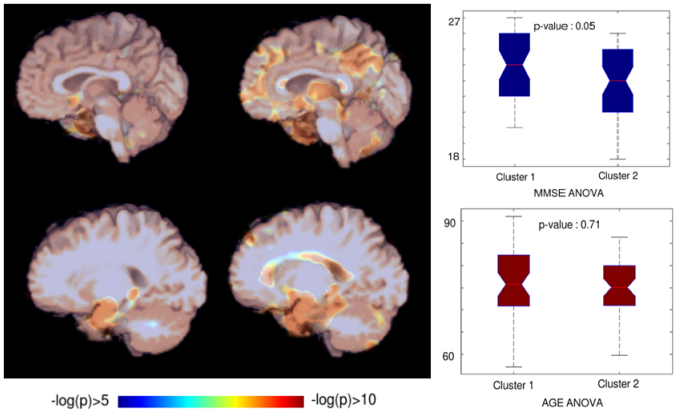


Figure 49: CHAMP results with number of nearest neighbors set (r) set to 12 and number of clusters set to 2.

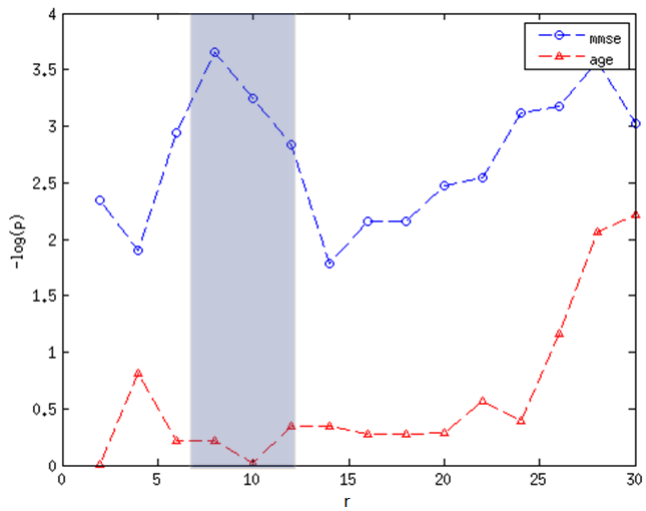


Figure 50: Negative logarithm of ANOVA p-values associated with clusters generated by CHAMP for various values of r and for $k = 2$.

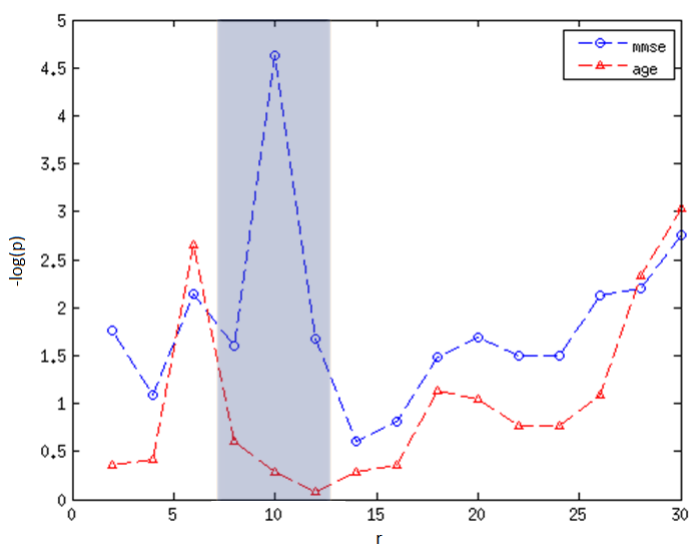


Figure 51: Negative logarithm of ANOVA p-values associated with clusters generated by CHAMP for various values of r and for $k = 3$.

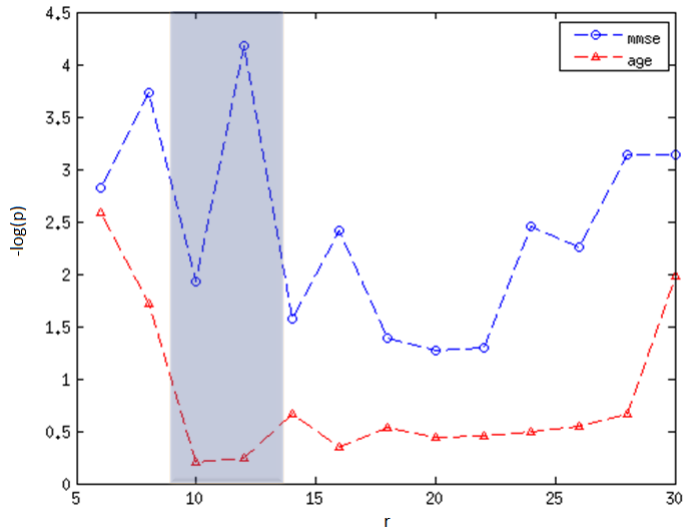


Figure 52: Negative logarithm of ANOVA p-values associated with clusters generated by CHAMP for various values of r and for $k = 4$.

4.5.4. Experiments on ADNI comparing CHAMP to prior art

In this section we present experiments that expose the value of the CHAMP approach by comparing it to existing image based heterogeneity analyses.

First we present an experiment comparing CHAMP difference map clustering to direct data clustering itself. This clustering method was used in Noh et al. (2014). Towards this end we perform data clustering using the grey matter RAVENS maps of patients. We compare the resulting clusters with respect to age and MMSE. These results are presented in figure 53. The resulting clusters differ significantly in terms of age but not in terms of MMSE.

Next we present an experiment comparing CHAMP clusters to clusters generated using 'image based' difference maps. The latter approach forms the basis of the heterogeneity analysis method proposed in Gaonkar et al. (2011). This approach estimates $\mathbf{x}^1_{j(i)}$ using r controls closest to a chosen patient, as measured by the Euclidean distance between $\mathbf{x}^2_i$ and vector representations of images from 'group 1'. We set r to 10 and computed these 'image based difference maps' using GM-RAVENS maps in ADNI data. Clustering was done with

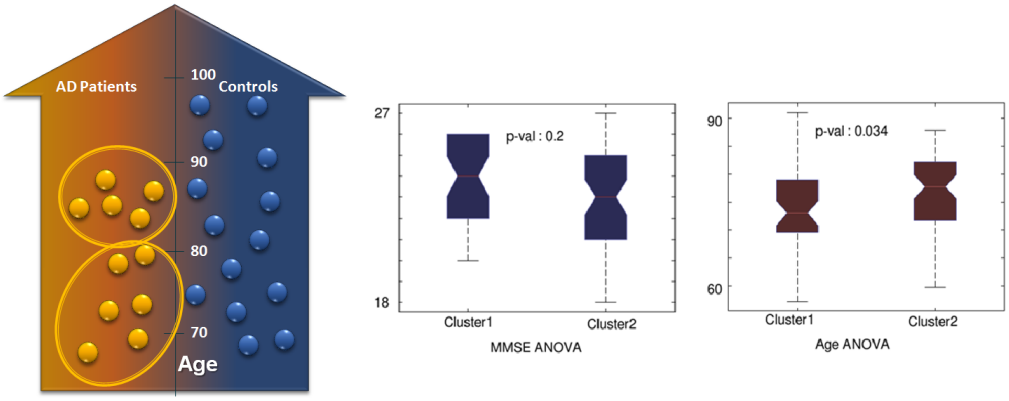


Figure 53: (Left) Illustration of why clustering ADNI data directly results in age based clusters. (Right) Comparison of MMSE and age using clusters generated with the ADNI patient data itself.

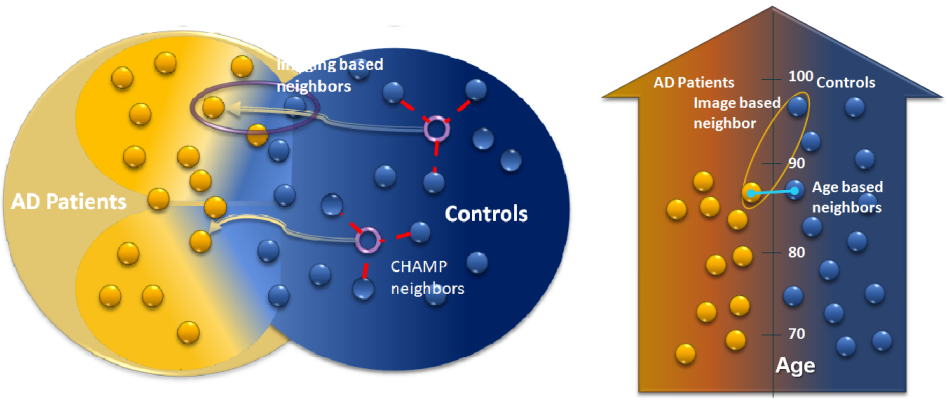


Figure 54: (Left) Why imaging based neighborhoods may not be appropriate for heterogeneity analysis. (Right) How imaging based neighborhoods are confounded by age in ADNI data

$k = 2$. The subgroups generated differ significantly in age but not MMSE. The associated results are presented in figure 45, where they are also contrasted with CHAMP results for the same parameter values.

Figure 45 shows why meta-data based difference maps should be used instead of image based difference maps. The p-value maps show that clusters generated using 'image-based differences' differ from each other mainly in terms of peri-ventricular lesions. Peri-ventricular lesions are a hallmark of dementia and aging and are not necessarily specific to Alzheimer's patients (Barber et al., 1999). Thus, these clusters mainly highlight the effect of aging. This confirms the trend presented by the ANOVA in Figure 45. On the other hand, clusters generated using CHAMP differ in the precuneus and the entorhinal cortex. Deficits in both of these are known to be associated with Alzheimer's disease (Karas et al., 2007; Gómez-Isla et al., 1996). This highlights the value of using meta data for defining matches prior to clustering in CHAMP.

4.6. Discussion

The method presented above is designed to highlight heterogeneity using neuroimaging data from large case-control cohorts. The driving principle behind our analysis is the clustering of differences between cases and controls. This is distinct from traditional approaches used for the analysis of these data. Traditionally, analyses used in neuroimaging have been based on methods assuming that the two groups differ by a single pattern, which is discovered via ROI-, VBA-, or MVPA-based analyses. These emphasize homogeneous disease patterns, or else find a "common denominator". However, it is known that diseases affect each individual in a different manner, and that there is substantial heterogeneity in the clinical presentation of any neurological disorder. Similarly, in fMRI task-activation studies, there might be heterogeneity in how individuals respond to external stimuli. Structural and func-

tional neuroimaging data present an unprecedented opportunity to quantitatively delineate case sub-populations that differ from one another neurophysiologically, a variability that CHAMP aims to tease out.

In what follows we discuss key concepts that form the basis of CHAMP. We discuss these concepts in relation to experiments and results presented above.

4.6.1. Key concept of CHAMP analysis 1: Using difference maps generated from data for clustering instead of using the data directly

Majority of the contemporary literature attempting to explore disease heterogeneity using imaging relies on either a) direct clustering of image data (Noh et al., 2014) or b) clinical definition of heterogeneity (Kramer and Miller, 2000; Dickerson et al., 2011; Butters et al., 1996). CHAMP uses a substantially different mechanism. It relies on clustering a data driven estimate of an individual specific effect of disease. One can think of this individual specific 'disease effect' as the high dimensional analog of the 'treatment effect' in the theory of matching estimators. This high dimensional disease effect is the difference map used by CHAMP.

The use of image difference maps to aggregate information related to population-wide group separation using neuroimaging data is a novel aspect of CHAMP. CHAMP aggregates these effects into a difference matrix which serves as a quantifier of the effects of a group difference on a population. This is distinct from the theory of matching estimators, where all the effects would be aggregated to obtain a mean effect.

This difference matrix is a unique and malleable representation of group difference information which opens up a lot of avenues for future work. CHAMP clusters rows of this matrix in order to explore disease heterogeneity using imaging. Quantitatively, the advantage of

clustering image difference maps stored in the difference matrix as compared to images themselves, can be seen by comparing figure 14 to figure 5. Clusters generated by using the difference maps differ significantly in terms of MMSE and clusters generated using the images differ in significantly in terms of age. This is encouraging since MMSE is a proven measure dementia and differences in MMSE between patients with a similar age distribution is a credible sign of disease heterogeneity.

4.6.2. Key concept of CHAMP analysis 2: Meta-data based difference maps versus image-based difference maps

CHAMP uses meta data for computing matches between the groups. Standard machine learning techniques which form the foundation of neuroimaging MVPA-analysis use data itself to compute 'nearness'. In this sense CHAMP is fundamentally different from MVPA-analyses which are considered state of the art in neuroimaging.

This makes it critical to make the right choice of meta-variables. It is this choice, rather than the imaging data, that essentially drives the matching process. In this section we present specific guidelines for picking these meta variables. To this end recall that the difference maps employed by CHAMP are high dimensional analogs of 'treatment effects' in the theory of matching estimators. Thus, we surmise that we may use this theory for guiding the selection of matching variables.

In line with this theory, we require that meta-variables used for matching be distributed independently of group membership. That is knowledge of the distribution of these meta-variables should not allow for prediction of group label. In general variables such as age,sex and ethnicity fit this criterion well. We have also required that meta-variables encode for a substantial degree of variation in the imaging data. For instance, age greatly influences anatomy as does gender. On the other hand socio-economic meta variables such as income

are unlikely to be associated with a similar degree of variation. Thus, matching based on age is likely to generate a better estimate of difference maps as compared to matching on incomes. A key exception to the use of age as a meta variable would be if one were trying to study heterogeneity in aging itself. In such a case one would need to use a different set of meta-variables. Possibly a combination of gender, ethnicity and heredity could achieve the required effect. More generally, the choice of meta variables should be a study specific decision and it must be made based on the specific aims of a given analysis. Thus, the investigator has a critical role in ensuring the appropriate choice.

4.6.3. Parameter selection for CHAMP analyses

CHAMP requires the user to select two parameters namely, r - the number of matches used to estimate the difference map and k - the number of clusters to generate.

Using a value of k which is very large may generate clusters containing small numbers of patients. If a cluster contains less than 10% of the data, it is likely that it consists mostly of outliers. Using $k = 1$ computes the mean of all the disease effects and yields p-value maps that are similar to VBM analyses. In the work presented here we have chosen 'k' values between 2 and 4. Using $k > 4$ yields clusters of low cardinality that are unlikely to yield significant nosological constructs. Based on the ANOVA results shown in figures 5,7 and 8; $k = 2, 3$ seem to yield cluster memberships correlated with the MMSE score. When k is increased to 4 two of the generated clusters present similar MMSE values and the overall significance of the ANOVA drops.

The number of matches r used to model the effect of a phenotypic difference is the other major parameter that must be selected before running CHAMP. Choosing a very small value of r carries the risk of generating noisy estimates of $\mathbf{x}^1_{j(i)}$. Hence cluster memberships generated using small values of r may not be reliable. On the other hand choosing, a very large value of r implies that $\mathbf{x}^1_{j(i)}$ is very close to the mean of group 1 itself. In such a setting

the clustering memberships are essentially based on the patient data rather than difference maps. It is advisable to stay in between these two extremes while applying CHAMP analysis. Results presented in figures 50 through 52 indicate that for values of r close to 10 CHAMP clusters differ most significantly in terms of MMSE and least significantly in terms of age. While there is no definitive trend at low values of r, at large values of r the ANOVA for age starts becoming significant. This is in line with our intuition that the clustering mimics clustering of raw patient data itself. Thus, we surmise that values of r close to 10 represent a balanced approach that is most effective in highlighting disease heterogeneity.

The guidelines presented above along with plots analogous to figures 10 to 13 yield an effective technique for parameter selection in a new data set. In the next section we conclude this chapter with a brief summary.

4.7. Conclusion

In summary, this chapter presents CHAMP, a method for exploration of heterogeneity in population neuroimaging studies. We have validated this method through experiments on simulated data and also demonstrated a potential use case of the method by analyzing the publicly available ADNI dataset. In the future, we hope that the neuroimaging community develops this method further and applies it to a diverse array of biomedical data.

CHAPTER 5

Conclusion

5.1. Summary

The aim of this dissertation was to address the problem of population wide neuroimaging analysis from a high dimensional machine learning centric perspective. Machine learning analyses enables the use of brain MR images to diagnose brain related disorders. However, these methods are often designed to act like black boxes. In this dissertation we have explored both supervised and unsupervised machine learning methods and developed techniques to connect the associated black box models with traditional interpretative statistics such as p-values.

Specifically, in chapters 2 and 3, we have developed a p-value based interpretation for support vector machine models. We have also presented applications of this framework to interpret diagnostic / classification models learnt using data from several large scale population studies. The availability of a p-value map in conjunction with a diagnostic SVM model allows non specialists to understand how the model works in a mathematically rigorous way. The availability of a p-value based inference mechanism is highly relevant, given that a majority of the clinical and research training relies on p-value based inference. We chose to focus on support vector machines because they are the most widely used supervised machine learning tool in the domain of medical image analysis.

The work presented in chapter 4, highlights the value of unsupervised machine learning to population wide neuroimaging analyses. Unsupervised methods have rarely if ever been used to directly quantify the difference between populations in neuroimaging. However, such analyses can be extremely valuable in the study of several psychiatric disorders. We have broached this topic using a simple framework in chapter 4. While the analyses presented in the chapter relies a simplistic methodological framework, it does present an imaging based approach to identify disease subtypes and understand heterogeneity in disease.

The following section presents oportunities for future methological work based on this dissertation. We also present some preliminary results associated with this work.

5.2. Discussion and future work

5.2.1. SVM for detecting abnormalities

The inference framework developed for the support vector machine contains elements that may be re-purposed for a slightly different problem in medical image analysis. Namely the problem of delineating abnormalities. To see how this may be done, recall that the SVM solves a convex optimization problem that uncovers a direction in hyperspace that differentiates two point clouds with maximal margin. What would happen if we were to force one of the point clouds to contain only one point? In such a case, the algorithm would compute a direction that maximally differentiates this point from the other point cloud (figure 55). The direction and the associated margin computed by the algorithm could potentially be used for abnormality detection and even abnormality segmentation. In the subsequent text, we explain how this could be achieved.

Let $i \in \{1, \cdots, n\}$ index a set of n normal controls. Also let $\mathbf{x}^i$ denote vectorized image

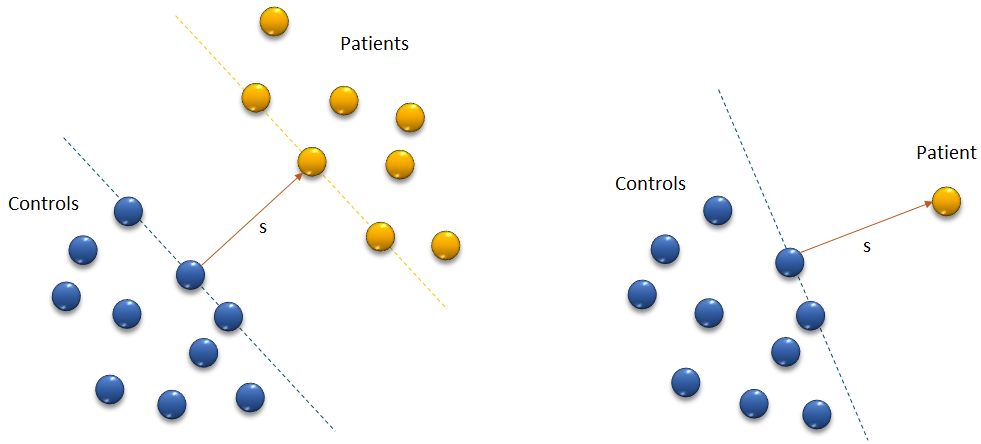


Figure 55: (Left) Typical support vector machine optimization with multiple control and patient samples (Right) support vector machine optimization with a single patient and many controls.

representations corresponding to these normal controls. In principle, one could apply the leave one out paradigm and train n support vector machines. Each SVM would generate a corresponding margin vector $\mathbf{s}^i$ quantifying how a specific control differs from the group. Now the collection of vectors $\mathbf{s}^i$ for $i \in \{1, \cdots, n\}$ quantifies how each control image differs from the rest of the group . We can use the theory presented in chapter 3 to estimate the distribution of the $\mathbf{s}^i$. If we let $\mathbf{X}$ to denote a matrix constituted as a stack of the $\mathbf{x}^i$'s, then, we would expect the components of $\mathbf{s}^i$ to be distributed normally according to equation (3.14). Given a vectorized image and a separating vector $\mathbf{s}^*$ corresponding to a new subject, abnormality detection requires us to answer two questions a) is this subject abnormal and b) what is the location of the abnormality.

The first of these questions may be answered by appealing to the distribution of the magnitude of $\mathbf{s}^j$. If the magnitude of $\mathbf{s}^*$ denoted by $||\mathbf{s}^*||_2$, is far larger than the magnitudes of the vectors $\mathbf{s}^i$, then it is likely that the new subject is not normal. On the other hand if the value of $||\mathbf{s}^*||_2$ is comparable to $||\mathbf{s}^i||_2$, $\forall i$ then it is likely that the new subject is in fact a control. Doing this would require us to further develop the analytic framework presented in chapter 3 with a focus on approximating the distribution of $||\mathbf{s}^i||_2$. This is

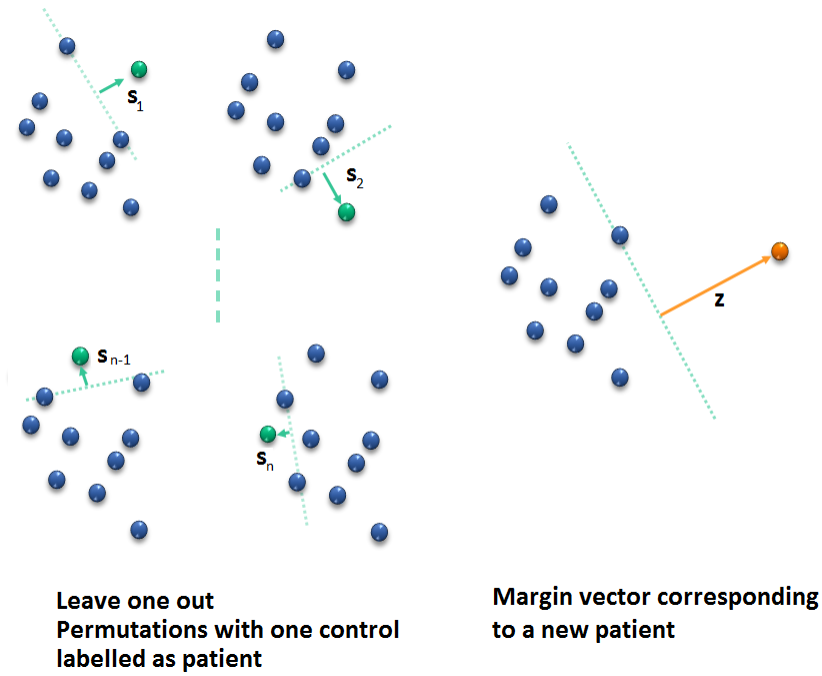


Figure 56: (Left) Illustration of proposed permutation procedure (Right) Expected margin vector corresponding to a patient image

a complex endeavor because the associated distribution is unlikely to be normal and also because it requires the estimation of the variance and expected value of a quadratic form with a specialized but unknown co-variance structure.

The second question, may be answered using the framework presented in chapter 3 itself. That is, given that a subject is 'abnormal', we may locate the abnormality by comparing components of $\mathbf{s}^*$ to the respective components of $\mathbf{s}^i$. Since, the analytic form of the distribution of these components has been worked out, we would expect this distribution to be normal. Specifically, following the logic presented in chapters 2 and 3, we would expect:

$$E(s_j) = 0 \tag{5.1}$$

and

$$var\left(\frac{w_j}{\mathbf{w^T w}}\right) \approx \frac{\sigma_j^2}{\left[\sum_{k=1}^{d} \sigma_k^2\right]^2} \tag{5.2}$$

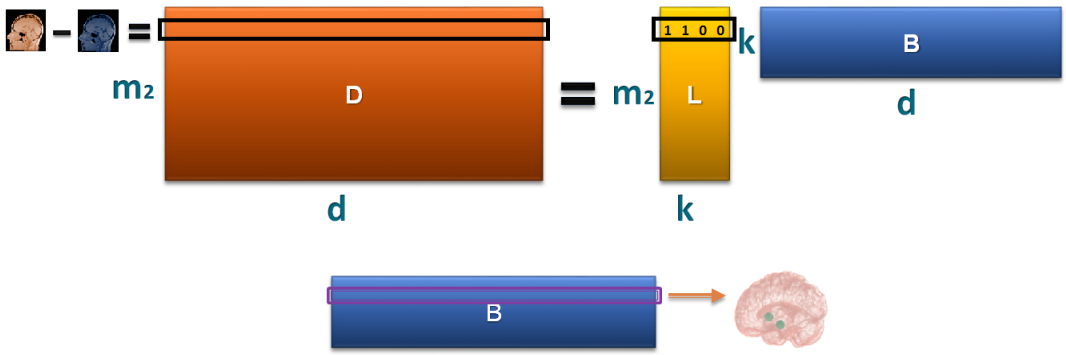


Figure 57: Illustration showing the application of matrix factorization to difference matrix data

with:

$$\sigma_j^2 = \frac{4(n-1)}{n^2} \sum_{i=1}^{m} C_{ij}^2 \tag{5.3}$$

with C_{ij} being the components of $\mathbf{C}$ given by (2.12).

In combination, the two questions presented above and their respective answers, yield a learning framework that could potentially turn the support vector machine based statistical analysis framework into a learning based abnormality detection and segmentation engine. Such an engine would mimic human radiologist training in that it could 'learn' a model of 'normalcy' from control images and use this model to identify what is abnormal. The prime advantage of such a mechanism over traditional segmentation techniques would be 1) the ability to learn from 'normals' 2) a p-value based interpretation of results 3) improved performance with an increased sample size of normal scans. We propose to develop this approach to maturity in future work.

5.2.2. Dictionary learning for heterogeneity analysis

Heterogeneity analysis presented in chapter 4 relies ultimately on consensus k-means clustering of a feature matrix to define patient clusters. We chose this approach because a)

k-means is one of the simplest and well understood clustering algorithms and b) it produces discrete cluster memberships for patients which can consequently be used to understand what each patient subgroup represents. However, this overlooks the fact that every patient may not neatly fall into one of the CHAMP 'subgroups' and it may in fact be expressing a pattern of deficit that is some combination of the patterns associated with the two different CHAMP clusters. In what follows we argue that the difference matrix representation used by CHAMP in conjunction with advanced matrix factorization algorithms may provide a powerful tool for understanding subject specific anatomical/functional effects of disease.

Recall that the difference matrix used by CHAMP is a representation of imaging group difference information which is distinct from traditional t-statistic/p-value maps generated by VBM (Ashburner and Friston, 2000) or SVM (Gaonkar and Davatzikos, 2013) analysis. This factor differentiates the method from other group based analysis methods in literature. The fact that CHAMP represents group differences as a matrix, opens up the possibility of applying dictionary learning/matrix factorization methods to population-wide difference analyses. Unlike k-means, such methods may be applied within the CHAMP paradigm to yield soft clustering. This allows each individual to be a part of multiple clusters. To understand this better first note that $\mathbf{D} \in \mathbf{R}^{m_2 \times d}$ where we stick to the notation from chapter 4 with m_2 as the number of patients and d as the number of voxels in the images. We may decompose this matrix as:

$$\mathbf{D} = \mathbf{LB} \tag{5.4}$$

where we will have $\mathbf{L} \in \mathbf{R}^{m_2 \times k}$ and $\mathbf{B} \in \mathbf{R}^{k \times d}$. Here k is a parameter akin to the number of clusters used in k-means. Rows of the matrix B are the dictionary learning equivalents of 'cluster centers'. This matrix is also known as the basis matrix is thought to uncover latent structure in the data. The matrix $\mathbf{L}$ is the loadings matrix. The loadings matrix defines how rows of B can combine to yield rows of D. We illustrate in figure 57 how matrix factorization might be used to uncover latent patterns that combine to manifest disease in a specific subject.

Literature on matrix factorization approaches is vast and we do not do not detail it here. However, we opine that sparse matrix factorization techniques should be most effective in the context of heterogeneity analysis. In summary, the use of matrix factorization in place of standard clustering presents the possibility that one could find a common set of base patterns that summarizes the effect of disease in a population and a combination of such base patterns could potentially delineate individual specific disease effects in a more concise manner than the existing framework.

A substantial amount of methodological and experimental work is necessary to develop the sparse matrix factorization idea into a mature approach. This is outside the scope of the current work, but nevertheless remains a promising avenue of future research.

5.2.3. Improved matching for heterogeneity analysis

The construction of the matrix $\mathbf{D}$ in chapter 4 uses simple and intuitive meta-data based nearest neighbor matching procedure. This procedure is simple and reveals interesting insights about the data. However, it does require a) an intelligent choice of meta variables and b) a careful selection of parameters before running the method. In the future both of these aspects may be improved upon.

5.2.3.1. Meta variable selection using machine learning

While a researcher still has to exercise his or her intuition while selecting meta variables, one may be able to guide the selection procedure by measuring how strongly each candidate meta-variable is associated with variability in the data. One possible approach would involve, learning a multivariate model to predict each meta-variable from imaging data. One would expect the imaging data to be more predictive of meta-variables which encode for a higher degree of variation. From this perspective, it is interesting to note that a support vector regression trained on the ADNI controls data does result in predictions that

are highly correlated with the actual age in cross-validation. This indirectly points to the fact that age might encode a large proportion of variance in the imaging data. Similar claims may be made for gender information. This is not necessarily true of say, a randomly generated meta-variable. Thus, machine learning techniques offer one possible approach to automate or at least guide the selection of meta-variables. Other possibilities include (but are not limited to) an aggregate measure of correlation between specific candidate meta-variables and voxel intensity values or the correlation between candidate meta-variables and the principal component loadings of the data. The apt choice of meta-variables is critical to the success of the CHAMP approach. Thus, it is a subject of paramount importance and should be a major thrust of future work.

5.2.3.2. Alternate matching schemes

Parameters used by the heterogeneity analysis in chapter 4 define how images get matched across groups. The matching scheme proposed in the chapter is simplistic and ultimately relies on a nearest neighbor matching. It is easy to imagine sophisticated versions of this matching framework. The simplest modification would be to use a weighted matching. In this case we would re-write equation (**??**) as:

$$\hat{\mathbf{x}}^1_{j(i)} = \sum_{l \in S_i} w_l \mathbf{x}^1_l. \tag{5.5}$$

where w_l would be appropriately chosen weights. The weights themselves may either be explicitly specified on the basis of the meta-data or some combination of meta-data and imaging. For instance, sticking to the notation in chapter 4, one could set w_l to:

$$w_l = e^{\gamma ||\mathbf{v}_i - \mathbf{v}_{j(i)}||^2} \tag{5.6}$$

where γ is an appropriately chosen constant. Such a weighting would estimate $\mathbf{x}^1_{j(i)}$ using a linear combination of all group 1 images with highest weights assigned to group 1 subjects

that are closest to $\mathbf{x}_i^2$ in terms of meta-data. This also opens up the possibility of learning w_l from some combination of data and meta-data, given a specific study population, and a specific target score with respect to which to explore heterogeneity. A learning based approach could possibly do away with parameter selection in CHAMP.

Another possibility for improving the matching procedure would be to devise a high dimensional analog of a point cloud matching algorithm for generating matches. Such an algorithm could potentially generalize to larger collections of meta-variables and even incorporate image information directly into the matching procedure. Each of these approaches require extensive experimental verification and consequent modifications to develop into mature heterogeneity analyses tools. Hence, it presents a promising avenue for future research.

5.2.4. Robustness to scanner and pre-processing variation

In this dissertation, we have presented results using several different datasets. Each of these datasets was acquired using a scanning protocol unique to it. Image preprocessing pipelines used also differed between datasets. We suspect that these variations should ultimately lead to slight variation in the results of the analysis. As such, this type of variation affects not only the methods presented here but also other well established population based analysis methods in literature. Addressing the variation in results due to variation in scanner types/preprocessing protocols is a wide and complex area of research that has not been broached in this dissertation. However, we have pointed this out here since we believe that robustness to such variation is an important aspect which needs to be addressed in the future.

5.2.5. Extension to other types of data

All the experiments and results presented in this dissertation are focussed on neuroimaging data. However, high dimension low sample size data appear in several other fields of scientific inquiry as well. For instance assays measuring the difference in expression of thousands of human genes between a few patients and controls constitute high dimension low sample size data in genomics. High dimension low sample size datasets may also occur in the field of natural language processing where they might be used for sentiment analysis. The methods developed here may be extended and applied in genomics as well as in language processing. In general they may be applied to any other field of scientific inquiry as long as the base assumptions underlying these methods can be adapted to the specific application.

5.3. Conclusion

In conclusion, we have presented a high dimensional perspective on population based neuroimaging analysis and enumerated several avenues along which our work may be developed further.

APPENDIX A

Appendix

A.1. Medical imaging modalities

T1 weighted magnetic resonance imaging (T1): T1 weighted scans are one of the basic pulse sequences in MRI. Magnetic resonance imaging or (MRI) is ultimately based on the physical phenomenon called Nuclear magnetic resonance (NMR). Each atomic nucleus is associated with a quantized set of spin quantum numbers. For instance the hydrogen nuclei which typically drive the signal in MR imaging can exist in two possible spin states (m=1/2 and m=-1/2). Normally, these states have the same energy level, that is , they are degenerate. However, placing them in a strong external magnetic field flips this balance and aligns the nuclei in one state or the other resulting in a net magnetization vector aligned with the applied external field. If an electromagnetic pulse corresponding to the energy difference between the two quantum states is now applied to the tissue, some of the protons flip causing the net magnetization vector to diverge from the external magnetic field. In the time after the application of the pulse the magnetization in the direction of the external field recovers to its original strength. Different tissues require different amounts of time to complete this recovery. These differences are the source of the T1-weighted MR image signal. In general tissue containing a higher amount of water, and thus more hydrogen nuclei, is darker on a T1-weighted image. Fatty tissue on the other hand is brighter in this

modality. Consequently, myelin sheaths that cover neuronal axons appear brighter and the cell body of the neuron appears darker.

T2 weighted magnetic resonance imaging (T2) While T1 relies on the decay of longitudinal magnetization, that is, the component of the magnetization that is aligned with the external field, a T2 weighted MR image relies on the decay of transverse magnetization. Transverse magnetization refers to the component of the net magnetization vector that is perpendicular to the external magnetic field. The pulse sequences used to acquire T2 images differ from those used to acquire T1 images in that they use longer relaxation and excitation times. Fatty tissue appears darker of T2 relative to tissue containing a higher amount of water. A sample T2 image is shown in figure 58.

Proton density images (PD) Another commonly used radiologic imaging modality is the proton density or PD image. This image is produced by controlling the selection of scan parameters to minimize the effects of T1 and T2, resulting in an image dependent primarily on the density of protons in the imaging volume. This image is essentially a quantitative summary of the number of protons per unit tissue. The higher the number of protons in a given unit of tissue,the brighter the signal on the proton density contrast image. Conversely the lower the number of protons in a given unit of tissue, the darker the signal on the proton density image.

Fluid Attenuated Inversion Recovery (FLAIR)

(FLAIR) Fluid attenuation inversion recovery is a special inversion recovery sequence used to remove the effects of fluid from the resulting images. For example, it can be used in brain imaging to suppress signals from cerebrospinal fluid (CSF) in the image, so as to bring out hyperintense lesions, such as multiple sclerosis (MS) plaques.

Blood oxygen level dependent functional magnetic resonance imaging (BOLD-fMRI) Neuronal firing in the brain is necessary for any processing task. The energy required by the neurons for performing these tasks is delivered to them in the form of

oxyhemoglobin. Neuronal processing involves oxygen exchange resulting in the formation of deoxyhemoglobin. The degree of magnetization in response to an applied external field differs between oxyhemoglobin and deoxyhemoglobin. This can be used as a source of signal in MRI. The measurement of this signal is the source of BOLD fMRI.

Diffusion tensor imaging Diffusion tensor imaging is a method that allows the mapping of the diffusion process of water molecules in vivo. Molecular diffusion of water in brain tissue is constrained by neuronal axons that compose white matter fibers. In general, water diffuses more rapidly along white matter fibers and less rapidly in a direction perpendicular to them. Therefore, diffusion patterns can reveal white matter fiber structure of the brain. Diffusion tensor imaging (DTI) is an imaging modality designed to capture this information. Advanced tractography algorithms can utilize the information contained in these DTI images to map white matter tracts in the human brain.

Positron emission tomography (PET) This is an imaging technique based on in vivo quantification of radioactivity. PET imaging relies on the introduction of a radioactive tracer based on a biologically active molecule into the body. A prime example of such a compound is 2-fluoro-2-deoxy-D-glucose (FDG). These compounds are a)differentially accumulated by specific tissue types and b)they radioactively decay inside the body to produce positrons. These positrons annihilate electrons inside the body to produce two gamma ray photons travelling in opposite directions that are detected by the scanner. The PET imaging system then reconstructs a 3D image of tracer concentration within the body.

Figure 58 shows representative slices from brain images captured using each of the previously described modalities.

Perfusion imaging Perfusion imaging measures parameters related to the passage of blood through blood vessels into the brain. These parameters may measure regional cerebral blood flow, blood volume or oxygenation. In neuroimaging perfusion studies may be done either with the aid of an external contrast inducing agent such as a gadolinium based agent or with

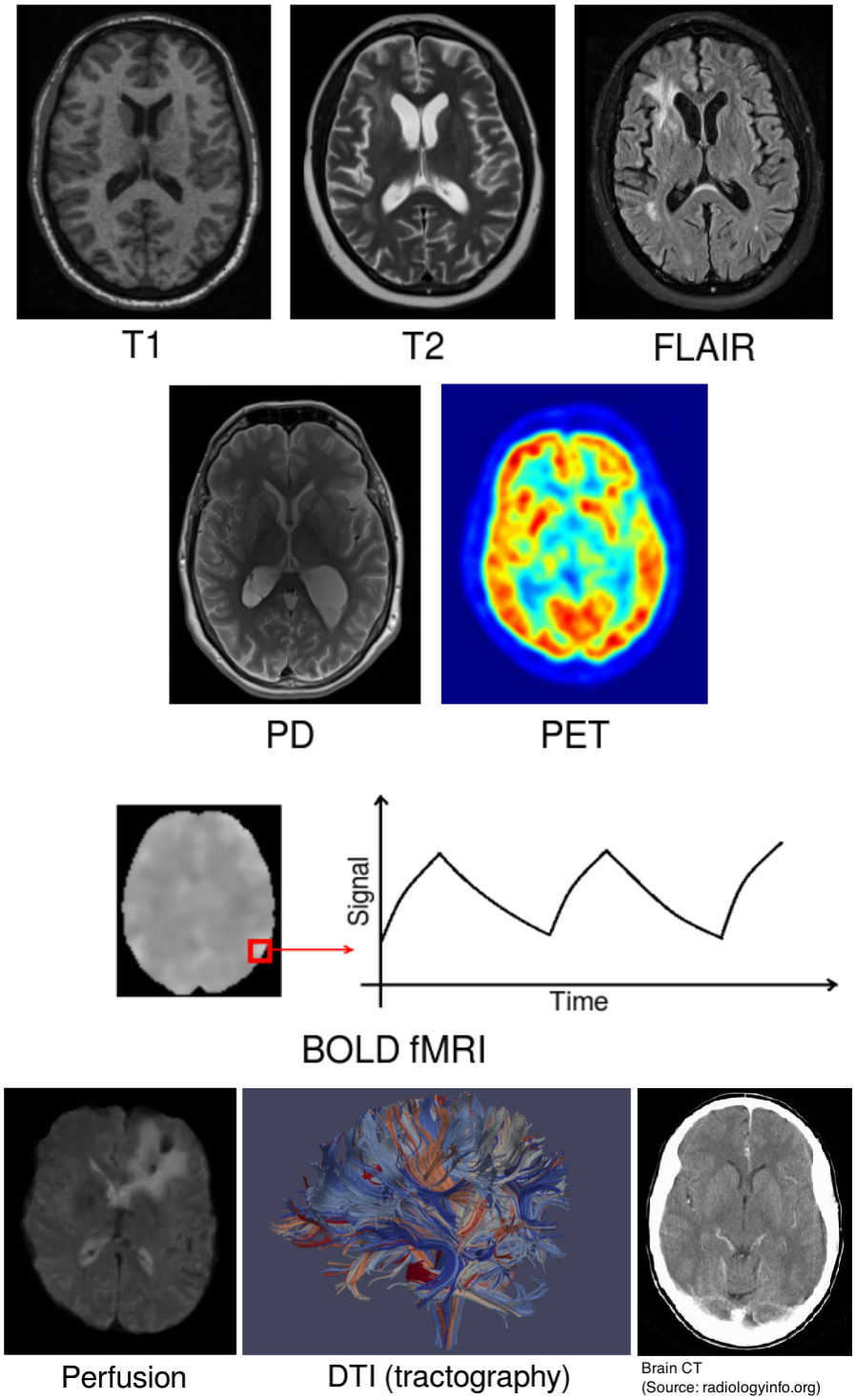


Figure 58: Illustrative slices of brain images captured using different imaging modalities.

the use of an intrinsic signal inducing phenomenon such as arterial spin labelling. Perfusion imaging of the brain allows us to study blood flow and is thus a promising tool in assessing disorders related to changes in blood flow and vascularization. It has shown promise in the study of stroke, brain tumors, and also in the study of neurodegenerative diseases.

CT scans A computed tomography or CT scan re-constructs the a 3D image of an organ using a series of radial sections obtained using X-ray imaging. Thus, CT produces a volume image corresponding to a various anatomical structures in the body based on their respective ability to block X-rays. Naturally, CT images are most useful for studying skeletal structures in the body. A brain slice computed using CT is shown in figure 58.

A.2. Image preprocessing techologies

A.2.1. Image enhancement

Often the first step in any image analysis pipeline involves improving the image quality using tools for image denoising, bias field correction or histogram equalization. As such there exists a substantial amount of literature on how best each of these steps may be done. For instance image de-noising often involves smoothing using a specific type of image filtering algorithm. A widely applied example of such an algorithm is presented in (**?**). Similarly, N3 bias correction (Sled et al., 1998) is an algorithm that changes voxel intensities to sharpen tissue peaks. It is commonly used for correcting inhomogeneities for MR images acquired on account of surface coils (see figure 60). Standard histogram equalization algorithms are included in toolkits like uch as ITK , MIPAV (McAuliffe et al., 2001) and FSL (Jenkinson et al., 2012) as well.

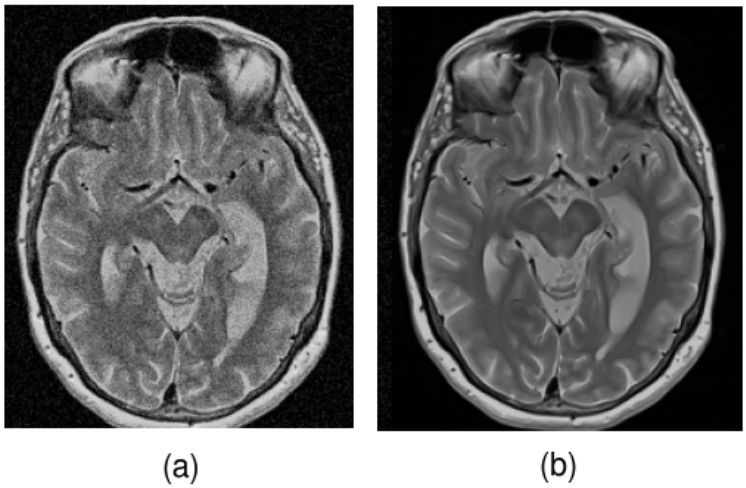


Figure 59: (a) Raw image (b) Image denoised using Gaussian smoothing

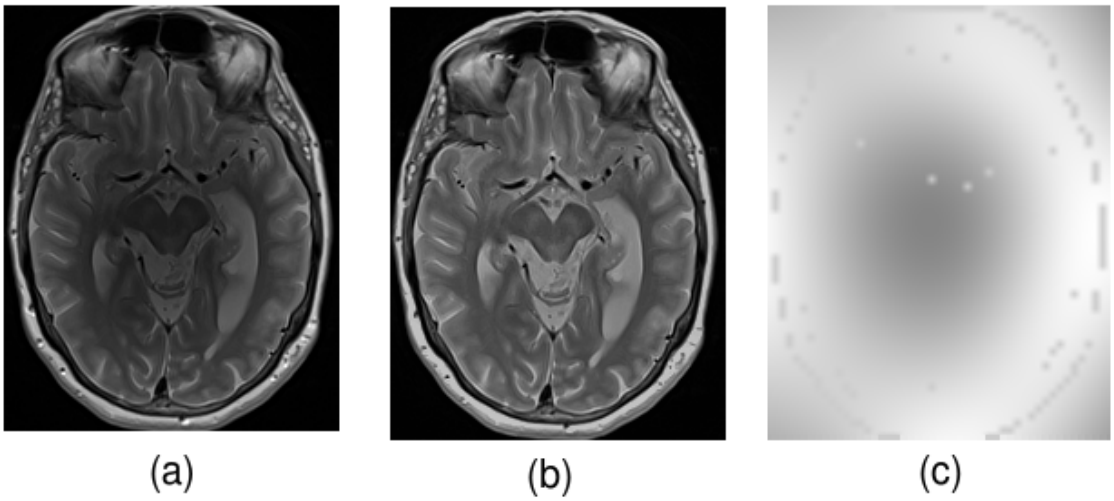


Figure 60: (a) PD image with bias (b) PD image after bias correction with n3 (c) bias field detected by n3

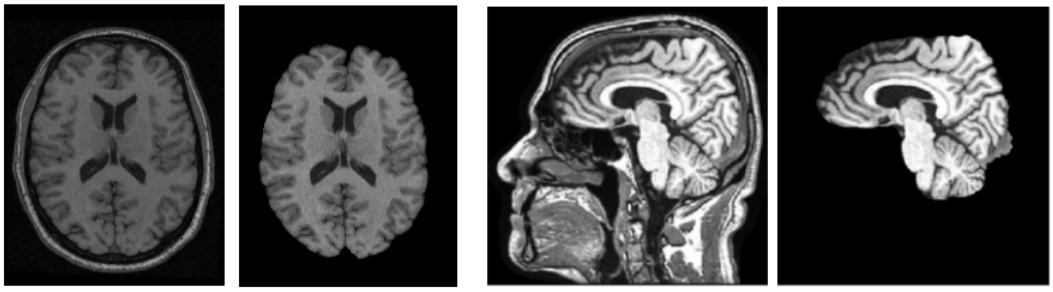


Figure 61: Illustrative example of the process of skull removal/ brain extraction

A.2.2. Brain extraction

Brain extraction, or skull stripping is a necessary preprocessing step in most neuroimaging analysis pipelines. It consists of the removal of the skull and the extracerebral tissues (e.g., scalp and dura) on brain magnetic resonance (MR) images. An illustrative example of extraction on a T1-weighted image is shown in figure 61. Several alternative methods for skull stripping are available in literature. One of the most popular approaches is the BET algorithm implemented as a part of FSL (Smith, 2002). Recent developments in skull stripping favor the use of learning based multi atlas techniques (Doshi et al., 2013; Lötjönen et al., 2010) and learning based methods(Iglesias et al., 2011). The data analysis presented throughout this thesis uses one of these methods, namely the weighted multi atlas skull stripping (MASS Doshi et al. (2013)) algorithm for brain extraction. To extract a brain from a given target image, the algorithm first registers a pre-selected set of template brain images to the target. Each of these template brain images are pre-segmented into brain and skull manually. A weighted label fusion technique is then used to extract the brain from the target image. All data processed through this pipeline is manually checked and corrected for errors if any. After, skull stripping and image enhancement, we obtain a set of relatively clean brain images. However, in order to compare images across a population, specialized image processing techniques are needed. Image segmentation and image registration are the two fundamental concepts that are required to be understood from the group analysis perspective. These are described next.

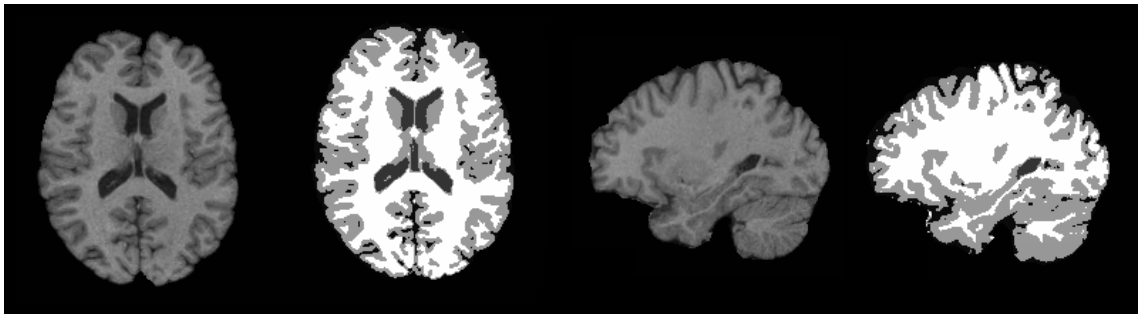


Figure 62: Illustrative example tissue segmentation of MR images into grey matter, white matter and cerebrospinal fluid

A.2.3. Image segmentation

In general, image segmentation is the partitioning of an image into multiple sets of pixels (segments), such that the resulting representation is more meaningful than the original image. The problem of automatic image segmentation has received tremendous attention from the medical imaging community in the past two decades. Specifically, in the case of structural brain images, it is used in the context of dividing the image into different tissue types such as grey matter, white matter and cerebrospinal fluids. There exist several different segmentation methods that address the task of brain image segmentation. A detailed review of segmentation methods is unnecessary for the purposes of this thesis. The adaptive k-means algorithm (Pham and Prince, 1999) and the multiplicative intrinsic component optimization (MICO) algorithm (Li et al., 2014) are geared towards segmenting grey matter, white matter and cerebrospinal fluids from adequately preprocessed brain images . Each of these algorithms perform segmentation by estimating the probability that a given voxel belongs to a specific tissue type. Sample segmentations of grey and white matter in the T1-MR image obtained using adaptive k-means are shown by figure 62.

A.2.4. Image registration

Image registration involves warping different images into a single template co-ordinate system so that a voxel by voxel comparison between the different images is meaningful and interpretable. The concept is illustrated in figure 63. The immense anatomical variation in human brains makes brain image registration a particularly challenging problem. For a comprehensive review of the literature of image registration techniques we refer the reader to a comprehensive review by (Sotiras et al., 2013). Typically, brain image registration is a two step procedure. Brain images are first transformed into a standardized template space via a global affine transformation or linear registration. This is followed by non linear deformations to match the intricate anatomical details between subject and template, also known as non linear registration. Mapping of a group of subjects to a template in this manner is an essential element of any large scale neuroimaging study. Without a common co-ordinate system it would be impossible to quantify the differences and similarities between different subjects. The template space to which the mapping is made is usually called an atlas. The atlas, for a particular study may be the brain image of a single subject chosen from the study, a population average or an independently annotated external image. One of the oldest such atlases is the Talairach atlas (Talairach and Tournoux, 1988). Another commonly used template is the population average atlas supplied by the Montreal Neurological Institute (MNI). We used an in house template for the analysis presented in this work. The area of brain image registration has been a particularly active topic of research in the medical imaging analysis community. Consequently there are several methods starting from intensity based registration (Collins et al., 1994; Collins and Evans, 1999; Ashburner et al., 1997), physical model based image registration (Christensen et al., 1997; Davatzikos, 1997; Gee et al., 1993; Toga, 1998), diffeomorphic frameworks (Vercauteren et al., 2009; Avants et al., 2008; **?**) and the more recent feature vector based registration (Shen and Davatzikos, 2002; Ou et al., 2011). The experiments presented here, use methods presented in (Shen and Davatzikos, 2002; Ou et al., 2011).

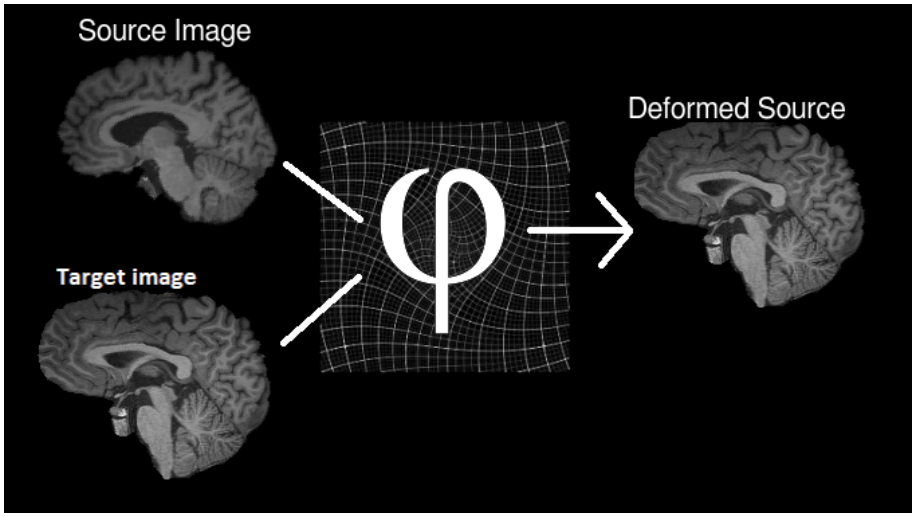


Figure 63: Concept of image registration

A.2.5. Feature extraction: Tissue density map generation

The registration process described above can yield one deformation field per image in a population based study. If registration was perfect the deformation field associated with each subject encodes all the information about the subject anatomy and the registered subject looks exactly the same as the template image. However, registration is rarely if ever perfect and the information about subject anatomy encoded in the deformation field alone is at best partial. Nevertheless, different ways of comparing deformation fields have been devised in literature. We will go through some of these measures in the literature review presented in the next section. However, in this section we present the regional analysis of volumes examined in normalized space (RAVENS) approach, that we use to compare anatomical information from different subjects. Briefly, for each tissue in the subject brain (grey matter/white matter/ cerebrospinal fluid) the RAVENS approach generates a tissue density map (see figure 64) in the template space. To do this the RAVENS analysis associates a counter with each location in template space. This counter increments every time a voxel associated with a specific tissue type in subject space is mapped to the location of the counter. Ultimately this leads to a tissue density map for each tissue for each subject.

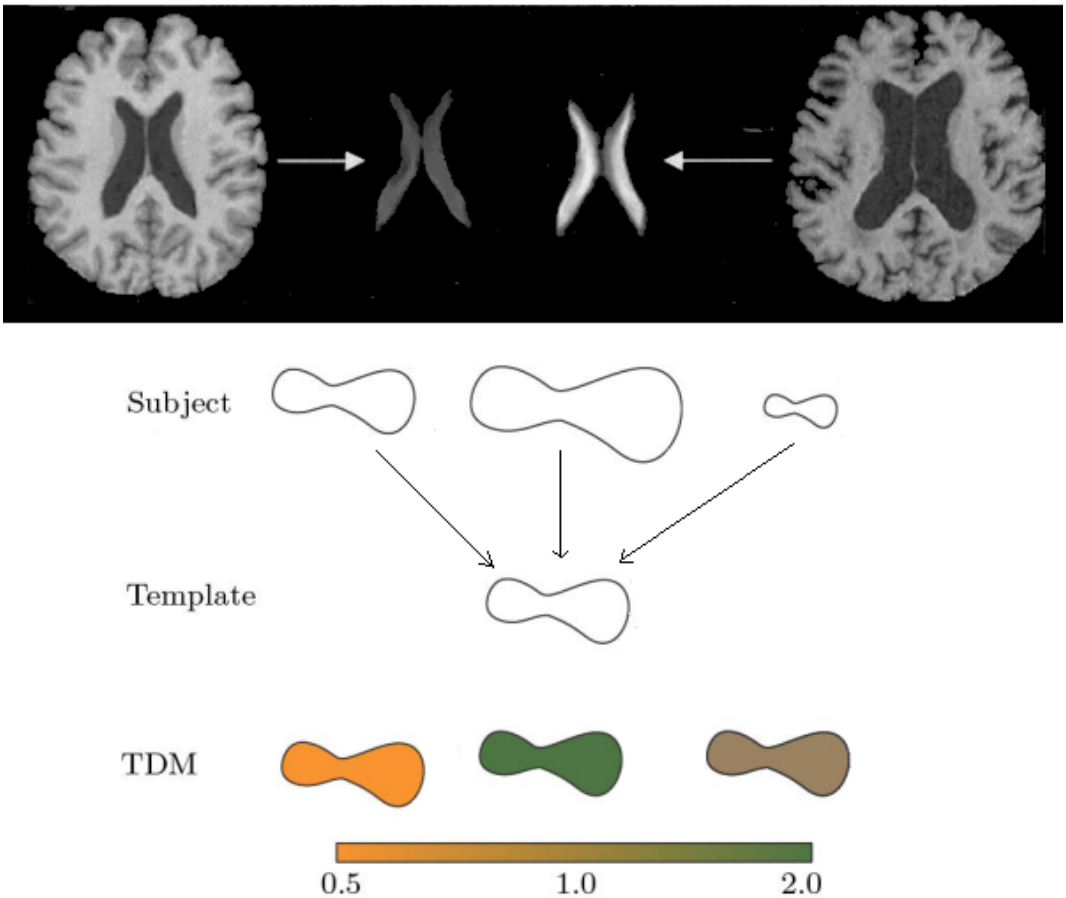


Figure 64: Illustrative example of tissue density maps

This allows the registration procedure to be utilized for comparing shape and volume in a principled manner across a population of images, by simply warping them to a common template space. The RAVENS approach is volume preserving and encodes shape differences in subject space as density differences in template space. Since, the map is volume preserving, every voxel in the subject space is mapped to some voxel in the template space. Thus, RAVENS maps immediately allow us to establish location correspondence between different subjects , which translates to feature correspondence in a machine learning setting. RAVENS maps associated with structural data have been widely used in experiments and analysis presented in this thesis. Hence, we have introduced the concept before the literature review. We refer the reader to original literature (Davatzikos, 1998; Davatzikos et al., 2001) for additional details.

BIBLIOGRAPHY

D. G. Amaral, C. M. Schumann, and C. W. Nordahl. Neuroanatomy of autism. *Trends in neurosciences*, 31(3):137–145, 2008.

N. C. Andreasen and S. Olsen. Negative v positive schizophrenia: definition and validation. *Archives of General Psychiatry*, 39(7):789, 1982.

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

J. Ashburner and K. J. Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

J. Ashburner, P. Neelin, D. Collins, A. Evans, and K. Friston. Incorporating prior knowledge into image registration. *Neuroimage*, 6(4):344–352, 1997.

A. P. Association, A. P. Association, et al. *Diagnostic and statistical manual-text revision (DSM-IV-TRim, 2000)*. American Psychiatric Association, 2000.

A. P. Association et al. *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*. bookpointUS, 2013.

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

R. Barber, P. Scheltens, A. Gholkar, C. Ballard, I. McKeith, P. Ince, R. Perry, and J. O'brien. White matter lesions on magnetic resonance imaging in dementia with lewy bodies, alzheimer's disease, vascular dementia, and normal aging. *Journal of Neurology, Neurosurgery & Psychiatry*, 67(1):66–72, 1999.

N. K. Batmanghelich, B. Taskar, and C. Davatzikos. Generative-discriminative basis learning for medical imaging. *Medical Imaging, IEEE Transactions on*, 31(1):51–69, 2012.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, Oct. 2007. ISBN 0387310738.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

M. A. Butters, O. L. Lopez, and J. T. Becker. Focal temporal lobe dysfunction in probable alzheimer's disease predicts a slow rate of cognitive decline. *Neurology*, 46(3):687–692, 1996.

V. Calhoun, T. Adali, V. McGinty, J. Pekar, T. Watson, and G. Pearlson. fmri activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage*, 14(5):1080–1088, 2001.

J. Cao, K. J. Worsley, et al. The detection of local shape changes via the geometry of hotelling's t^2 fields. *The Annals of Statistics*, 27(3):925–942, 1999.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformation of brain anatomy. *Medical Imaging, IEEE Transactions on*, 16(6):864–877, 1997.

M. Chung, K. Worsley, T. Paus, C. Cherif, D. Collins, J. Giedd, J. Rapoport, and A. Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3): 595–606, 2001.

M. K. Chung. *Computational Neuroanatomy: The Methods*. World Scientific, 2013.

R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, 2008.

D. Collins and A. Evans. Animal: Automatic nonlinear image matching and anatomical labeling. *Brain warping*, pages 133–142, 1999.

D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *Journal of computer assisted tomography*, 18(2):192–205, 1994.

R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot. 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.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

R. Das. A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572, 2010.

C. Davatzikos. Spatial transformation and registration of brain images using elastically deformable models. *Computer Vision and Image Understanding*, 66(2):207–222, 1997.

C. Davatzikos. Mapping image data to stereotaxic spaces: applications to brain mapping. *Human Brain Mapping*, 6(5-6):334–338, 1998.

C. Davatzikos. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *Neuroimage*, 23(1):17–20, 2004.

C. Davatzikos and R. N. Bryan. Using a deformable surface model to obtain a shape representation of the cortex. *Medical Imaging, IEEE Transactions on*, 15(6):785–795, 1996.

C. Davatzikos, M. Vaillant, S. M. Resnick, J. L. Prince, S. Letovsky, and R. N. Bryan. A computerized approach for morphological analysis of the corpus callosum. *Journal of computer assisted tomography*, 20(1):88–97, 1996.

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, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–668, 2005.

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, 2008.

C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski. Prediction of mci to ad conversion, via mri, csf biomarkers, and pattern classification. *Neurobiology of aging*, 32(12):2322–e19, 2011.

B. C. Dickerson, D. A. Wolk, et al. Dysexecutive versus amnesic phenotypes of very mild alzheimer's disease are associated with distinct clinical, genetic and cortical thinning characteristics. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(1):45–51, 2011.

N. U. Dosenbach, B. L. Schlaggar, et al. Distinct neural signatures detected for adhd subtypes after controlling for micro-movements in resting state functional connectivity mri data. 2013.

J. Doshi, G. Erus, Y. Ou, B. Gaonkar, and C. Davatzikos. Multi-atlas skull-stripping. *Academic radiology*, 20(12):1566–1576, 2013.

S. Durston. A review of the biological bases of adhd: what have we learned from imaging studies? *Mental retardation and developmental disabilities research reviews*, 9(3):184–195, 2003.

C. Ecker, V. Rocha-Rego, P. Johnston, J. Mourao-Miranda, A. Marquand, E. M. Daly, M. J. Brammer, C. Murphy, and D. G. Murphy. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. *Neuroimage*, 49 (1):44–56, 2010.

J. A. Etzel, J. M. Zacks, and T. S. Braver. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos. Compare: classification of morphological patterns using adaptive regional elements. *Medical Imaging, IEEE Transactions on*, 26(1):93–105, 2007.

S. O. R. Fedkiw. Level set methods and dynamic implicit surfaces. 2003.

R. Filipovych, B. Gaonkar, and C. Davatzikos. A composite multivariate polygenic and neuroimaging score for prediction of conversion to alzheimer's disease. In *Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on*, pages 105–108. IEEE, 2012.

M. B. First, R. L. Spitzer, M. Gibbon, and J. B. Williams. *Structured Clinical Interview for DSM-IV® Axis I Disorders (SCID-I), Clinician Version, Administration Booklet.* American Psychiatric Pub, 2012.

M. F. Folstein, S. E. Folstein, and P. R. McHugh. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

A. Fuchs, A. Joffe, and J. Teugels. Expectation of the ratio of the sum of squares to the square of the sum: exact and asymptotic results. *Theory of Probability & Its Applications*, 46(2):243–255, 2002.

B. Gaonkar and C. Davatzikos. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *NeuroImage*, 78:270–283, 2013.

B. Gaonkar, K. Pohl, and C. Davatzikos. Pattern based morphometry. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 459–466. Springer, 2011.

C. Gaser, H.-P. Volz, S. Kiebel, S. Riehemann, and H. Sauer. Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research. *Neuroimage*, 10(2):107–113, 1999.

Y. Ge, R. I. Grossman, J. S. Babb, M. L. Rabin, L. J. Mannon, and D. L. Kolson. Age-related total gray matter and white matter changes in normal adult brain. part i: volumetric mr imaging analysis. *American journal of neuroradiology*, 23(8):1327–1333, 2002.

J. C. Gee, M. Reivich, and R. Bajcsy. Elastically deforming 3d atlas to match anatomical brain images. *Journal of computer assisted tomography*, 17(2):225–236, 1993.

J. N. Giedd, J. W. Snell, N. Lange, J. C. Rajapakse, B. Casey, P. L. Kozuch, A. C. Vaituzis, Y. C. Vauss, S. D. Hamburger, D. Kaysen, et al. Quantitative magnetic resonance imaging of human brain development: ages 4–18. *Cerebral cortex*, 6(4):551–559, 1996.

T. Gómez-Isla, J. L. Price, D. W. McKeel Jr, J. C. Morris, J. H. Growdon, and B. T. Hyman. Profound loss of layer ii entorhinal cortex neurons occurs in very mild alzheimer's disease. *The Journal of neuroscience*, 16(14):4491–4500, 1996.

P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, June 2005. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2005.00510.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2005.00510.x`.

B.-C. Ho, N. C. Andreasen, S. Ziebell, R. Pierson, and V. Magnotta. Long-term antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia. *Archives of general psychiatry*, 68(2):128–137, 2011.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

L. Ibanez, W. Schroeder, L. Ng, and J. Cates. The itk software guide. 2003.

J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu. Robust brain extraction across datasets and comparison with publicly available methods. *Medical Imaging, IEEE Transactions on*, 30(9):1617–1634, 2011.

M. Ingalhalikar, J. Yang, C. Davatzikos, and R. Verma. Dti-droid: Diffusion tensor imaging-deformable registration using orientation and intensity descriptors. *International Journal of Imaging Systems and Technology*, 20(2):99–107, 2010.

C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

H. Jiang, P. van Zijl, J. Kim, G. D. Pearlson, and S. Mori. Dtistudio: resource program for diffusion tensor computation and fiber bundle tracking. *Computer methods and programs in biomedicine*, 81(2):106–116, 2006.

Y. Jiao, R. Chen, X. Ke, K. Chu, Z. Lu, and E. H. Herskovits. Predictive models of autism spectrum disorder based on brain regional cortical thickness. *Neuroimage*, 50(2):589–599, 2010.

G. Karas, P. Scheltens, S. Rombouts, R. van Schijndel, M. Klein, B. Jones, W. van der Flier, H. Vrenken, and F. Barkhof. Precuneus atrophy in early-onset alzheimer's disease: a morphometric structural mri study. *Neuroradiology*, 49(12):967–976, 2007.

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

C. Kawas, S. Resnick, A. Morrison, R. Brookmeyer, M. Corrada, A. Zonderman, C. Bacal, D. D. Lingle, and E. Metter. A prospective study of estrogen replacement therapy and the risk of developing alzheimer's disease the baltimore longitudinal study of aging. *Neurology*, 48(6):1517–1521, 1997.

Y. Kawasaki, M. Suzuki, F. Kherif, T. Takahashi, S.-Y. Zhou, K. Nakamura, M. Matsui, T. Sumiyoshi, H. Seto, and M. Kurachi. Multivariate voxel-based morphometry successfully differentiates schizophrenia patients from healthy controls. *Neuroimage*, 34(1): 235–242, 2007.

S. Klöppel, B. Draganski, C. V. Golding, C. Chu, Z. Nagy, P. A. Cook, S. L. Hicks, C. Kennard, D. C. Alexander, G. J. Parker, et al. White matter connections reflect changes in voluntary-guided saccades in pre-symptomatic huntington's disease. *Brain*, 131(1): 196–204, 2008a.

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

N. Koutsouleris, E. M. Meisenzahl, C. Davatzikos, R. Bottlender, T. Frodl, J. Scheuerecker, G. Schmitt, T. Zetzsche, P. Decker, M. Reiser, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of general psychiatry*, 66(7):700–712, 2009.

J. H. Kramer and B. L. Miller. Alzheimer's disease and its focal variants. In *Seminars in neurology*, volume 20, pages 447–454. Copyright© 2000 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662, 2000.

S. A. Ladoucette and J. L. Teugels. Asymptotics for ratios with applications to reinsurance. *Methodology and Computing in Applied Probability*, 9(2):225–242, 2007.

R. K. Lenroot and P. K. Yeung. Heterogeneity within autism spectrum disorders: what have we learned from neuroimaging studies? *Frontiers in human neuroscience*, 7, 2013.

C. Li, J. C. Gore, and C. Davatzikos. Multiplicative intrinsic component optimization (mico) for mri bias field estimation and tissue segmentation. *Magnetic Resonance Imaging*, 2014.

J. M. Lötjönen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage*, 49(3):2352–2365, 2010.

D. MacDonald, N. Kabani, D. Avis, and A. C. Evans. Automated 3-d extraction of inner and outer surfaces of cerebral cortex from mri. *NeuroImage*, 12(3):340–356, 2000.

J. L. Marroquín, B. C. Vemuri, S. Botello, E. Calderon, and A. Fernandez-Bouzas. An accurate and efficient bayesian method for automatic segmentation of brain mri. *Medical Imaging, IEEE Transactions on*, 21(8):934–945, 2002.

M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus. Medical image processing, analysis and visualization in clinical research. In *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*, pages 381–386. IEEE, 2001.

M. J. McKeown. Detection of consistently task-related activations in fmri data with hybrid independent component analysis. *NeuroImage*, 11(1):24–35, 2000.

M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fmri data by blind separation into independent spatial components. Technical report, DTIC Document, 1997.

D. McLeish and G. O'Brien. The expected ratio of the sum of squares to the square of the sum. *The Annals of Probability*, pages 1019–1028, 1982.

T. Nikolcheva, S. Jäger, T. A. Bush, and G. Vargas. Challenges in the development of companion diagnostics for neuropsychiatric disorders. 2011.

Y. Noh, S. Jeon, J. M. Lee, S. W. Seo, G. H. Kim, H. Cho, B. S. Ye, C. W. Yoon, H. J. Kim, J. Chin, et al. Anatomical heterogeneity of alzheimer disease based on cortical thickness on mris. *Neurology*, 83(21):1936–1944, 2014.

Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4): 622–639, 2011.

V. C. Pangman, J. Sloan, and L. Guse. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Applied Nursing Research*, 13(4):209–213, 2000.

D. L. Pham and J. L. Prince. An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensity inhomogeneities. *Pattern Recognition Letters*, 20(1):57–68, 1999.

G. Rabinovici, W. Seeley, E. Kim, M. Gorno-Tempini, K. Rascovsky, T. Pagliaro, S. Allison, C. Halabi, J. Kramer, J. Johnson, et al. Distinct mri atrophy patterns in autopsy-proven alzheimer's disease and frontotemporal lobar degeneration. *American journal of Alzheimer's disease and other dementias*, 22(6):474–488, 2008.

P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, 55(3):1120–1131, 2011.

S. M. Resnick, C. Davatzikos, M. A. Kraut, and A. B. Zonderman. Longitudinal changes in mri volumes in older adults. *Neuroimage*, 11(5):S153, 2000.

S. M. Resnick, D. L. Pham, M. A. Kraut, A. B. Zonderman, and C. Davatzikos. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *The Journal of Neuroscience*, 23(8):3295–3301, 2003.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 2005.

S. Ryali, T. Chen, K. Supekar, and V. Menon. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *Neuroimage*, 59(4):3852–3861, 2012.

M. R. Sabuncu and K. Van Leemput. The relevance voxel machine (rvoxm): a bayesian method for image-based prediction. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 99–106. Springer, 2011.

T. D. Satterthwaite, M. A. Elliott, K. Ruparel, J. Loughead, K. Prabhakaran, M. E. Calkins, R. Hopson, C. Jackson, J. Keefe, M. Riley, et al. Neuroimaging of the philadelphia neurodevelopmental cohort. *Neuroimage*, 86:544–553, 2014.

I. N. C. G. H. Sauer. Heterogeneity of brain structural variation and the structural imaging endophenotypes in schizophrenia. *Neuropsychobiology*, 66:44–49, 2012.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

S. R. Searle. *Linear models*. John Wiley & Sons, 2012.

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

A. Shiino, T. Watanabe, K. Maeda, E. Kotani, I. Akiguchi, and M. Matsuda. Four subgroups of alzheimer's disease based on patterns of atrophy using vbm and a unique pattern for early onset disease. *Neuroimage*, 33(1):17–26, 2006.

J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *Medical Imaging, IEEE Transactions on*, 17(1):87–97, 1998.

S. M. Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *Medical Imaging, IEEE Transactions on*, 32(7):1153–1190, 2013.

A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, 9(1):319, 2008.

R. G. Steen, C. Mull, R. Mcclure, R. M. Hamer, and J. A. Lieberman. Brain volume in first-episode schizophrenia systematic review and meta-analysis of magnetic resonance imaging studies. *The British Journal of Psychiatry*, 188(6):510–518, 2006.

D. Sun, T. G. van Erp, P. M. Thompson, C. E. Bearden, M. Daley, L. Kushan, M. E. Hardt, K. H. Nuechterlein, A. W. Toga, and T. D. Cannon. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological psychiatry*, 66 (11):1055–1060, 2009.

J. A. K. Suykens and J. Vandewalle. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, 9(3):293–300, June 1999. ISSN 13704621. doi: 10.1023/A:1018628609742. URL `http://dx.doi.org/10.1023/A:1018628609742`.

J. Talairach and P. Tournoux. *Co-Planar Stereotaxic Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging (Thieme Classics)*. Thieme, Jan. 1988. ISBN 0865772932. URL `http://www.worldcat.org/isbn/0865772932`.

P. Thompson and A. W. Toga. Anatomically driven strategies for high-dimensional brain image warping and pathology detection. *Brain warping*, pages 311–336, 1998.

P. M. Thompson, J. L. Stein, S. E. Medland, D. P. Hibar, A. A. Vasquez, M. E. Renteria, R. Toro, N. Jahanshad, G. Schumann, B. Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior*, 8(2):153–182, 2014.

P. Todd. Matching estimators. *The New Palgrave Dictionary of Economics*, 2, 2008.

A. W. Toga. *Brain warping*. Academic press, 1998.

M. T. Tsuang. Heterogeneity of schizophrenia. *Biol Psychiatry*, 10(4):465–474, 1975.

V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural computation*, 12(9):2013–2036, 2000.

V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 2. Wiley New York, 1998.

E. Varol, B. Gaonkar, G. Erus, R. Schultz, and C. Davatzikos. Feature ranking based nested support vector machine ensemble for medical image classification. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 146–149. IEEE, 2012.

T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

R. Verma, S. Mori, D. Shen, P. Yarowsky, J. Zhang, and C. Davatzikos. Spatiotemporal maturation patterns of murine brain quantified by diffusion tensor mri and deformation-based morphometry. *Proceedings of the national academy of sciences of the United States of America*, 102(19):6978–6983, 2005.

J. Ye and T. Xiong. Svm versus least squares svm. *Journal of Machine Learning Research - Proceedings Track*, pages 644–651, 2007.

P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.