# ADVANCING STATISTICAL INFERENCE FOR POPULATION STUDIES IN NEUROIMAGING USING MACHINE LEARNING

Erdem Varol

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

_____

Christos Davatzikos
Professor of Radiology,
University of Pennsylvania

Graduate Group Chairperson

_____

Alejandro Ribeiro
Associate Professor of Electrical and Systems Engineering
University of Pennsylvania

DISSERTATION COMMITTEE

Alejandro Ribeiro, Associate Professor of Electrical and Systems Engineering
Danielle Bassett, Associate Professor of Bioengineering
Christos Davatzikos, Professor of Radiology

Dedicated to my Dede (1928-2017)

"Deep in the human unconscious is a pervasive need for a logical universe

that makes sense. But the real universe is always one step beyond logic."

—from 'The Sayings of Muad'Dib' by the Princess Irulan

# Acknowledgments

The story of this thesis begins with my arrival in Philadelphia on a hot and humid August day many years ago. As I stumbled into the office of my future advisor Christos Davatzikos, I did not know what a voxel was, and neither had I ever used a Linux system. I'm very grateful that Christos saw the potential in me and provided the setting and support in which my intellectual growth was able to take form. Countless discussions we've had about the nuances of the wacky methods that I have proposed have allowed me to filter out the noise from the signal regarding ideas.

While Christos provided the foundation on which I could build this thesis on, the construction of this thesis would not have been stable had it not been for the unwavering meticulous mentoring of Aristeidis Sotiras. One of the most important lessons I learned from Aris is that the abstract of a paper should be like an introduction, the introduction should be like a review paper and captions of figures should be stand-alone abstracts that will be worthy of submission to a conference. This attention to detail kept me grounded and allowed me to focus on the real goals in my Ph.D. I hope he can cease to utter the words, "sweet baby Jesus" in reference to his frustration with me from now on.

I am thankful for Alejandro Ribeiro and Danielle Bassett for agreeing to serve on my

dissertation committee. I truly appreciated the numerous discussions I've had with Alejandro early on in my Ph.D. while serving as his teaching assistant. The mathematical rigor with which he assessed my ideas is a quality in him that I admire and aspire to achieve. Dani's comments and suggestions further allowed me to challenge my own set of thoughts and provided further reminders to me that neuroscience is indeed much more complicated than we can ever model.

I especially thank Guray Erus for years of help and advice during my Ph.D. He was the second person after Christos that I met in the lab. Guray provided invaluable feedback in almost all of my manuscripts including my thesis. Importantly, he enabled me to attempt to express my scientific thoughts in Turkish, our native language, an area that I'm still not very successful at to this day.

The data, and incidentally, the results of this thesis would not have been possible without the crucial work done by the data analysts, Jimit "Master" Doshi, Michael Hsieh and Martin Rozycki and the system administrator, Mark Bergman. I thank them sincerely for their daily efforts that often are not recognized by readers of our manuscripts. Furthermore, I appreciate the work done by Paraskevi Parmpi and Amanda Shacklett. They made sure the administrative issues were resolved promptly so the lab could focus on its scientific endeavors.

The mental strength necessary for pushing through doctoral adversity partly came from the camaraderie of my fellow students, Aoyan "Codeboy" Dong, Ke "Mr. Handsome" Zeng and Mrs. Harini Eavani, sitting beside me every day. The often extended conversations about the meaning of life, whether we should eat at Sitar or get chicken and rice, when are we going to graduate and how sad we are, got me through the days

# ABSTRACT

ADVANCING STATISTICAL INFERENCE FOR POPULATION STUDIES IN

NEUROIMAGING USING MACHINE LEARNING

Erdem Varol

Christos Davatzikos

Modern neuroimaging techniques allow us to investigate the brain in vivo and in high resolution, providing us with high dimensional information regarding the structure and the function of the brain in health and disease. Statistical analysis techniques transform this rich imaging information into accessible and interpretable knowledge that can be used for investigative as well as diagnostic and prognostic purposes.

A prevalent area of research in neuroimaging is group comparison, i.e., the comparison of the imaging data of two groups (e.g. patients vs. healthy controls or people who respond to treatment vs. people who don't) to identify discriminative imaging patterns that characterize different conditions. In recent years, the neuroimaging community has adopted techniques from mathematics, statistics, and machine learning to introduce novel methodologies targeting the improvement of our understanding of various neuropsychiatric and neurodegenerative disorders.

However, existing statistical methods are limited by their reliance on ad-hoc assumptions regarding the homogeneity of disease effect, spatial properties of the underlying signal and the covariate structure of data, which imposes certain constraints about the sampling of datasets.

- First, the overarching assumption behind most analytical tools, which are commonly used in neuroimaging studies, is that there is a single disease effect that differentiates

the patients from controls. In reality, however, the disease effect may be heterogeneously expressed across the patient population. As a consequence, when searching for a single imaging pattern that characterizes the difference between healthy controls and patients, we may only get a partial or incomplete picture of the disease effect.

- Second, and importantly, most analyses assume a uniform shape and size of disease effect. As a consequence, a common step in most neuroimaging analyses it to apply uniform smoothing of the data to aggregate regional information to each voxel to improve the signal to noise ratio. However, the shape and size of the disease patterns may not be uniformly represented across the brain.

- Lastly, in practical scenarios, imaging datasets commonly include variations due to multiple covariates, which often have effects that overlap with the searched disease effects. To minimize the covariate effects, studies are carefully designed by appropriately matching the populations under observation. The difficulty of this task is further exacerbated by the advent of big data analyses that often entail the aggregation of large datasets collected across many clinical sites.

The goal of this thesis is to address each of the aforementioned assumptions and limitations by introducing robust mathematical formulations, which are founded on multivariate machine learning techniques that integrate discriminative and generative approaches. Specifically,

1. First, we introduce an algorithm termed **HYDRA** which stands for *heterogeneity through discriminative analysis*. This method parses the heterogeneity in neuroimaging studies by simultaneously performing clustering and classification by use of

piecewise linear decision boundaries.

2. Second, we propose to perform *regionally linear multivariate discriminative statistical mapping* (**MIDAS**) toward finding the optimal level of variable smoothing across the brain anatomy and tease out group differences in neuroimaging datasets. This method makes use of overlapping regional discriminative filters to approximate a matched filter that best delineates the underlying disease effect.

3. Lastly, we develop a method termed *generative discriminative machines* (**GDM**) toward reducing the effect of confounds in biased samples. The proposed method solves for a discriminative model that can also optimally generate the data when taking into account the covariate structure.

We extensively validated the performance of the developed frameworks in the presence of diverse types of simulated scenarios. Furthermore, we applied our methods on a large number of clinical datasets that included structural and functional neuroimaging data as well as genetic data. Specifically, HYDRA was used for identifying distinct subtypes of Alzheimer's Disease. MIDAS was applied for identifying the optimally discriminative patterns that differentiated between truth-telling and lying functional tasks. GDM was applied on a multi-site prediction setting with severely confounded samples. Our promising results demonstrate the potential of our methods to advance neuroimaging analysis beyond the set of assumptions that limit its capacity and improve statistical power.

# Contents

# List of Tables

# List of Figures

| | |
|---|---|
| **AD** | Alzheimer's disease |
| **ADNI** | Alzheimer's disease neuroinitiative |
| **APOE** | Apolipoprotein-E |
| **ARI** | Adjusted Rand index |
| **ASD** | Autism spectrum disorder |
| **AUC** | Area under the receiver operating characteristic curve |
| **CN** | Cognitively normal |
| **CSF** | Cerebrospinal fluid |
| **DBM** | Deformation based morphometry |
| **DPP** | Determinental point process |
| **DTI** | Diffusion tensor imaging |
| **FDR** | False discovery rate |
| **fMRI** | Functional magnetic resonance imaging |
| **FPR** | False positive rate |
| **FWHM** | Full width at half maximum |
| **GDM** | Generative discriminative machine |
| **GLM** | General linear model |
| **GM** | Gray matter |
| **GRBF** | Gaussian radial basis function |
| **HYDRA** | Heterogeneity through discriminative analysis |
| **LS-SVM** | Least squares support vector machine |
| **MCI** | Mild cognitive impairment |
| **MIDAS** | Regionally linear multivariate discriminative statistical mapping |

| | |
|---|---|
| **MMSE** | Mini mental state examination |
| **MRI** | Magnetic resonance imaging |
| **MVPA** | Multivariate pattern analysis |
| **ODVBA** | Optimally discriminative voxel based analysis |
| **OLS** | Ordinary least squares regression |
| **P-SVM** | Support vector machine based statistical significance testing |
| **P-VBM** | Permutation testing based voxel based morphometry |
| **RAVENS** | Regional analysis of volume examined in normalized space |
| **ROI** | Region of interest |
| **SCZ** | Schizophrenia |
| **sMRI** | Structural magnetic resonance imaging |
| **SNP** | Single nucleotide polymorphism |
| **SVM** | Support vector machine |
| **TBM** | Tensor based morphometry |
| **TPR** | True positive rate |
| **VBA** | Voxel based analysis |
| **VBM** | Voxel based morphometry |
| **WM** | White matter |

# Chapter 1

# Introduction

## 1.1 Overview

Neuroimaging techniques enable a detailed non-invasive in vivo exploration of the brain. There is a variety of imaging modalities that provide complementary information about the brain structure and function. Structural magnetic resonance imaging (sMRI) allows the scrutiny of static anatomical structures. Functional magnetic resonance imaging (fMRI) enables the measurement of dynamic activity through measurements of blood flow. Diffusion tensor imaging (DTI) helps to understand the structural connections across the brain.

The analysis of neuroimaging data has been able to shed light on the complex structure and function of the human brain under normal or pathological conditions. Group studies are amongst the most common ways of studying changes in the brain, and they involve the analysis of differences between a control group and a patient group. Analysis techniques for group studies typically fall under two categories: voxel-based analyses and multivariate pattern analysis techniques.

Voxel-based analysis techniques perform statistical tests on a voxel by voxel basis. Such mass univariate tools are used to tease out of the data anatomical and functional entities that describe brain structure and function in an unbiased, hypothesis-free way. These techniques can be further classified depending on the type of information the statistical tests are performed on. Specifically, deformation-based morphometry (DBM) [24, 56] and tensor-based morphometry (TBM) [50, 128] compare the deformation fields, or the derivatives of deformation fields, between different populations, respectively. DBM and TBM both rely on highly accurate registration of brain images, which may not always be possible given the large variation of human brains. On the other hand, voxel-based morphometry (VBM) analysis [3, 160, 62, 58] conducts voxel wise t-tests to compare groups of tissue density maps across different populations with the goal to investigate focal differences in brain anatomy. The generation of tissue density maps is typically accompanied by spatial smoothing of the signal to account for registration errors and to Gaussianize the data. This process makes VBM robust to small registration errors, which makes it one of the most widely used methods for population neuroimaging analysis.

Nevertheless, voxel-based analysis techniques ignore multivariate relations between brain regions that may best characterize population differences. Instead, multivariate pattern analysis (MVPA) methods [6, 106] take advantage of dependencies among brain regions, which leads to increased sensitivity. Statistical mapping frameworks, such as *Searchlight* [89], aim to capture multivariate relations in local neighborhoods of voxels to more accurately detect the underlying differences between groups. Nonetheless, Searchlight does not account for signal that spans distant locations in the brain [44]. For this reason, there exist machine learning methods, such as support vector machine (SVM) [147], that

enable the analysis of the entire brain [85, 88, 154]. These methods mainly focus on selecting voxels or regions that maximize classification accuracy and may not capture all the differences between groups.

Critically, current group analysis techniques either make some assumptions regarding disease effects or are bound by limitations on sample distributions. Specifically:

1. A common assumption of group analyses is that there is a single disease process that affects all samples in the disease group in a unified way, thus resulting in a single imaging pattern of brain differences that discriminates patients from controls.

2. Group analysis methods commonly assume that the spatial extent and the shape of the underlying disease effect is uniform across the brain. Consequently, a smoothing filter with a single bandwidth is applied on imaging maps prior to analysis.

3. A limitation in group analyses is the requirement to match control and disease samples for covariates (e.g., for age and sex) that can have an effect that may overlap with the disease effects one searches for. Otherwise, the assumption is that the unmatched covariates have no confounding effects on the results.

These assumptions are limiting neuroimaging analysis techniques in utilizing the available rich imaging data to its full potential. Particularly, disease processes are rarely homogeneous. There is ample evidence for the heterogeneity of pathological phenotypes presented by many diseases, such as Alzheimer's disease [111, 92], Schizophrenia [46, 114, 86], Autism spectrum disorder [142, 75], and Attention-deficit hyperactivity disorder [155]. As a consequence, current approaches may miss heterogeneous disease effects in the data when searching for a single disease effect pattern. These approaches can only find differences in the central tendency, such as a common imaging pattern of difference when

comparing two populations. Thus, the derived imaging patterns are at best incomplete, and at worst, misleading.

Also, the spatial extent and shape of the differences between controls and patients are seldom uniform across the brain anatomy. Thus, applying a single bandwidth of Gaussian smoothing to the imaging data will fail to amplify the signal that has a spatial extent greater than the width of smoothing kernel and conversely will smear out the signal that is narrower than the width of the kernel width [82]. In both cases, this will result in loss of statistical power by way of reducing sensitivity and specificity, respectively.

Lastly, in real-world datasets, it is often difficult to have a perfect covariate match between groups without pruning very expensive and hard to acquire data. As a result, either the statistical power of group analysis suffers from the reduced sample size that is balanced for covariates, or the sample is confounded by covariate imbalances [125].

## 1.2    Aims of this thesis

The general goal of this thesis is to develop techniques to move beyond the aforementioned assumptions and propose a set of advanced machine learning tools for robust analysis of neuroimaging data. This goal is divided into the three following aims that are detailed below.

### Aim 1: Inference in the presence of population heterogeneity

Brain disorders often exhibit a heterogeneous clinical presentation: autism spectrum disorder (ASD) encompasses neurodevelopmental disabilities characterized by deficits in social communication and repetitive behaviors [57]; schizophrenia can be subdivided into dis-

tinct groups by separating its symptomatology to discrete symptom domains [18]; Alzheimer's disease (AD) can be separated into three subtypes on the basis of the distribution of neurofibrillary tangles [112]; and mild cognitive impairment (MCI) may be further classified based on the type of specific cognitive impairment [157].

Disentangling disease heterogeneity may greatly contribute to our understanding of disease mechanisms and lead to more accurate diagnosis and prognosis, as well as targeted treatment. However, most commonly used neuroimaging analysis approaches assume a single unifying pathophysiological process governing the presence of disease and perform a monistic analysis to identify it. Such approaches typically aim to either identify voxels that characterize group differences through mass-univariate statistical techniques [3] or use MVPA to identify the multivariate imaging pattern that best discriminates between two populations [153]. Thus, the heterogeneity of the disease is completely ignored, which results in deriving imaging patterns that are at best incomplete, and at worst misleading.

Recognizing this limitation, few research efforts have focused on revealing the inherent disease heterogeneity. These methods can be mainly classified into two groups. The first class assumes an a priori subdivision of the diseased samples into coherent groups, based on independent criteria, and opts to identify group-level anatomical differences using univariate statistical methods [87, 156]. Thus, multivariate effects are ignored, while the a priori definition of disease subtypes is either difficult to obtain (*e.g.*, from autopsy near the date of imaging), or noisy and non-specific (*e.g.*, cognitive or clinical evaluations). The second class focuses on the diseased population and maps it to distinct anatomical subtypes by applying multivariate unsupervised clustering driven by considering all image elements [157, 118]. These methods tend to group patients along the direction of largest

variability, which may be confounded by effects such as age and sex, and thus may not be induced by pathology.

**To tackle these challenges, the second aim of the thesis is to develop a method for detecting and characterizing heterogeneity through the data-driven identification of disease subgroups.**

## Aim 2: Inference through optimal spatial filtering

Group analysis studies how distinct clinically-defined groups of individuals differ in brain anatomy and function, aiming to understand the pathophysiological processes that steer these differences. Towards this goal, mass-univariate [3] as well as MVPA techniques [89, 55] have been developed to summarize and understand imaging patterns reflecting a clinical change.

Mass-univariate techniques, such as VBM, have been widely used for neuroimaging analysis. However, mass-univariate techniques ignore multivariate relations in the data, while also suffering from multiple comparison problems. Critically, local smoothing is typically applied to reduce voxel-wise noise, account for errors in spatial alignment of images and Gaussianize the data before performing statistical analyses. However, this smoothing is seldom adapted to the anatomical structures of the brain and may obscure the effects of interest. A narrow blurring kernel cannot effectively account for noise in the data, thus reducing the statistical power. Contrarily, a wide blurring kernel diffuses signal, potentially leading to false conclusions about the real loci of the effect. Additionally, it may introduce signal from regions that have no group difference, thus reducing sensitivity in detecting group differences.

MVPA methods characterize group differences by harnessing multivariate relationships in the data. They can be distinguished into two classes according to whether they perform local or global learning. Local learning techniques, such as Searchlight [89], analyze the information content of local neighborhoods, while global learning methods, such as SVM [55], perform inference by modeling signal relationships across the entire brain. Local techniques are computationally expensive, while they may also lead to serious interpretation errors [44]. Global techniques, by construction, select regions sufficient for discrimination and may not fully reflect the group difference [68].

**Toward addressing the above limitations of univariate and multivariate techniques, the last aim of this thesis is to develop a method for performing statistical inference through optimal spatial filtering of data and regional discriminative analysis.**

## Aim 3: Inference in the presence of confounds

Univariate statistical methods, such as general linear models, effectively account for confounds by explicitly parametrizing them in the model. However, there is no clear consensus on how to reduce confounding effects within MVPA predictive settings. Confounding effects are an important problem in MVPA prediction methods as powerful machine learning methods may learn the covariate structure rather than group effects, which may lead to overfitting and failure to generalize.

Prior approaches have either 1) ignored confounds, or have taken them into account either 2) implicitly, or 3) explicitly [125]. The first approach is to ignore the confounds and proceed with the predictive learning task using the imaging features. The second approach is to implicitly account for the confounds by correcting a posteriori the learned

model using the underlying covariate structure [68]. Lastly, confounds can be adjusted for explicitly. Weighting schemes [136, 139, 101] and residualization approaches [40] explicitly account for confounds prior to the learning model. The limitation of these approaches is that they either compromise generalization for interpretability, or interpretability for generalization. Furthermore, they seldom allow for immediate statistical inference due to lack of insight into the probability distribution of the model parameters.

**Toward addressing the above limitations, the first aim of the thesis is to develop a framework for performing multivariate statistical inference and pattern analysis that is robust to confounding variations.**

## 1.3   Main Contributions

In this thesis, we move beyond the aforementioned commonly applied assumptions by introducing three novel machine learning techniques for reliable and efficient analysis of neuroimaging data.

1. **Statistical inference in the presence of disease heterogeneity:** We introduced a novel convex polytope based learning method that is used to disentangle disease subtypes in a semi-supervised fashion. This method is termed **HYDRA**, which is an acronym for *heterogeneity through discriminative analysis*. This method can be kernelized, which eases the computational burden on high dimensionality datasets. This work is validated using simulated data and applied to an imaging and genetic study of Alzheimer's disease.

2. **Statistical inference through optimal spatial filtering:** We introduced a novel framework that utilizes multiple overlapping local learners, which act as adaptive filters,

to optimally tease out group differences. This method is termed **MIDAS**, which stands for *regionally linear multivariate discriminative statistical mapping*. A key novelty of this method is that its resulting statistic is equipped with an analytical form of null distribution, which enables rapid statistical inference in large neuroimaging datasets. This method is extensively validated using simulated data and is tested using an fMRI dataset of truth-telling and lying to delineate correlated brain regions as well as a sMRI dataset that studies the effects of aging on cognition.

3. **Statistical inference in the presence of confounds:** We introduce a novel discriminative model that encompasses a generative regularization term, which explicitly removes the effects of confounds yielding a confound invariant model. This framework is termed *generative discriminative machine* or **GDM** for short. A key novelty of this method is that the null distribution of the resulting statistical model can be analytically approximated, which allows for accurate statistical inference and significance testing. We demonstrated the robustness of the proposed approach by using data from neuroimaging studies of Schizophrenia and Alzheimer's Disease to carefully design settings influenced by different confounding factors.

By moving beyond commonly applied assumptions in neuroimaging analysis, these frameworks aim to derive data-driven disease subtypes, attain more specific and sensitive imaging biomarkers, and control for confounding variations, respectively. Taken together, these contributions demonstrate great potential in improving our understanding of pathology, enabling therapeutic innovation and improving diagnosis and prognosis.

## 1.4  Organization of this thesis

The three main methodological contributions of this thesis are described in Chapters 2, 3 and 4. In Chapter 2, we describe the HYDRA method for disentangling heterogeneous populations and its validation on simulated data as well as its applications to structural MRI and genetics datasets. Chapter 3 details the MIDAS method that estimates the optimal spatial filtering for statistical inference and its validation on simulated data as well as clinical applications on functional and structural MRI data. Chapter 4 describes the GDM method for adjusting for confounds in neuroimaging datasets and its applications to structural MRI datasets. Chapter 5 summarizes all the contributions of this thesis and discusses possible directions of future research.

# Chapter 2

# Inference in the presence of population heterogeneity: HYDRA

## 2.1 Introduction

Automated analysis of spatially aligned medical images has become the main framework for studying the anatomy and function of the human brain. This is typically performed by either employing voxel-based (VBA) or multivariate pattern analysis (MVPA) techniques.

VBA complements region of interest (ROI) volumetry by providing a comprehensive assessment of anatomical differences throughout the brain, while not being limited by *a-priori* regional hypotheses. VBA typically performs mass-univariate statistical tests on either tissue composition or deformation fields, aiming to reveal regional anatomical or shape differences [5, 60, 3, 33, 25, 51, 78, 90, 26, 138, 13, 59, 77, 107, 2]. However, voxel-wise methods often suffer from low statistical power and more importantly, ignore multivariate relationships in the data.

On the other hand, MVPA techniques have gained significant attention due to their ability to capture complex relationships of imaging signals among brain regions. This property allows to better characterize group differences and could potentially lead to improved diagnosis and personalized prognosis. As a consequence, machine learning methods have been used with increased success to derive highly sensitive and specific biomarkers of diseases on individual basis [109, 84, 32, 153, 39, 130, 105, 42, 69, 28].

A common assumption behind both VBA and MVPA methods is that there is a single pattern that distinguishes the two contrasted groups. In other words, most computational neuroimaging analyses assume a single unifying pathophysiological process and perform a monistic analysis to identify it. However, this approach ignores the heterogeneous nature of diseases, which is supported by ample evidence. Typical examples of brain disorders that are characterized by a heterogeneous clinical presentation include both neurodevelopmental and neurodegenerative disorders: Autism Spectrum Disorder (ASD) comprises neurodevelopmental disorders characterized by deficits in social communication and repetitive behaviors [57, 76]; Schizophrenia and Parkinson's Disease can be subdivided into distinct groups by separating its symptomatology to discrete symptom domains [18, 63, 87, 115, 163, 99]; Alzheimer's Disease (AD) can be separated into three subtypes on the basis of the distribution of neurofibrillary tangles [112]; and Mild Cognitive Impairment (MCI) may be further classified based on the type of specific cognitive impairment [73, 157].

Disentangling disease heterogeneity may significantly contribute to our understanding and lead to a more accurate diagnosis, prognosis, and targeted treatment. However, few research efforts have been focused on revealing the inherent disease heterogeneity.

12

These approaches can be categorized into two distinct classes. The first class assumes an *a priori* subdivision of the diseased samples into coherent groups, based on independent (*e.g.*, clinical) criteria, and opts to identify group-level anatomical or functional differences using univariate statistical methods [73, 87, 115, 156, 163]. As a consequence, multivariate relationships in the data are ignored. Moreover, and more importantly, these methods depend on an *a priori* disease subtype definition, which may be either difficult to obtain (*e.g.*, from autopsy near the date of imaging), or noisy and non-specific (*e.g.*, cognitive or clinical evaluations). Methods belonging to the second class apply multivariate clustering (typically driven by all image elements) directly to the diseased population towards segregating subsets of distinct anatomical subtypes [63, 157, 99, 118]. Such an approach aims to cluster brain anatomies instead of pathological patterns. Thus, it has the potential risk of estimating clusters that reflect normal inter-individual variability, some of which is due to sex, age, and other confounds, instead of highlighting disease heterogeneity.

To tackle the aforementioned limitations, it is necessary to develop a principled machine learning approach that can simultaneously identify a class of pathological samples and separate them into coherent subgroups based on multivariate pathological patterns. To the best of our knowledge, one approach has been previously proposed in this direction [47]. That work tackled disease subtype discovery by simultaneously solving classification and clustering in a semi-supervised maximum margin framework. It jointly estimated two hyperplanes, one that separates the diseased population from the healthy one, and another hyperplane that splits the estimated diseased population into two groups. Thus, only one linear classifier was used to separate patients from controls, thereby limiting its ability to capture heterogeneous pathologic processes. Moreover, it arbitrarily assumed that exactly

13

two disease subgroups exist, rather than attempting to determine the number of subtypes from the data.

Here, we propose a novel non-linear semi-supervised[1] machine learning algorithm for integrated binary classification and subpopulation clustering aiming to reveal **H**eterogeneit**Y** through **D**isc**R**iminative **A**nalysis (HYDRA). To the best of our knowledge, ours is the first algorithm to deal with anatomical/genetic heterogeneity in a supervised-clustering fashion with an arbitrary number of clusters. The proposed approach is motivated by recent machine learning methods that derive non-linear classifiers through the use of multiple-hyperplanes[52, 65, 148, 83, 143, 120]. Classification is performed through the separation of healthy controls from pathological samples by a convex polytope that is formed by combining multiple linear max-margin classifiers. Heterogeneity is disentangled by implicitly clustering pathologic samples through their association to single linear sub-classifiers. Multiple dimensions of heterogeneity may be captured by varying the number of estimated hyperplanes (faces of the polytope). This is in contrast to non-linear kernel classification methods which may accurately fit heterogeneous data in terms of disease prediction, but do not provide any explicit clustering information that can be used to determine subtypes of pathology. HYDRA is a hybrid between unsupervised clustering and supervised classification methods; it can simultaneously fit maximum margin classification boundaries and elucidate disease subtypes, which is not possible with neither unsupervised clustering methods nor non-linear kernel classifiers.

Note that a preliminary version of this work was presented in [149]. The current chapter extends our previous work in multiple ways: i) A more sophisticated initialization

---

[1]The term semi-supervised is in reference to lack of disease subtype labels that must be inferred from data

scheme based on Determinantal Point Processes is employed (Sec. 2.3.1); ii) The sensitivity to initialization due to the non-convexity of the objective function has been improved by using multiple initializations and consensus strategies (Sec. 2.3.4); iii) A symmetric version of the algorithm is developed towards accounting for the heterogeneity of the healthy controls and avoiding over-learning (Sec. 2.2.4). iv) A detailed description of the proposed methodology is provided. v) We extensively evaluate our method, HYDRA, by using additional (imaging and genetic) datasets and comparing it to unsupervised clustering and non-linear classification methods.

The remainder of this chapter is organized as follows. In section 2.2, we detail the proposed approach. Next, we experimentally validate our method using synthetic (Sec. 2.4) and clinical (Sec. 2.5) data. We discuss the results in Sec. 2.6, while section 2.6 concludes the chapter with our final remarks.

## 2.2   Method

In high dimensional spaces, the modeling capacity of linear Support Vector Machines (SVMs) is theoretically rich enough to discriminate between two homogeneous classes. However, while two classes are linearly separable with high probability, the resulting margin may be small. This case arises for example when one class is generated by a multimodal distribution that models a heterogeneous process (see Fig. 2.1a). This may be remedied by the use of non-linear classifiers, allowing for larger margins and thus, better generalization. However, while kernel methods, such as Gaussian Radial Basis Function (GRBF) kernel SVM, provide non-linearity, they lack interpretability when aiming to characterize heterogeneity.

(a)          (b)          (c)
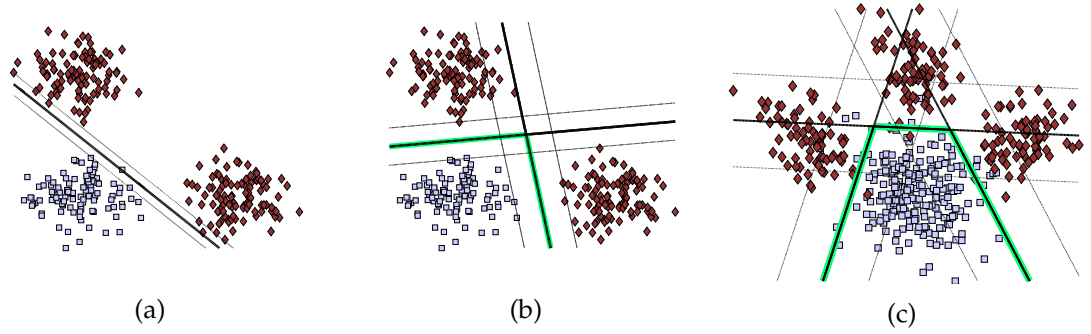
Figure 2.1: Illustrating the effect of heterogeneity when separating a positive class (denoted by gray squares) from a heterogeneous negative class (denoted by red rhombuses). (a) Linear SVM separates the positive class from a heterogeneous negative class (presence of two clusters) by a small margin. (b) Our method classifies each cluster separately, resulting in a larger margin. (c) Heterogeneity introduced by the presence of three clusters modeling distinct deviations from normality. Each deviation is captured by a different face of the convex polytope. Solid lines correspond to the classifier, dashed lines indicate margin while highlighted linear segments define the separating convex polytope.

Here, we take advantage of the previous intuition to design a novel machine learning technique that will provide larger margins while being able to elucidate heterogeneity. We introduce non-linearity using multiple linear classifiers that form locally linear hyperplanes whose linear segments separate the clusters of negative samples from the positive class (see Fig. 2.1b). In this way, subjects are explicitly clustered by being assigned to different hyperplanes, giving rise to interpretable directions of variability that may be useful in discovering heterogeneity.

Suppose that our dataset consists of $n$ binary labelled $d$-dimensional data points ($\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^{n}, \mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$). Without loss of generality, we assign the negative class to the pathological population whose heterogeneity we seek to reveal. Let us note that while there may be heterogeneity in the healthy population, we focus here on revealing disease heterogeneity. Our aim is twofold. First, we aim to estimate $k$ hyperplanes that form a convex polytope that separates the two classes with a large margin. Second, we aim to

16

assign each pathological sample to the hyperplane that best separates it from the normal controls. The main idea is that samples that belong to different pathological subgroups will be assigned to different hyperplanes, each of which reflects a respective pathological process (see Fig. 2.1c). Towards fulfilling the aims mentioned above, we introduce the proposed approach by extending standard linear maximum margin classifiers.

### 2.2.1 Large Margin Classification

For completeness, let us briefly introduce standard linear maximum margin classifiers. Maximum margin classifiers aim to estimate a hyperplane that separates the two classes by a half space, while ensuring that the distance (or margin) from the decision boundary for each sample is maximized. More formally, suppose that the set $\mathcal{F}$ comprises the set of all linear classifiers $\mathbf{w}$ such that for the given dataset $\mathcal{D}$ all samples are correctly classified, or $\forall i,\ y_i(\mathbf{w}^T\mathbf{x}_i) + b \geq 1$. The goal is to find the classifier $\mathbf{w}$ belonging to the set $\mathcal{F}$ that maximizes the margin between classes. The margin is defined as the orthogonal distance between the two hyperplanes:

$$\mathbf{w}^T\mathbf{u} + b = -1, \text{ and } \mathbf{w}^T\mathbf{v} + b = +1,$$

where the set of points $\mathbf{u}, \mathbf{v}$ that satisfy the equations, represent points from both classes with active constraints. Notice that setting $\mathbf{u} = -\frac{1+b}{\|\mathbf{w}\|_2^2}\mathbf{w}$ and $\mathbf{v} = \frac{1-b}{\|\mathbf{w}\|_2^2}\mathbf{w}$ satisfies the previous equations. Since $\mathbf{u}, \mathbf{v}$ are parallel, the orthogonal distance between the hyperplanes is simply $\|\mathbf{u} - \mathbf{v}\|_2 = \frac{2}{\|\mathbf{w}\|_2}$, which is the margin for SVM [147].

The optimal classifier is estimated by solving an optimization problem. However, in-

stead of maximizing the margin, its inverse ($\frac{\|\mathbf{w}\|_2^2}{2}$) is typically minimized subject to the

separability constraints. This results in the well known SVM objective:

$$\underset{\mathbf{w},b,\xi}{\text{minimize}} \; \frac{\|\mathbf{w}\|_2^2}{2} + C \sum_{i=1}^{n} \xi_i$$

subject to

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

where $\xi = (\xi_1, \ldots, \xi_n)$. The second term of the objective $\left(C \sum_{i=1}^{n} \xi_i\right)$ accounts for slack when

classes are non-separable.

### 2.2.2 Convex Polytope Classification

Standard SVMs assume that there is a single pattern (encoded by the estimated hyper-

plane) that distinguishes the two classes. However, this assumption is violated in the case

of heterogeneity. We aim to model heterogeneity by utilizing multiple linear hyperplanes,

each one corresponding to a different pathological pattern. By combining multiple linear

classifiers in a piecewise fashion, we extend linear max-margin classifiers to the non-linear

case. Thus, we consider the extended hypothesis class that consists of the set of sets of $K$

hyperplanes, generalizing the geometry of the classifier to that of a convex polytope [143].

Due to the interior/exterior asymmetry of the polytope, it is necessary to confine one class

to its interior while restricting the other class to its exterior. Without loss of generality,

we confine the positive class to the interior of the polytope. Thus, the search space $\mathcal{F}_K$ is

defined as:

$$\mathcal{F}_K \triangleq \Big\{ \{\mathbf{w}_j, b_j\}_{j=1}^K \mid \forall j, \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 \text{ if } y_i = +1,$$

$$\exists j : \mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 \text{ if } y_i = -1 \Big\}.$$

In other words, $\mathcal{F}_K$ comprises all sets of $K$ classifiers such that all classifiers correctly classify all members of the positive class, while for every negative sample, there is at least one classifier that correctly classifies it.

The latter gives rise to an assignment problem, where samples that have been affected by the same pathological process are assigned to the same hyperplane. This can also be seen as a clustering task since samples that have been assigned to the same hyperplane can be equivalently considered as clustered together. Thus, if $\mathbf{S}^- = [s_{i,j}] \in \{0,1\}^{n^- \times K}$ denotes the binary matrix that describes the assignment of the $i$-th negative class sample ($n^-$ in number) to the $j$-th face of the polytope, then the search space becomes:

$$\mathcal{F}_K(\mathbf{S}^-) \triangleq \Big\{ \{\mathbf{w}_j, b_j\}_{j=1}^K \mid \forall j, \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 \text{ if } y_i = +1,$$

$$\mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 \text{ if } y_i = -1 \text{ and } s_{i,j} = 1 \Big\}.$$

Given the assignment $\mathbf{S}^-$, there are $K$ margins; each one corresponding to one face of the polytope. Analogous to the SVM formulation, the margin for the $j$-th face of the polytope is $\frac{2}{\|\mathbf{w}_j\|_2}$. However, due to the piecewise nature of the convex polytope, there are multiple notions of margin for the surface of the polytope. In this work, aiming to keep the problem tractable, we maximize the average margin across all the faces of the polytope:

19

$\bar{m} = \frac{1}{K} \sum_{j=1}^{K} \frac{2}{\|\mathbf{w}_j\|_2}$. Thus, for a given dataset $\mathcal{D}$ and assignment $\mathbf{S}^-$ for the negative class, the objective becomes:

$$\underset{\{\mathbf{w}_j, b_j\}_{j=1}^K}{\text{maximize}} \quad \frac{1}{K} \sum_{j=1}^{K} \frac{2}{\|\mathbf{w}_j\|_2}$$

subject to

$$\mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 \qquad\qquad\qquad \text{if } y_i = +1 \text{ for } j = 1, \ldots, K$$

$$\mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 \qquad\qquad\qquad \text{if } y_i = -1 \text{ and } s_{i,j} = 1$$

Note that, given the assignments, the objective, and the constraints are separable into $K$ independent subproblems. Each subproblem is analogous to the SVM formulation after adding the slack terms $\xi_{i,j}$, or:

$$\underset{\mathbf{w}_j, b_j, \xi_j}{\text{minimize}} \quad \frac{\|\mathbf{w}_j\|_2^2}{2} + C \sum_{i=1}^{n} \xi_{i,j}$$

subject to

$$\mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 - \xi_{i,j} \qquad\qquad\qquad \text{if } y_i = +1$$

$$\mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 + \xi_{i,j} \qquad\qquad\qquad \text{if } y_i = -1 \text{ and } s_{i,j} = 1$$

$$\xi_{i,j} \geq 0 \qquad\qquad\qquad\qquad\qquad \text{for } i = 1, \ldots, n$$

where $C$ is a penalty parameter on the training error. If we now use the definition of the slack terms as $\xi_{i,j} = \max\{0, 1 - y_i(\mathbf{w}_j^T \mathbf{x}_i + b_j)\}$, and consider all hyperplanes ($\{\mathbf{W}, \mathbf{b}\} \triangleq$

$\{\mathbf{w}_j, b_j\}_{j=1}^K$) at the same time, we get:

$$\underset{\{\mathbf{w}_j, b_j\}_{j=1}^K}{\text{minimize}} \sum_{j=1}^K \frac{\|\mathbf{w}_j\|_2^2}{2} + C \sum_{\substack{i|y_i=+1 \\ j}} \frac{1}{K} \max\{0, 1 - \mathbf{w}_j^T \mathbf{x}_i - b_j\}$$

$$+ C \sum_{\substack{i|y_i=-1 \\ j}} s_{i,j} \max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} \tag{2.1}$$

So far, we have assumed that the assignment matrix $\mathbf{S}^-$ is known. However, this is not the case in practice, and $\mathbf{S}^-$ has to be estimated too.

Attempting to solve for both $\{\mathbf{W}, \mathbf{b}\}$ and $\mathbf{S}^-$ results in a non-convex objective function which is combinatorially difficult to optimize. Furthermore, optimization for the binary assignment $\mathbf{S}^-$ is itself non-convex since it constitutes an integer programming task. To make the problem tractable, we take two steps. First, we relax the binary assignment ($s_{i,j} \in \{0, 1\}$) to a soft assignment ($s_{i,j} \in [0, 1]$, $\sum_{j=1}^K s_{i,j} = 1$, $\forall i$). Given this relaxation, the objective becomes block-wise convex with respect to the groups of variables $\{\mathbf{W}, \mathbf{b}\}$ and $\{\mathbf{S}^-\}$. We then use this relaxed objective function to obtain locally optimal solutions by iteratively solving for $\{\mathbf{W}, \mathbf{b}\}$ and $\{\mathbf{S}^-\}$. The details of the iterative optimization are given in 2.3.

**Prediction**

Once the polytope classifier $\{\mathbf{W}, \mathbf{b}\}$ is trained, predicting the class $y^*$ of a new instance $\mathbf{x}^*$ is straightforward:

$$y^* = \text{sign}(\min_j \mathbf{w}_j^T \mathbf{x}^* + b_j)$$

In other words, if $\mathbf{x}^*$ is in the interior of the polytope defined by the estimated hyperplanes ($\{\mathbf{W}, \mathbf{b}\}$), then it is classified as positive by all classifiers corresponding to the faces of the polytope ($\mathbf{w}_j^T \mathbf{x}^* + b_j > 0$), resulting in an overall positive class prediction ($y^* = +1$). Otherwise, if $\mathbf{x}^*$ is in the exterior of the polytope, then it is classified as negative by at least one classifier corresponding to a face of the polytope ($\mathbf{w}_j^T \mathbf{x}^* + b_j < 0$), resulting in an overall negative class prediction ($y^* = -1$). Analogously, the prediction score is simply the minimum of the prediction scores of all classifiers corresponding to the faces of the polytope: ($\min_j \mathbf{w}_j^T \mathbf{x}^* + b_j$). Moreover, a new sample may be assigned to the existing clusters by computing the assignment index $s_{*,j}$ using Eq. 2.3.

### 2.2.3   HYDRA Algorithm

Given the solutions of $\{\mathbf{W}, \mathbf{b}\}$ and $\mathbf{S}^-$ outlined in Sec. 2.3.2 and Sec. 2.3.3, we solve for the maximum margin convex polytope in an iterative fashion. This is the main workhorse behind the proposed framework that aims to elucidate **H**eterogeneit**Y** through **D**isc**R**iminative **A**nalysis (HYDRA) and is outlined in Algorithm 1. However, due to the non-convex nature of the problem, it is necessary to take additional steps to ensure the high quality of the solution.

Our approach towards enhancing the quality of the solution is twofold. First, particular care is taken to initialize the iterative algorithm in such a way that clustering solutions that exhibit disease-related diversity are promoted. This is made possible by employing Deterministic Point Processes (DPP) [91] to sample diverse directions of pathology, which can subsequently be used to estimate the initial clustering assignments (see 2.3.1 for details).

Second, acknowledging the fact that, in non-convex settings, the estimated solution

**Algorithm 1 — HYDRA**

---

**Input:** $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \{-1, +1\}^n$ (training signals), $C$ (loss penalty), $K$ (number of clusters/hyperplanes)
**Output:** $\mathbf{W} \in \mathbb{R}^{d \times K}$, $\mathbf{b} \in \mathbb{R}^{1 \times K}$ (Classifier); $\mathbf{S}^- \in [0, 1]^{n^- \times K}$ (Clustering Assignment)
**Initialization:** Initialize $\mathbf{S}^-$ by Algorithm 2
**Loop:** Repeat until convergence (or a fixed number of iterations)
• Fix $\mathbf{S}^-$ — Solve for $\mathbf{W}, \mathbf{b}$ by weighted LIBSVM (sample weights set by Eq. 2.4)
• Fix $\mathbf{W}, \mathbf{b}$ — Solve for $\mathbf{S}^-$ using Eq. 2.3

---

may vary greatly depending on the initialization, we employ a multi-initialization strategy that is coupled with a fusion step. Multiple runs of the Algorithm 1 are performed using different initializations generated by the previously described DPP sampling process, as well as different subsets of the population. The estimated clusters constitute hypotheses that capture perturbations of the underlying group topography. These clustering hypotheses are aggregated by taking into account the consensus of the respective solutions, producing the final clustering result that is free of noisy perturbations and emphasizes the underlying group structure (see 2.3.4 for details).

### 2.2.4 Symmetric HYDRA algorithm

The algorithm that we have so far outlined is asymmetric. The patients lie on the exterior of the polytope while the controls are constrained on the interior of the polytope. This property may result in over-fitting when classifying. This can be remedied by symmetrizing the algorithm. One can run the Algorithm 1 twice, once using the actual labels $Y$ and once using the negated labels: $-Y$. In that case, one can use the estimated output polytopes $[\mathbf{W}^+, \mathbf{b}^+]$ and $[\mathbf{W}^-, \mathbf{b}^-]$ to make predictions using the following formula:

$$y^* = \text{sign}\left(\left(\min_j \mathbf{w}_j^{+T} \mathbf{x}^* + b_j^+\right) - \left(\min_j \mathbf{w}_j^{-T} \mathbf{x}^* + b_j^-\right)\right), \tag{2.2}$$

where both classifiers are taken into account.

Note that the symmetric model does not affect the clustering of the patients since the two runs of Algorithm 1 are independent of each other. The difference is that the symmetric model provides two clusterings, one for the patients, and one for the controls.

## 2.3 Optimization

Similar to other clustering methods, HYDRA algorithm requires an initialization step followed by iterations of assignment and convex polytope solutions. To make the clustering robust, we further find the consensus of the clustering results obtained in multiple runs of HYDRA. Here we detail the techniques used for each of these steps. Initialization is found in 2.3.1, assignment step is found in 2.3.2, convex polytope solution is in 2.3.3 and consensus is found in 2.3.4.

As mentioned in the main text, HYDRA is geometrically asymmetric, requiring one of the groups to lie inside the polytope. We provide the solution for the symmetric version of HYDRA in 2.2.4.

Lastly, HYDRA can be solved in the dual domain if the sample size is relatively lower than the dimensionality. The dual solution is in 2.3.6.

### 2.3.1 Initialization

Due to the non-convex nature of the maximum margin polytope problem, the initialization is crucial in directing the iterative algorithm towards favorable solutions. Since we are interested in elucidating discriminative patterns between controls and patients, simply initializing by clustering the patients may not be sufficient. This is because standard clus-

24

**Algorithm 2 — Initialization — Determinantal Point Processes**

---

**Input:** $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \{-1, +1\}^n$ (training signals), $K$ (number of clusters), $m$ (number of hyperplanes samples to draw)

**Output:** $\mathbf{S}^{-0} \in [0,1]^{n^- \times K}$ (Initial Clustering Assignment)

- Randomly draw $m$ pairs of negative ($\mathbf{x}^-$) and positive ($\mathbf{x}^+$) samples (with replacement): $\{\mathbf{x}_i^-, \mathbf{x}_i^+\}_{i=1}^m$
- Obtain $m$ hyperplanes by taking the difference between members of the same pair: $\mathbf{u}_i = (\mathbf{x}_i^+ - \mathbf{x}_i^-)/\|\mathbf{x}_i^+ - \mathbf{x}_i^-\|_2$
- Sample $K$ hyperplanes $\{\mathbf{w}_j^0\}_{j=1}^K$ from $\{\mathbf{u}_i\}_{i=1}^m$ by Determinantal Point Processes [91]
- Set rows of $\mathbf{S}^-$ such that $s_{i,\arg\min_j \mathbf{w}_j^{0T} \mathbf{x}_i} = 1$, otherwise set $s_{i,j} = 0$

---

tering may group patients by following global patterns, such as the brain volume, or even more subtle patterns that nonetheless reflect normal inter-individual variability and not variability in the disease process. On the contrary, patients should be assigned to initial clusters by considering their difference map with respect to controls. In other words, since we aim to explore different directions of deviation from normal anatomy without concern for the magnitude of that deviation, we initially group patients into clusters based on the regions in which they differ from the controls and not the magnitude of their difference. To achieve this, we initialize the assignments of patients into clusters by sampling $K$ unit length hyperplanes obtained by considering the space of all pairwise differences between patients and controls. We choose $K$ unique hyperplanes by applying Determinantal Point Processes (DPP) [91]. DPP is a sampling technique that aims to obtain samples that are as diverse as possible. This type of sampling ensures that the differences we sample reflect unique biomarkers instead of repeated biomarkers with varying magnitudes. This is crucial in preventing clustering patients into groups that are not related to variability in the disease process. The steps of the initialization algorithm are given in Algorithm 2.

### 2.3.2 Assignment Step Solution

For $\{\mathbf{W}, \mathbf{b}\}$ fixed, the problem of estimating $\mathbf{S}^-$ is an assignment problem that can be cast as a linear program (LP). The LP problem has infinite solutions when the loss function $\max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\}$ is equal to 0 for multiple classifiers $j$ and for the same sample $i$. In this case, we choose the solution that is proportional to the margin:

$$
s_{i,j} = \begin{cases} 0 & \text{if} \ \ \max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} > 0 \\[2ex] \dfrac{1 + \mathbf{w}_j^T \mathbf{x}_i + b_j}{\sum_j (1 + \mathbf{w}_j^T \mathbf{x}_i + b_j) \mathbf{1}(\max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} \leq 0)} & \text{otherwise} \end{cases} \tag{2.3}
$$

where $\mathbf{1}(\cdot)$ is the indicator function. Let us note here that the obtained clustering is inherently different from the result that is obtained by standard clustering techniques. Instead of grouping together samples based on the similarity of their appearance, we aggregate here samples that are best separated by the same classifier. Thus, the inferred clustering is driven by discrimination. The more pronounced the pathology is, the easier it is to disentangle the underlying heterogeneity in the imaging profiles.

### 2.3.3 Convex Polytope Solution

For $\mathbf{S}^-$ fixed, the solution to $\{\mathbf{W}, \mathbf{b}\}$ can be obtained using $K$ calls to a modified version of LIBSVM [21][2] that allows for adaptive sample weightings. The adaptive weight $c_{i,j}$ of

---

[2]http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/weights/

sample $i$ for the classifier $j$ is calculated as:

$$c_{i,j} = \begin{cases} Cs_{i,j} & \text{if } y_i = -1 \\ \\ \frac{C}{K} & \text{if } y_i = +1 \end{cases} \tag{2.4}$$

In case the dataset is highly unbalanced (*i.e.*, one of the classes is overrepresented) samples in each class can be further weighted by their inverse relative proportion within the training set.

### 2.3.4 Consensus Solution

While DPP initialization serves as the first step in avoiding poor locally optimal solutions, consensus clustering serves as the second layer to eliminate unstable clusterings that may arise due to the non-convexity of the objective function. In noisy, or high dimensional data, the clustering obtained via Algorithm 1 may depend greatly on the initialization. To decrease this dependency and obtain stable clustering results that characterize the disease heterogeneity, we opt for a multi-initialization strategy, endowed by a fusion step. First, multiple runs of Algorithm 1 result in a number of clustering hypotheses. Then, we aim to fuse the respective hypotheses by harnessing the wisdom of the crowd to obtain an aggregate clustering. A consensus is achieved by grouping together samples that co-occur (*i.e.*, they are assigned to the same clustering) across different clustering hypotheses. In practice, we first compute a co-occurrence matrix of the subjects based on each clustering result and then perform spectral clustering using it.

---

**Algorithm 3 — Consensus Clustering**

---

**Input:** $\{\mathbf{S}^{-p} \in [0,1]^{n^- \times K}\}_{p=1}^{P}$ ($P$ clusterings from Algorithm 1), $K$ (number of clusters)

**Output:** $\mathbf{S}^- \in [0,1]^{n^- \times K}$ (Final Clustering Assignment)

• Compute co-occurrence matrix $\mathbf{A}$ using Eq. 2.5

• Spectral clustering on $\mathbf{A}$:

  • Compute Laplacian matrix $\mathbf{L} = \text{diag}\left(\sum_{l=1}^{n^-} \mathbf{A}_{i,l}\right) - \mathbf{A}$

  • Compute the $K$ eigenvectors $(\mathbf{v}_1, \ldots, \mathbf{v}_K)$ that correspond to $K$ smallest eigenvalues of $\mathbf{L}$ ($\lambda_1 \leq \ldots \leq \lambda_K$)

  • $\mathbf{S}^- \leftarrow$ K-means($[\mathbf{v}_1 \ldots \mathbf{v}_K]$)

---

**Co-occurrence Matrix**

Given $P$ clusterings $\{\mathbf{S}^{-p}\}_{p=1}^{P}$ obtained by running Algorithm 1 $P$ times, the co-occurence matrix $\mathbf{A}$ is given by:

$$\mathbf{A}_{i,l} = \sum_{p=1}^{P} \sum_{j=1}^{K} s_{i,j}^{p} s_{l,j}^{p} \quad i,l = 1 \ldots n, i \neq l \tag{2.5}$$

$$\mathbf{A}_{i,i} = 0 \quad i = 1 \ldots n$$

In other words, each $il$-th entry of the matrix enumerates the number of cases that the $i$-th and $l$-th sample were assigned to the same cluster.

**Spectral Clustering**

The consensus clustering involves the calculation of the Laplacian matrix from the co-occurrence matrix $\mathbf{A}$ and the computation of the $K$ eigenvectors ($[\mathbf{v}_1 \ldots \mathbf{v}_k]$) that correspond to the $K$ smallest eigenvalues ($\lambda_1 \leq \ldots \leq \lambda_K$). Then, the aggregate clustering of subjects is obtained by running K-means in the obtained subspace. The implementation of consensus clustering is outlined in Algorithm 3. It should be noted that the consensus clustering presented herein is analogous to spectral clustering [116].

### 2.3.5 Dual Optimization

Due to the high dimensional, low sample size nature of neuroimaging data, it would be useful to operate in the dual domain to ease the computational burden. The dual formulation of HYDRA can be obtained by converting Eq. 2.1 to:

$$
\underset{\{\alpha_{i,j}\}_{j=1,\ldots,K}^{i=1,\ldots,n}}{\text{maximize}} \quad \sum_{j=1}^{K}\sum_{i=1}^{n}\alpha_{i,j} - \frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{n}\sum_{l=1}^{n}\alpha_{i,j}\alpha_{l,j}y_iy_l\mathbf{x}_i^T\mathbf{x}_l
$$

subject to

$$
\sum_{i=1}^{n}\alpha_{i,j}y_i = 0 \quad j = 1,\ldots,K
$$

$$
C/K \geq \alpha_{i,j} \geq 0 \quad \text{if } y_i = -1 \quad j = 1,\ldots,K
$$

$$
Cs_{i,j} \geq \alpha_{i,j} \geq 0 \quad \text{if } y_i = +1 \quad j = 1,\ldots,K
$$

The advantages of this formulation are two-fold. First, it allows us to solve for only $n \times K$ variables $\{\alpha_{i,j}\}_{j=1,\ldots,K}^{i=1,\ldots,n}$ instead of $K \times d$ variables, which may be prohibitively large. Second, via the kernel trick, we may substitute $\mathbf{x}_i^T\mathbf{x}_j$ with any kernel satisfying the Mercer condition. In terms of implementation, this formulation is readily adaptable to the weighted LIBSVM [21] implementation. Similar to the case of the primal problem, the weights are given by Eq. 2.4.

This formulation does not affect the assignment step solution since the assignment step requires only the prediction score for each subject corresponding to the $K$ hyperplanes. Since the hyperplanes are defined as $\mathbf{w}_j = \sum_{i=1}^{n} y_i\alpha_{i,j}\mathbf{x}_i$, the prediction score for each hy-

perplane $\mathbf{w}_j$ can be simply calculated as:

$$\mathbf{w}_j^T \mathbf{x}_l = \sum_{i=1}^{n} y_i \alpha_{i,j} \mathbf{x}_i^T \mathbf{x}_l$$

which can be readily obtained from the Gram matrix that stores the inner products be-

tween data points. Furthermore, the bias terms $b_j$ can be solved in the dual by:

$$b_j = y_l - \sum_{i=1}^{n} \alpha_{i,j} y_i \mathbf{x}_i^T \mathbf{x}_l$$

using any labeled sample $(\mathbf{x}_l, y_l)$ such that $C > \alpha_{i,l} > 0$. The solutions for $\{\alpha_{i,j}, b_j\}$ can be

directly used in Equation 2.3 to solve for the assignments $\mathbf{S}^-$. In addition, the prediction

for the dual version of HYDRA is:

$$y^* = \text{sign}\left( \min_j \sum_{i=1}^{n} y_i \alpha_{i,j} \mathbf{x}_i^T \mathbf{x}^* + b_j \right)$$

### 2.3.6   Dual Symmetric Prediction

In the case of the symmetric version of the algorithm, the final prediction can be obtained

as:

$$y^* = \text{sign}\left[ \left( \min_j \sum_{i=1}^{n} y_i \alpha_{i,j}^+ \mathbf{x}_i^T \mathbf{x}^* + b_j^+ \right) \right.$$
$$\left. - \left( \min_j \sum_{i=1}^{n} y_i \alpha_{i,j}^- \mathbf{x}_i^T \mathbf{x}^* + b_j^- \right) \right]$$

## 2.4 Experiments using Simulated Data

We first validated the proposed method using synthetic data. We used a two-dimensional toy dataset to provide insight into the workings of the proposed approach. Then, we quantitatively validated the proposed approach against common clustering and classification approaches in a simulated dataset where heterogeneity has been introduced. We evaluated the ability of HYDRA to distinguish between two classes and demonstrated its potential to reveal relevant subgroups.

Let us note that for all experiments, the classification was performed using the symmetric version of HYDRA, while the clustering of the negative class was used to reveal disease heterogeneity. The final clustering was the consensus result of twenty repetitions. The primal formulation was employed when tackling low-dimensional data, while the dual formulation was preferred in the case of high-dimensional data (see 2.3.6 for the dual formulation).

### 2.4.1 Toy Example

To illustrate the behavior of our method, we generated a synthetic two-dimensional dataset with thousand instances (see Fig. 2.2). The first half of the samples were drawn from a unimodal distribution, simulating the healthy control population (denoted by magenta squares). The other half consisted of a crescent-shaped cluster of points, corresponding to the heterogeneous disease group (denoted by rhombuses colored using different variants of blue). To provide a more comprehensive setting, we additionally considered two different separability cases between the two populations. In the first case (see Fig. 2.2a), the two classes overlapped highly, resulting in low separability. In the second case (see Fig.
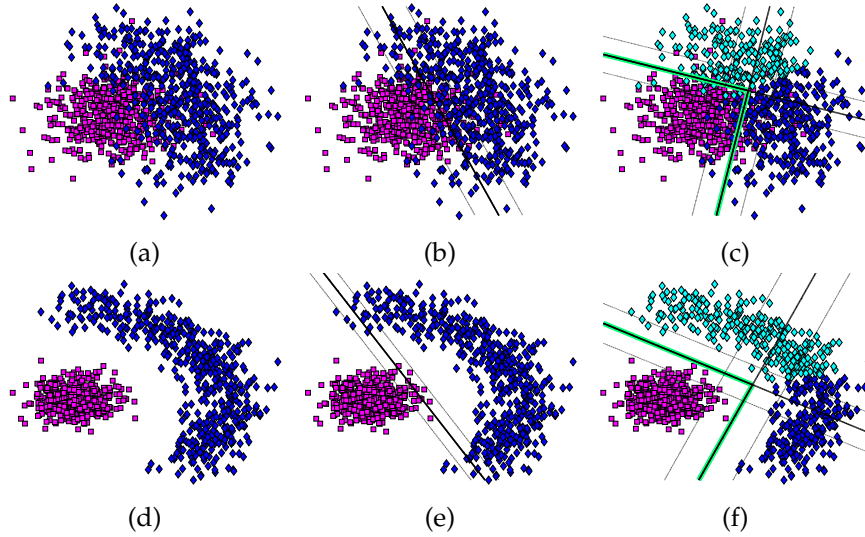
Figure 2.2: Positive (squares) and negative (rhombuses) instances in a continuous two-dimensional feature space. Instances of the two classes either (a) overlap and are not linearly separable, or (b) are highly separable. Linear SVM is used to classify the low (b) and high (e) separability toy dataset. Similarly, HYDRA ($K=2$) is applied to the low (c) and high (f) separability toy dataset. Dark gray lines correspond to the estimated separating hyperplanes, while light gray lines denote the estimated margins. Note the increase of the margin that is made possible through the use of multiple linear classifiers that form a convex polytope denoted by the highlighted line segments. The classes, as well as the estimated subgroups, are encoded using different colors.

2.2d), the two groups did not overlap and were separated by a significant margin, thus increasing separability.

To further clarify the advantages of the proposed framework, we compared the performance of HYDRA (using two hyperplanes, $K = 2$) against the performance of standard linear SVM. The results of the experiments are shown in Fig. 2.2. There are two important observations to make. First, the introduced non-linearity in HYDRA allows for improved separability between the two groups in both scenarios (see Fig. 2.2b, 2.2c, 2.2e and 2.2f). This increase is more important in the case of low-separability between classes (see Fig. 2.2b and 2.2c), where the linear SVM was not able to fully separate them. In the case of high-separability, the hyperplane that was estimated by the linear SVM effectively sepa-

rated positive from negative samples. However, it did so by a relatively small margin (see Fig. 2.2b). On the other hand, HYDRA harnessed the non-linear structure of the data and separated them with a high margin that led to improved generalization performance (see Fig. 2.2f).

Second, and most importantly, HYDRA separated the negative class into two subgroups that differ from the positive class in two distinct directions. This clustering is directly related to the hyperplanes that separate the two classes. As a consequence, the obtained clustering is obtained in a supervised fashion, and thus, it is driven by discriminating patterns that capture disease heterogeneity. This is in contrast to standard clustering techniques that group together samples based on appearance, which is not necessarily related to disease variability.

### 2.4.2   Simulated High-Dimensional Heterogeneous Data

Despite ample evidence of disease heterogeneity, the lack of labeled ground-truth poses a fundamental obstacle in validating the proposed approach. Thus, to overcome these limitations, we construct a simulated validation setting that allows for quantitative comparisons with other algorithms.

Aiming to replicate the common high-dimensional low sample size regime that is prevalent in neuroimaging studies, we generated a synthetic dataset with three hundred instances (or subjects) that are sampled as images with features on a $64 \times 64$ grid. The positive class (healthy group) was generated by randomly sampling 150 samples from a multivariate unimodal Gaussian distribution with zero mean and unit variance ($\mathcal{N}(0, 1)$). The negative class (disease group) was generated by drawing 150 samples from a tri-modal distri-
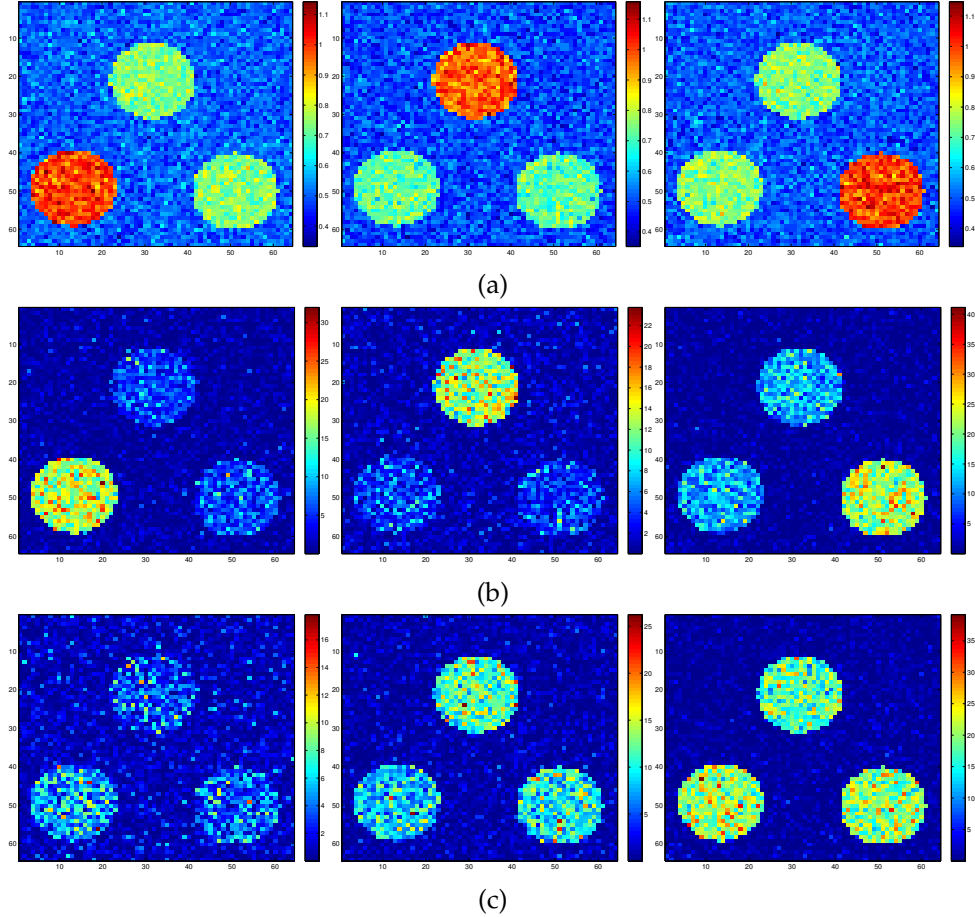
Figure 2.3: (a) Patterns of simulated heterogeneity. Mean difference images between the positive class and the three negative class subgroups, respectively. (b) The results that were obtained using HYDRA ($K = 3$) are visualized by performing group comparison between each estimated subgroup and the positive class. The negative logarithm of the estimated $p$-values is shown. (c) Similarly, the groups that were obtained using K-means ($K = 3$) are reported. Note that the groups estimated by HYDRA capture distinct focal effects that align well with the simulated ones, while the ones estimated by K-means mix the focal effects and recapitulate different stages of disease progression.

bution, where each mode simulates a different focus of disease progression (see Fig. 2.3a).

Each focal effect had a radius of 10 pixels, with a variance of 0.5 units. To simulate the

effect of disease progression, an age effect was simulated. This was generated by adding

unit variance random noise to simulate progression. Therefore, there were three distinct

focal effects in each subgroup, the subgroup specific effect with variance 1.5 units and the

34

non-specific effects with unit variance. Additionally, 10% of the labels were mislabeled to simulate misdiagnosis and label noise.

**Validation Measures**

HYDRA is in principle an exploratory analysis tool, aiming to reveal disease heterogeneity. However, it operates by simultaneously performing classification and clustering. Thus, it is of interest to understand how well the proposed method accomplishes each step.

To validate the classification performance, we computed the Area Under the receiver operating characteristic Curve (AUC) [16]. The AUC statistic summarizes the quality of the performance of a binary classifier. It is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Thus, an AUC equal to one indicates a perfect classifier. We calculated the distribution of AUC values by performing 100 realizations of 10-fold cross-validation. During each iteration, the data were partitioned into ten folds. Each fold was successively used as a test set while the remaining folds were used to train the method. The optimal parameter $C$ of the method was estimated by performing a grid search over $C \in \{2^{-5}, \ldots, 2^3\}$ using an internal round of 10-fold cross-validation.

The clustering performance of our approach was assessed by taking into account the stability of the obtained results. The adjusted Rand Index (ARI) [74] was used to quantify the similarity between different clustering results. This index is corrected for grouping by chance, resulting in a more conservative estimation of the overlap. A value equal to one indicates a perfect clustering. We calculated the ARI in a cross-validated fashion, following the previously described cross-validation scheme. However, in our calculations, we took

into account only the clustering stability between training folds. Any pair of training folds shared 80% of the subjects, allowing us to compute how consistently the common subjects were placed in the same clusters despite the variations due to the $\sim 10\%$ difference in the sample composition across folds. In detail, given the optimal $C$ value that was estimated during the inner-fold cross-validation, we trained the model, yielding a clustering of the negative subjects in the training set. This procedure was repeated for all realizations of the 10-fold cross-validation, yielding a set of clusterings of the negative subjects of the respective training sets. Finally, we computed the average pairwise ARI between the estimated clusterings.

Let us note that the classification accuracy and the clustering stability are only surrogate measures that allow us to elucidate the underpinnings of the proposed method. HYDRA does not directly target increased classification accuracy, but instead, it focuses on detecting disease subgroups. Moreover, while clustering stability is desirable, it does not necessarily imply that the estimated clusters correspond to the underlying heterogeneity. Quantitatively evaluating the relevance of the clustering to the intrinsic heterogeneity is in general not feasible. However, in this simulated scenario, the ground truth was available by default. Thus, we calculated the ARI between the estimated clusters and the simulated ones. Moreover, to further assess the performance, we conducted group analysis between the estimated subgroups and the positive class. The derived $p$-value maps allow for the visualization of the estimated clusters and their comparison to the generated ones.

## Comparison with existing methods

To further validate HYDRA, we compared it to common classification and clustering approaches.

As far as classification is concerned, we first compared our method against linear SVMs. In fact, our method is a generalization of the linear SVM framework. By setting the parameter $K$ equal to one, our method reduces to a linear SVM classifier. Parameter selection (*i.e.*, fixing $C$ value) was performed using the same strategy as the one for the proposed framework.

Moreover, because HYDRA establishes a non-linear separation boundary between the two classes, we contrasted its performance against the GRBF kernel SVM. The free parameters were determined through a nested cross-validation strategy. A grid search was performed over the parameter space defined by the regularization parameter $C$ ($C \in \{2^{-5}, \ldots, 2^3\}$) and the parameter $\sigma$ that controls the bandwidth of the RBF kernel ($\sigma \in \{2^{-5}, \ldots, 2^3\}$).

Verifying that HYDRA achieves comparable accuracy with commonly used classifiers, thus retaining discriminative power, is important because discrimination is inextricably tied to the cluster definition. However, the main focus of the method is on discovering clusters in the abnormal cohort. To validate the clustering potential of our framework, we included the performance of the K-means clustering [102] (20 replicates were used). We also examined the potential of the approach that performs classification on top of the clustering results. In particular, we first used K-means to cluster samples from one class and then trained a linear SVM for each cluster. This procedure was performed for both the negative and positive classes. The out of sample prediction was obtained using Eq. 2.2. This approach [65] is termed here **K-means/SVM**. Similar to the previous cases, nested
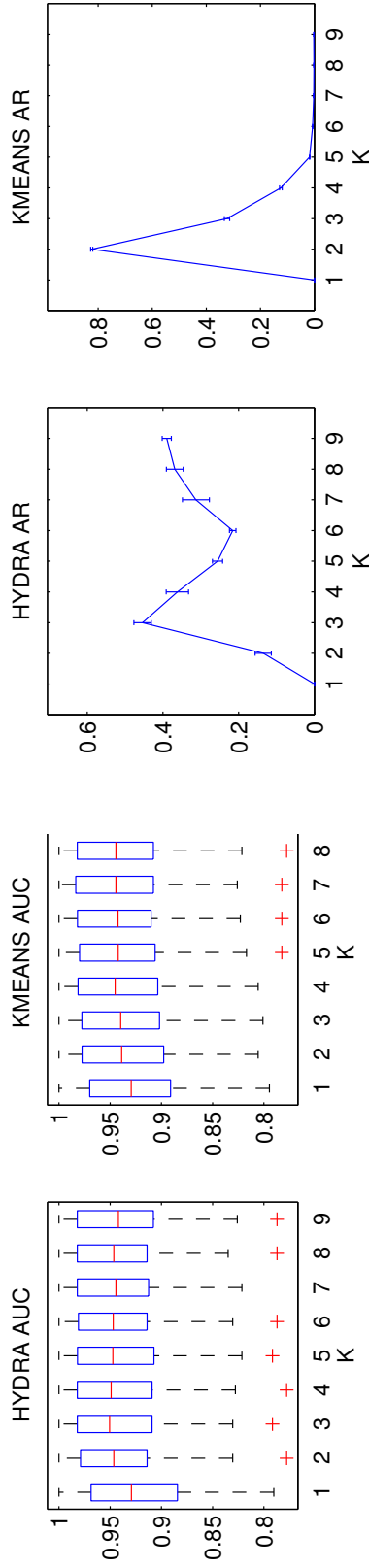
(a) AUC, Left: HYDRA, Right: K-means/SVM



(b) ARI, Left: HYDRA, Right: K-means

Figure 2.4: Simulated data results: (a) Cross-validated AUC for HYDRA (left) and K-means/SVM (right) binary classification. (b) Cross-validated ARI for the clustering result of HYDRA (left) and K-means (right). The results are reported for different values of the parameter $K$. Error bars are centered around the mean and indicate variance. Both the classification accuracy and the cluster stability were maximized at $K = 3$ for HYDRA, agreeing with the intrinsic dimensionality of the heterogeneous group. The classification accuracy obtained by K-means/SVM remained relatively stable for different values of $K$. However, the clustering stability was maximized for $K = 2$, demonstrating that higher reproducibility does not necessarily imply successful heterogeneity detection.

cross-validation was performed for selecting the $C$ parameter. Note also that we run K-means and HYDRA for the same value of the parameter $K$ that varied from one to nine ($K \in \{1, \ldots, 9\}$).

**Results**

| Decoding simulated focal effects | | | | | |
|---|---|---|---|---|---|
| Data | Method | K | AUC | ARI | ARI with Ground Truth |
| Synthetic Data | Gaussian SVM | — | $0.9327 \pm 0.0368$ | — | — |
| | Linear SVM | 1 | $0.9258 \pm 0.0498$ | — | — |
| | HYDRA | 2 | $0.9404 \pm 0.0471$ | $0.1353 \pm 0.1464$ | 0.3487 |
| | | 3* | $\mathbf{0.9423 \pm 0.0460}$ | $0.3620 \pm 0.1514$ | **0.6175** |
| | K-means/SVM | 2* | $0.9347 \pm 0.0484$ | $\mathbf{0.8237 \pm 0.0641}$ | -0.0076 |
| | | 3 | $0.9369 \pm 0.0470$ | $0.3235 \pm 0.0985$ | 0.0233 |

Table 2.1: Table summarizing the results for the simulated dataset. Cross-validated classification accuracy is reported for Gaussian SVM, linear SVM, HYDRA, and K-means/SVM. Cross-validated cluster stability and overlap with the ground truth are reported for HYDRA and K-means. * denotes the value of the parameter $K$ that was chosen based on the cluster stability analysis. All models achieved comparable classification performance in terms of AUC. However, HYDRA was able to correctly identify the ground truth clusters. Note that while K-means achieved the highest reproducibility, it estimated clusters that did not correspond to the generated focal effects.

The results of the cross-validated classification accuracy are reported in Fig. 2.4a. We note that the classification results depend on the value of the parameter $K$. The high dimension and low sample size setting allowed linear SVM to separate the two classes with high accuracy. However, the non-linearity that is introduced by Gaussian SVM, as well as by HYDRA and K-means/SVM, resulted in a slight improvement in the classification performance (see also Table 2.1). We should underline that a statistically significant improvement of the performance was observed only for HYDRA results ($p$-value for t-test comparison between $K = 3$ HYDRA results and linear SVM equals to 0.016). Lastly, we observe that the classification accuracy that was obtained by HYDRA peaks at $K = 3$ and

relatively decreases for higher values of $K$. This indicates that HYDRA was able to estimate the intrinsic dimensionality of the pathological class correctly.

As far as the clustering reproducibility is concerned, we note a significant difference between HYDRA and K-means (see Fig. 2.4b). Note that K-means obtained the highest reproducibility, yet the estimated clusters did not reflect the simulated focal effects. K-means consistently grouped the data into two clusters, while HYDRA segregated the data with higher stability into three subgroups (see also Table 2.1). The importance of this difference was further emphasized by the fact that K-means results were significantly different from the HYDRA clustering. HYDRA clusters overlapped highly with the simulated ones while K-means results did not match the generated subgroups (see Table 2.1). This is because K-means, being blind to class information, was driven by global patterns that were confounded by the variations stemming from covariate effects rather than relevant heterogeneity. On the contrary, HYDRA was able to identify the heterogeneous groups by exploiting patterns that encode directions along which the two groups differ.

To further appraise the differences between the two methods, we report in Fig. 2.3b and Fig. 2.3c the group differences between the positive class and the three subgroups K-means and HYDRA estimated, respectively. By visually comparing them to the group differences for the simulated groups (see Fig. 2.3a), we observe that HYDRA recovered the three modes of differences with high certainty. Contrarily, K-means captured global effects that reflect the overall progression of the simulated pathology (note the relevant increase of the group differences in Fig. 2.3c), instead of teasing out distinct pathological directions.

Our synthetic validation setting provides two key insights. First, while all methods were able to successfully separate the two groups, only HYDRA was able to distinguish

| | Anatomic heterogeneity in Alzheimer's disease | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | AD vs. CN (n = 300) | | | AD subgroups (n = 123) | | | |
| | CN (n = 177) | AD (n = 123) | p-value$^c$ | Group 1 (n = 29) | Group 2 (n = 63) | Group 3 (n = 31) | p-value$^d$ |
| Age (years) | 75.87 ± 5.18 | 74.66 ± 7.39 | 0.09 | 78.93 ± 5.75 | 73.70 ± 7.63 | 72.61 ± 6.85 | 0.0011 |
| Sex (female), n (%) | 87 (49.15) | 62 (50.4) | 0.83 | 8 (27.5) | 32 (50.7) | 22 (70.9) | 0.0031 |
| MMSE | 29.12 ± 1.03 | 23.57 ± 1.88 | 1.01e-100 | 23.96 ± 1.97 | 23.15 ± 1.99 | 24.06 ± 1.34 | 0.0388 |
| APOE $\varepsilon4$ genotype$^a$, n (%) | 48 (27.12) | 82 (66.67) | 1.71e-12 | 21 (72.41) | 38 (60.32) | 23 (74.19) | 0.3121 |
| CSF $A\beta$ (pg/mL)$^b$ | 209.2 ± 53.92 | 143.2 ± 42.29 | 1.468e-14 | 157.3 ± 49.49 | 144 ± 42.59 | 127.9 ± 28.66 | 0.09907 |
| CSF t-tau (pg/mL)$^b$ | 68.21 ± 24.66 | 122.5 ± 58.07 | 2.865e-13 | 97.37 ± 40.17 | 127.4 ± 55.16 | 139.4 ± 71.27 | 0.06547 |
| CSF p-tau (pg/mL)$^b$ | 24.36 ± 13.64 | 40.79 ± 19.11 | 2.102e-09 | 31.26 ± 10.76 | 44.91 ± 23.18 | 42.95 ± 14.4 | 0.03558 |

Table 2.2: Demographic and clinical characteristics of healthy controls (CN), AD patients (left) and the estimated structural MRI driven subtypes of AD (right). MMSE stands for mini mental state examination score. $^a$ – Denotes subjects with at least one APOE $\varepsilon4$ allele present. $^b$ – denotes the cerebrospinal fluid (CSF) concentrations of Amyloid-beta 1 to 42 peptide (A$\beta$), total tau (t-tau), and tau phosphorylated at the threonine 181 (p-tau). $^c$ – p-value estimated using two-tailed t-test to compare AD with CN. $^d$ – p-value estimated using analysis of variance (ANOVA) to compare the three estimated AD subgroups.

between pathological subgroups. Thus, to effectively disentangle disease heterogeneity, one should focus on discriminating patterns rather than global image appearance. Second, and most importantly, analyzing the clustering stability allows for the estimation of the intrinsic dimensionality of the pathological group. Therefore, we adopt hereafter this popular approach [10, 94] to perform model selection.

## 2.5 Experiments using Clinical Data

Having shown the interest of the proposed approach in synthetic data, we next applied our method to data from the Alzheimer's Disease Neuroimaging Initiative[3] (ADNI). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease[4]. Here, our goal was to investigate both the anatomical and the genetic heterogeneity in Alzheimer's Disease.

### 2.5.1 Visualization of Heterogeneity

**Anatomical heterogeneity**

To visualize the neuroanatomical heterogeneity of both the anatomically and genetically-defined disease clusters, voxel-based analyses (VBA) were performed between the controls and patient groups.

---

[3]adni.loni.usc.edu
[4]www.adni-info.org

To perform VBA, MRI scans were first pre-processed using previously validated and published techniques [60]. The preprocessing pipeline includes: (1) alignment to the Anterior and Posterior Commissures plane; (2) skull-stripping [36]; (3) N3 bias correction [137]; (4) tissue segmentation into gray matter (GM), white matter, cerebrospinal fluid, and ventricles using MICO [100]; (5) deformable mapping [121] to a standardized template space [81]; (6) formation of regional volumetric maps called RAVENS maps [33], generated to enable analyses of volume data rather than raw structural data; (7) the RAVENS were normalized by individual intracranial volume to adjust for global differences in intracranial size and smoothed for incorporation of neighborhood information using an 8-mm Full Width at Half Maximum Gaussian filter.

The GM RAVENS were used for all VBA experiments, where a general linear model (GLM) was applied voxel-wise to estimate the disease effect on the voxel value using age and sex as covariates. False Discovery Rate (FDR) correction for multiple comparisons was used for all voxel-based analyses. Only results surviving the statistical threshold at $q < 0.05$ are shown.

**Genetic heterogeneity**

In addition to anatomical heterogeneity, the genetic differences between the subgroups of AD were assessed by performing ANOVA on genetic markers, followed by a Bonferroni test for multiple comparisons. Only results surviving the statistical threshold at $q < 0.05$ are reported.

### 2.5.2 Anatomical Heterogeneity of Alzheimer's Disease

**Participants and MRI data preprocessing**

The first dataset comprises MRI scans that were made available by the ADNI study[5]. T1-weighted MRI volumetric scans were obtained at 1.5 Tesla for 123 AD patients and 177 normal controls (CN) (see demographic information given in Table 2.2).

A low-level representation was extracted by automatically partitioning the MRI scans of all participants into 153 ROIs spanning the entire brain. The ROI segmentation was performed by applying a new multi-atlas label fusion method [37]. The derived ROIs were used as features for all clustering and classification methods.

**Correction for age and sex effects**

To remove age and sex related differences between patient groups while retaining disease-associated neuroanatomical variation, the strategy outlined in [40] was used. Within each cross-validation training fold, we calculated voxel-level $\beta$-coefficients for age and sex in control subjects' ROIs using partial correlation analysis. Then, all subjects were residualized using these coefficients to correct for age and sex effects not attributable to disease related factors.

**Evaluation of results for structural MRI AD data**

Classification results are reported in Fig. 2.5a. The standard linear SVM achieved a highly accurate classification performance (AUC for $K = 1$ is greater than 0.9), which emphasizes the high separability between AD patients and healthy controls. Similar to linear SVM,

---

[5]http://adni.loni.usc.edu/data-samples/mri/

(a)

(b)
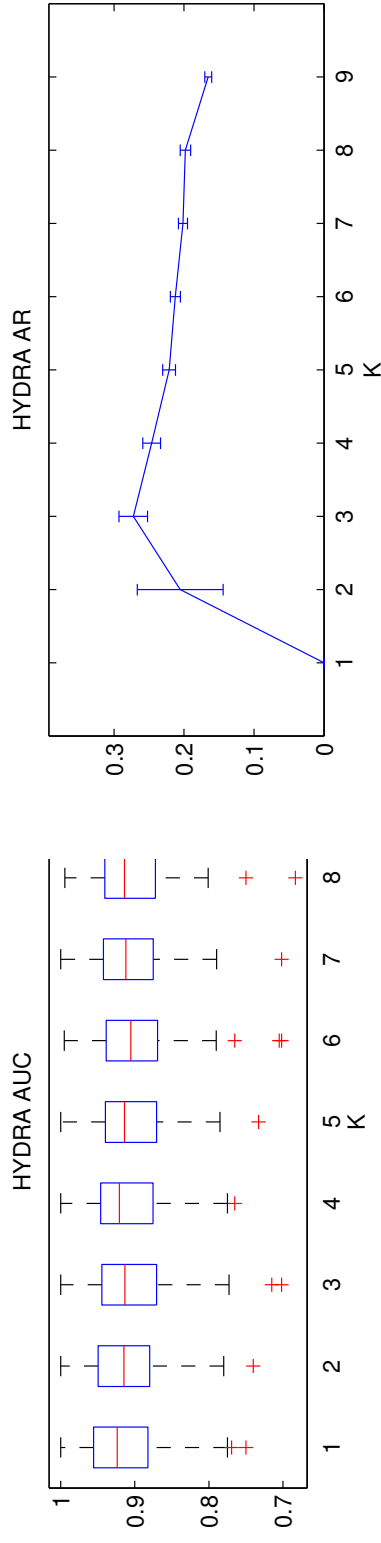
Figure 2.5: Anatomical Data: (a) Cross-validated classification accuracy. (b) Cross-validated cluster stability. Results are reported for different values of the parameter $K$. Error bars are centered around the mean and indicate variance. Classification accuracy remains relatively stable for different values of $K$ (no statistically significant differences between the reported AUC values were observed). Cluster stability exhibits a distinct peak at $K = 3$, suggesting the existence of three distinct disease subgroups.

HYDRA was able to separate the two groups with high accuracy but, contrary to the simulated case, it did not improve on the results of linear SVM. This is most likely because the data were already linearly separable. However, the classification performance of the proposed method remained relatively stable for different values of $K$ (no statistically significant differences between the results were found), demonstrating that HYDRA was able to retain the important discriminative information that is necessary for disease subtype clustering. Furthermore, the stable AUC at $K \geq 2$ may indicate a possible plateau in the AD vs. control classification rate [28]. Lastly, we should emphasize that HYDRA aims to increase the margin with $K$, which is indeed achieved. This has two important implications: i) that there is heterogeneity in the data; and ii) that HYDRA successfully harnesses this heterogeneity to improve the margin.

The clustering stability results are presented in Fig. 2.5b, while the AUC and ARI values for the HYDRA model at $K = 1, 2, 3$ are given in Table 2.3. The stability analysis suggests that three clusters are appropriate for capturing the intrinsic dimensionality for representing the disease heterogeneity. At finer levels (higher values of $K$), these three clusters are partitioned into smaller clusters, giving rise to a hierarchical structure. This observed hierarchy provides further evidence that the data has an inherent structure that HYDRA effectively reveals.

The optimal clustering is visualized through the use of VBA (see Fig. 2.6B, 2.6C and 2.6D). The commonly performed voxel-wise group difference analysis between all healthy subjects and all patients (see Fig. 2.6A) provides the necessary baseline for comparison. It should be noted that the statistical significance of the group comparisons between the controls and the subgroups of AD may be biased due to sample splitting. Thus, these

| Experiment | | Classification/Clustering Performance | |
|---|---|---|---|
| Data | K | AUC | ARI |
| *MRI* | 1 | **0.9149 ± 0.0563** | — |
| | 2 | 0.9123 ± 0.0517 | 0.2054 ± 0.2477 |
| | 3* | 0.9021 ± 0.0572 | **0.2724 ± 0.1430** |
| *Genotype* | 1 | **0.7296 ± 0.1033** | — |
| | 2* | 0.7047 ± 0.1105 | **0.7986 ± 0.2266** |
| | 3 | 0.6990 ± 0.1121 | 0.6412 ± 0.3124 |

Table 2.3: Table summarizing the classification and clustering performance of HYDRA for the experiments using structural MRI and genetic data, respectively. Results are reported for three values of the parameter K. The optimal value of the parameter *K* that was estimated by performing model selection based on clustering stability is denoted by *. The differences in AUC were statistically insignificant between *K* = 1 and *K* = 3 for MRI data (two-tailed t-test *p*-value equals to 0.115) and between *K* = 1 and *K* = 2 for genetic data (two-tailed t-test *p*-value equals to 0.102). This suggests that discriminative signal was preserved, allowing for clinically relevant clusters to be found.

comparisons should serve a qualitative visualization function, rather than a quantitative one. For this reason, we do not state the statistical significance levels for these differences.

We observe that at the *K* = 3 cluster level (see Fig. 2.6) the estimated subgroups are associated with distinct patterns of structural brain alterations: i) diffuse atrophy subtype (see Fig. 2.6B) exhibiting a typical AD pattern, similar to the one that is found by commonly applied monistic VBA (see Fig. 2.6A). This subtype was characterized by atrophy in nearly all cortical regions and increased lesion load in the periventricular white matter; ii) lateral parietal/temporal subtype (see Fig. 2.6C) in which bilateral parietal lobe, bilateral temporal cortex, bilateral dorsolateral frontal lobe, precuneus were mainly involved, and few periventricular white matter lesions were present; iii) medial temporal dominant subtype (see Fig. 2.6D) involving predominantly bilateral medial temporal cortex.

The estimated subgroups were associated with distinct demographic, cognitive and cerebrospinal fluid (CSF) biomarker characteristics. The first subgroup comprised 24% of
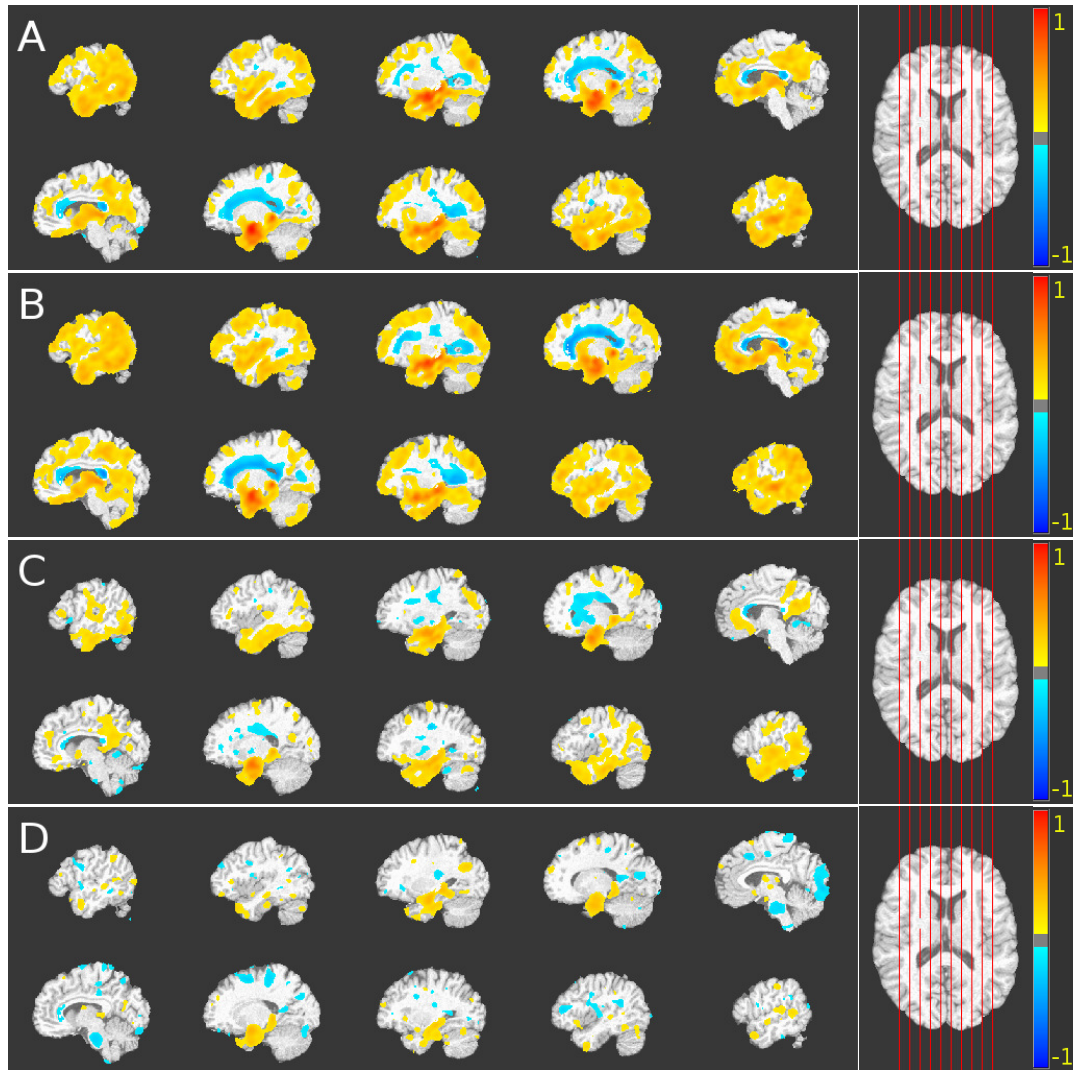
Figure 2.6: Comparison between group differences obtained using commonly applied monistic analysis and the results that were obtained using our method for heterogeneity detection in structural MRI data. The voxel-based analysis was performed using GM RAVENS. Color-maps indicate the scale for the t-statistic. Colder colors indicate relative GM volume increases (CN < pathological population), while warmer colors correspond to relative GM volume decreases (CN > pathological population). Images are displayed in radiological convention. Axial views of the VBA results obtained from GM group comparisons of (A) CN vs. AD; (B) CN vs. first AD subgroup; (C) CN vs. second AD subgroup; and (D) CN vs. third AD subgroup are shown. The first subgroup exhibited diffuse atrophy; the second subgroup was characterized by bilateral parietal lobe, precuneus, and bilateral dorsolateral frontal lobe atrophy, while the third subgroup exhibited bilateral medial temporal dominant atrophy.

AD subjects. It included relatively more male participants (21 males, 8 females) of relatively increased age (78.9 ± 5.75). Members of this group achieved a Mini-Mental State Examination (MMSE[6]) score of 23.97 ± 1.97, while the frequency of APOE $\varepsilon$4 allele carriers was 72.4%. In addition, this group had the highest CSF Amyloid-beta 1 to 42 peptide ($A\beta$) concentration, 157.3 pg/mL, and the lowest CSF total tau (t-tau) and CSF tau phosphorylated at threonine 181 (p-tau) concentrations, 97.3 pg/mL and 31.2 pg/mL, respectively, on average compared to the other subgroups.

The second subgroup was the largest one, consisting of 51% of AD subjects, 60.32% of whom are APOE $\varepsilon$4 carriers. Both sexes were nearly equally represented (31 males and 32 females), having a mean age of 73.7 years (±7.63 standard deviation). Its members performed relatively worse regarding MMSE (23.16 ± 1.99). The average CSF p-tau concentration for this group was the highest compared to the other subgroups at 44.9 pg/mL.

The last subgroup included the 25% of AD patients. Contrary to the previous subgroup, it was dominated by females (9 males and 22 females) of relatively younger age (72.62 ± 6.85) with a rather higher frequency of APOE $\varepsilon$4 allele carriers (74.19%). MMSE performance of this subgroup was 24.06 ± 1.34. The CSF $A\beta$ concentration was the lowest for this group at 127.9 pg/mL while the CSF t-tau concentration was the highest at 139.4 pg/mL, on average, compared to the other subgroups.

Comparing the genetic profiles of these three subgroups of AD yielded further insight on the differences between the pathologies exhibited by each subgroup. One-way ANOVA was performed for each of the single nucleotide polymorphisms (SNPs) identified in two recent genome-wide association studies that reported loci associated with AD

---

[6]MMSE is a quantified clinical assessment for dementia [49]

[93] and cognitive decline [135] (see C). Three SNPs were statistically significantly different: rs10948363, which is related to gene CD2AP, rs11023139, which is related to gene SPON1, and rs7245858, which is related to gene LOC390956.

For SNP rs10948363, which is related to gene CD2AP, 58% of the first subgroup and 74% of the third subgroup were carriers of the minor G allele, while 39% of the second subgroup were carriers of this risky allele.

For SNP rs11023139, which is related to gene SPON1, 29% of the first subgroup were carriers of the minor A allele, while 2% of the second subgroup and 11% of the third subgroup were carriers of this allele.

Lastly, for SNP rs7245858, which is related to gene LOC39095, 23% of the first subgroup were carriers of the minor A allele, while 2% of the second subgroup and 4% of the third subgroups were carriers of this allele.

### 2.5.3   Genetic Heterogeneity of Alzheimer's Disease

**Genotype data**

The second dataset comprises genotypes for 103 AD patients and 139 normal controls (see demographic information in Table 2.4), obtained from the ADNI study[7]. ADNI genotyping is performed using the Human610-Quad Bead-Chip (Illumina, Inc., San Diego, CA) which results in a set of 620,901 single nucleotide polymorphisms (SNPs) and copy number variation markers (for details see [133]).

Due to the weak or spurious signal in most of the genome, we opted to only use SNP loci that were associated with Alzheimer's disease or cognitive decline in recent large scale

---

[7]http://adni.loni.usc.edu/data-samples/genetic-data/

Genetic heterogeneity in Alzheimer's Disease

| | AD vs. CN ($n = 243$) | | | AD subgroups ($n = 103$) | | |
|---|---|---|---|---|---|---|
| | CN ($n = 139$) | AD ($n = 103$) | $p$-value[c] | Group 1 ($n = 68$) | Group 2 ($n = 35$) | $p$-value[d] |
| Age (years) | 76.19±4.85 | 75.04±7.59 | 0.15 | 74.46±6.56 | 76.18±9.27 | 0.27 |
| Sex (female), $n$ (%) | 62 (44.60) | 49 (47.57) | 0.64 | 33 (48.52) | 16 (45.71) | 0.78 |
| MMSE | 29.16±1.01 | 23.54±1.95 | 1.85e-80 | 23.60±1.88 | 23.42±2.10 | 0.67 |
| APOE $\varepsilon 4$ genotype[a], $n$ (%) | 36 (25.89) | 72 (69.90) | 9.56e-13 | 67 (98.52) | 5 (14.28) | 8.96e-33 |
| CSF A$\beta$ (pg/mL)[b] | 206.1 ± 54.61 | 147.2 ± 43.82 | 1.093e-09 | 133.6 ± 28.47 | 174.2 ± 56.04 | 0.0004245 |
| CSF t-tau (pg/mL)[b] | 71.11 ± 24.89 | 121.9 ± 59.62 | 6.456e-09 | 129.5 ± 57.31 | 107 ± 62.71 | 0.1738 |
| CSF p-tau (pg/mL)[b] | 25.02 ± 13.69 | 40.7 ± 19.86 | 1.026e-06 | 42.58 ± 19.92 | 36.95 ± 19.7 | 0.3051 |

Table 2.4: Demographic and clinical characteristics of healthy controls, AD patients (left) and the estimated genetic-driven sub-types of AD (right). [a] – Denotes subjects with at least one APOE $\varepsilon 4$ allele present. [b] – denotes the cerebrospinal fluid (CSF) concentrations of Amyloid-beta 1 to 42 peptide (A$\beta$), total tau (t-tau), and tau phosphorylated at the threonine 181 (p-tau). [c] – $p$-value estimated using two-tailed t-test to compare AD with CN. [d] – $p$-value estimated using analysis of variance (ANOVA) to compare the two estimated AD subgroups.
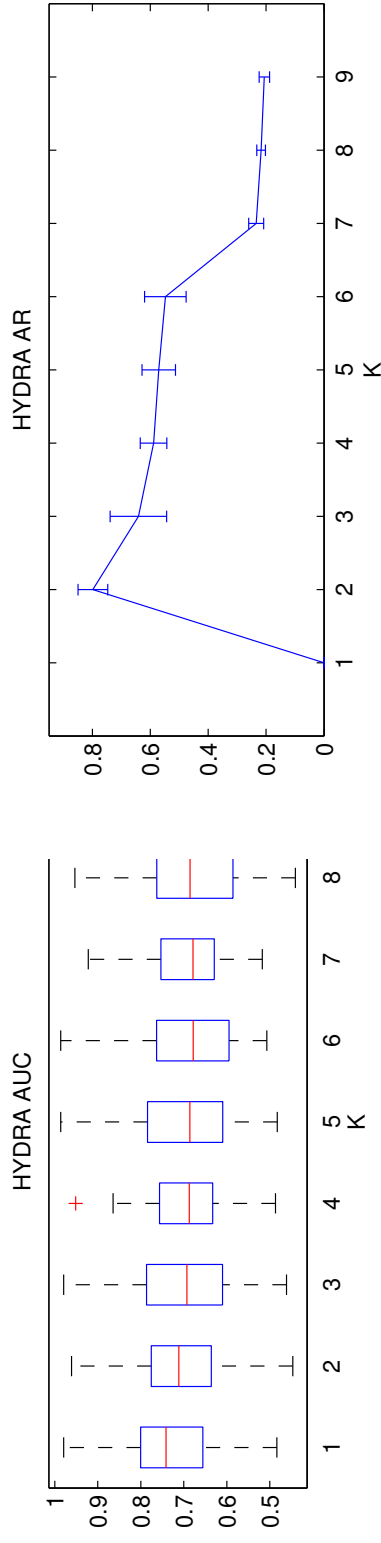
(a)

(b)

Figure 2.7: Genetic Data: (a) Cross-validated classification accuracy. (b) Cross-validated cluster stability. Results are reported for different values of the parameter $K$. Error bars are centered around the mean and indicate variance. Classification accuracy slightly decreases. However, the results for $K = 1$ and $K = 2$ were not statistically significant different. Cluster stability exhibited a distinct, high peak at $K = 2$, suggesting the existence of two distinct disease subgroups.

genome wide association studies [93, 135]. This resulted in a reduced set of 66 SNPs (see table in C)that were represented through the use of two binary variables encoding the presence of major-major or major-minor alleles, thus raising the total number of features to 132.

**Evaluation of results for genotype AD data**

Classification results are reported in Fig. 2.7a. The standard linear SVM discriminated fairly between healthy controls and AD patients (AUC for $K = 1$ equals to 0.72). Compared to the result that was obtained using imaging features, this highlights the difficulties associated with disease classification in the genotype domain. HYDRA was able to separate the two groups with similar accuracy for $K = 2$ (AUC equals to 0.70). The classification accuracy dropped for higher values of $K$. However, the difference between the results for $K = 1$ and $K = 2$ was statistically insignificant ($p = 0.10$).

The clustering stability results are presented in Fig. 2.7b, while the AUC and ARI values for the HYDRA model at $K = 1, 2, 3$ are given in Table 2.3. The stability analysis suggested that two clusters are appropriate for capturing the intrinsic dimensionality for representing the genetic heterogeneity associated with AD. Similar to the anatomically-driven clustering results, these two clusters are successively partitioned to smaller clusters for higher values of $K$, showing a hierarchical organization. This suggests that the data has the structure that HYDRA reveals.

The optimal genotype clustering is visualized by contrasting the imaging phenotypes of the estimated subgroups against the healthy control population through VBA (see Fig. 2.8A and Fig. 2.8B).
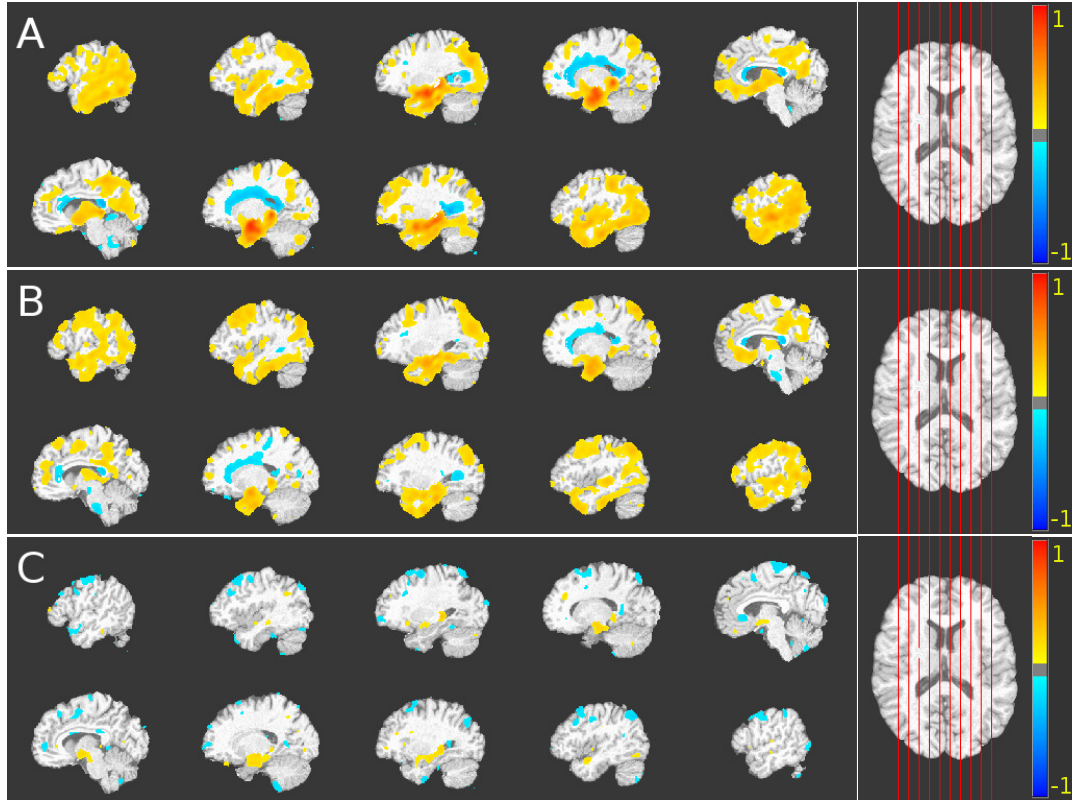
Figure 2.8: Comparison between group differences obtained using commonly applied monistic analysis and the results that were obtained using our method for heterogeneity detection in genetic data. The voxel-based analysis was performed using GM RAVENS. Color-maps indicate the scale for the t-statistic. Images are displayed in radiological convention. Axial views of the VBA results obtained from GM group comparisons of (A) CN vs. first AD subgroup; (B) CN vs. second AD subgroup; and (C) first AD subgroup vs. second AD subgroup are shown. For (A) and (B), colder colors indicate relative GM volume increases (CN < AD subgroups), while warmer colors correspond to relative GM volume decreases (CN > AD subgroups). Similarly, for (C), warmer colors indicate relative GM volume increases (first AD subgroup < second AD subgroup), while colder colors correspond to relative GM volume decreases (first AD subgroup > second AD subgroup). Both groups exhibit atrophy in the temporal lobe and posterior medial cortex while white matter lesions are present in the periventricular area. However, the first AD subgroup, which mainly comprises APOE $\varepsilon$4 carriers, is characterized by significantly more hippocampus and entorhinal cortex atrophy and less superior frontal lobe atrophy.

We observe that at the $K = 2$ cluster level, the estimated subgroups were associated with distinct patterns of structural brain alterations: i) increased temporal lobe atrophy subtype (see Fig. 2.8A) including posterior medial cortex atrophy and increased white

matter lesion load; ii) increased superior frontal lobe atrophy subtype (see Fig. 2.8B) including temporal lobe atrophy and periventricular white matter lesions.

The first subgroup exhibited reduced GM volumes in the hippocampus and entorhinal cortex (Fig. 2.8A), while the second subgroup exhibited reduced GM volumes in the superior frontal lobe (Fig. 2.8B). The difference between the brain images in the two subgroups is visualized in Fig. 2.8C.

The sex and age composition of the two estimated subgroups was similar for both cases. The proportion of the females in the first subgroup was 48.52%, while for the second one was 45.71% (see also Table 2.4). The average age of the first subgroup was 74.5, while for the second one was 76.2 years old.

In addition to anatomical differences, the two subgroups exhibited significantly different levels of APOE $\varepsilon$4 allele and CSF biomarkers. While the first subgroup was composed of 98% APOE $\varepsilon$4 carriers, only 14% of the second subgroup were APOE $\varepsilon$4 carriers. Also, the first group had lower $A\beta$ concentration, 133.6 pg/mL, and higher t-tau and p-tau concentrations, 129.5 pg/mL and 42.5 pg/mL, respectively, on average compared to the second subgroup.

Further analysis of the genetic differences between the two subgroups yielded two additional loci of interest. While 32% of the first subgroup were carriers of the risk related A allele of the SNP rs6656401 (related to gene CR1), 49% of the second subgroup was composed of carriers of this allele.

The second locus that differed between the two subgroups was the SNP rs6733839, which is related to gene BIN1. While 72.06% of the first subgroup consisted of risk related C allele carriers of rs6733839, 85.71% of the second group comprised carriers of this allele.

However, similar to the voxel-based analysis of the differences between the subgroups of AD patients, these statistical findings should be approached with care as there might be bias due to sample splitting. The statistical power needed to make a definite statement about the genetic differences between the subtypes of AD may require a much higher sample size.

## 2.6    Discussion & Conclusion

**Synopsis**

In this chapter, we presented HYDRA, a method for disentangling heterogeneity in a principled semi-supervised machine learning framework. HYDRA aims to generalize the basic assumption of computational neuroimaging studies from a single separating pattern to many patterns, thus addressing one of the major challenges that characterizes many studies, namely the presence of heterogeneity. HYDRA attempts to find patterns associated with the underlying disease process, or more generally with the difference between two groups. These different patterns could potentially identify different dimensions of the underlying disease process and hence lead to diagnostic subcategories.

The proposed approach seamlessly integrates clustering and discrimination in a coherent framework by solving for a non-linear classifier that bears common geometric properties with convex polytopes. Discrimination is achieved by constraining one class in the interior of the polytope, while at the same time maximizing the margin between examples and class boundary. On the other hand, clustering is performed by associating disease samples to different faces of the polytope and hence to different disease processes. Thus,

each face of the polytope informs us about the distinct foci of disease effects that distinguish the patients from the healthy control subjects. This coupling between clustering and classification allows for segregating patients based on disease patterns rather than global anatomy.

In our experiments, we demonstrated the ability of the proposed approach to discern disease foci in both synthetic and clinical datasets without undermining its predictive power. Moreover, our method is endowed with improved generalization performance due to its maximum margin property of the method and the low complexity of the model (compared to standard non-linear classifiers, *e.g.*, Gaussian kernel SVM). The latter allows it to efficiently handle small sample size high dimensionality data that are commonly encountered in neuroimaging studies by exploiting the dual model representation and operating in the inner product space.

**Model selection**

Choosing an appropriate number of hyperplanes, or corresponding disease subtypes is an important and difficult model selection question. The difficulty is underlined by the fact that there is no ground truth available against which one may test a clustering result. However, we presented a strategy based on examining the clustering stability [10, 94]. The basic premise behind this strategy is that as one gets closer to the intrinsic dimensionality of the pathological group, the clustering algorithm should obtain similar results for different datasets generated by sampling the initial population. The group structure should remain relatively stable accounting for the fact that the datasets have been generated by the same factors.

**Anatomical heterogeneity of AD**

Applying the proposed framework to structural imaging data from ADNI, resulted in the definition of three AD subgroups. Our results largely agree with a recent study employing surface-based morphometry to study AD heterogeneity based on cortical thickness [118] and bear similarity to the subtypes that were recently identified in a pathologic study based on the distribution and density of neurofibrillary tangles [112]. The first subgroup is similar to the diffuse atrophy subtype reported in [118] and the typical AD group in [112]. The second subgroup is comparable to the parietal dominant in [118] and the first subtype in [112]. The third subgroup maps to the medial temporal subtype of [118] and the third group of [112].

The agreement of the results, despite the differences in the design of the studies, emphasizes the fact that AD should be considered as a neuroanatomically heterogeneous disease, characterized by multiple pathological dimensions. Among the pathological dimensions revealed in this study, only the first one (Fig. 2.6B) bore important resemblance with a typical AD pattern involving signature AD regions, while the other two (Fig. 2.6B and Fig. 2.6C) exhibited distinct pathological patterns. These dimensions may reflect distinct pathways leading to AD, associated with distinct disease processes that may constitute potential therapeutic targets.

Aiming to elucidate the recovered pathological dimension of AD further, we found that the anatomically defined clusters exhibit significant differences in their genotypes, demographic characteristics and CSF biomarker distributions.

The first subgroup comprised more male participants of relatively older age. 72.4% of its members were APOE $\varepsilon 4$ allele carriers, while SNPs rs11023139 and rs7245858 were

carried relatively more by members of this subgroup than members of the other two; 29% of the first subgroup were carriers of the minor A allele for rs11023139 and 23% of the first subgroup were carriers of the minor A allele for rs7245858, respectively (see Sec. 2.5.2). This subgroup was characterized by the most widespread pattern of atrophy, yet the most normal CSF biomarker levels. Moreover, the cognitive performance of its members was comparable to one of the other subgroups. The older age of the group, the relatively more normal levels of CSF biomarkers as well as the protective nature of rs11023139, which has been associated with a slower rate of cognitive decline [135], suggest a protracted disease progression. The possible long disease progression may have allowed for compensatory mechanisms to develop resulting in a cognitive performance that is comparable to the other groups despite the extended atrophy.

The second subgroup was the largest one (comprising 51% of AD subjects), with nearly equal sex proportions. However, it comprised proportionally fewer APOE $\varepsilon$4 carriers (60.32%), fewer carriers of the risky allele of SNP rs10948363 (39%), and almost no carriers of the minor A allele of SNP rs10948363 (2%) and SNP rs7245858 (2%). This was the group whose members performed worse regarding MMSE.

The third subgroup included predominantly females of relatively younger age. Most of the patients (74.19%) were APOE $\varepsilon$4 allele carriers, while also 74% of them were carriers of the minor G allele of the SNP rs10948363, whose corresponding gene is CD2AP. CD2AP is a scaffolding protein that is involved in cytoskeletal reorganization and intracellular trafficking [41] and has been previously associated with late-onset AD [113]. Moreover, a direct link between CD2AP and amyloid $\beta$ toxic effects has been noted in yeast, nematodes, and rat cortical neurons after study of the role of several genes in amyloid $\beta$ and tau pathways

[145]. This along with the fact that this group exhibits the most abnormal levels of CSF t-tau and $A\beta$ concentration may explain why members of this group are diagnosed as AD, despite being of younger age and exhibiting more focal atrophy. The sex difference in the population of this subgroup may result from the gender difference in the AD-promoting effect of the APOE genotype [122]. Given that APOE $\varepsilon$4 preferably affects medial temporal lobe structures, women may have a more vulnerable medial temporal cortex than men, giving rise to this specific subtype.

**Genetic heterogeneity of AD**

Applying the proposed framework to genetic data from ADNI, resulted in the identification of two AD subgroups. These groups were essentially dichotomized based on the presence of APOE $\varepsilon$4 allele (98% of the members of the first subgroup carry it, while only 14% of the second subgroup do). However, the two groups exhibit additional genetic differences, as well as anatomical differences and distinct distributions of CSF biomarkers.

Genetic differences were found for the SNP rs6656401 (related to gene CR1) and the SNP rs6733839 (related to gene BIN1). Genetic variations at CR1 have been associated with the risk of cerebral amyloid angiopathy and decreased entorhinal cortex volume [14, 17]. Increased expression of the BIN1 gene has been recently implicated in modulating tau pathology [22], while BIN1 has also been associated with entorhinal and temporal pole cortex thickness [14].

Anatomical differences were mainly found in the hippocampal and entorhinal cortex, where the first group was characterized by significantly more atrophy. The anatomical differences between the subgroups may be explained by the genetic variations. APOE $\varepsilon$4

has been related to increased atrophy in hippocampus [67, 71], entorhinal [80] and medial frontal cortex [45]. Given that, the first subgroup is expected to exhibit more atrophy in these areas.

The two groups were characterized by differences in the distribution of the CSF biomarkers. This difference was more significant for the CSF $A\beta$, which was significantly reduced in the first group. This difference may also be attributed to the effect of APOE $\varepsilon4$, which has been previously associated with reduced levels of CSF $A\beta$ and t-tau[124, 140].

While the dominant presence of APOE $\varepsilon4$ in the first subgroup provides the means to interpret the anatomical and CSF biomarker differences between the two subgroups, the relatively higher expression of the SNPs related to CR1 and BIN1 genes in the second subgroup (where APOE $\varepsilon4$ allele is less expressed) may be an indication that these genes may be part of an alternative pathway for AD pathogenesis in the absence of APOE $\varepsilon4$ expression. The atrophy exhibited by the second subgroup in the entorhinal cortex seen in Fig. 2.8B) may be a product of CR1 expression since APOE $\varepsilon4$ is mostly absent in this subgroup. While this hypothesis remains to be validated, this underlines the value of data-driven, multivariate, exploratory techniques in forming new hypotheses.

**Limitations and future work**

There are some limitations to this work. First, the lack of ground truth for the clinical datasets does not allow us to quantitatively validate the proposed method. However, on the one hand, when AD patients were clustered based on imaging information, the identified patterns of abnormality aligned well with findings based on neuropathology reported in [112] and the subtypes defined based on cortical thickness in [118]. Moreover, the

anatomically defined subgroups also exhibited genetic differences, which provides additional evidence for the validity of the obtained clustering. On the other hand, when clustering based on genetic information, we identified subpopulations that exhibited meaningful anatomical differences. In summary, our results were consistent with the existing picture of pathological neurodegeneration and the function of the related SNPs.

Nevertheless, the sample size that is necessary for drawing reliable conclusions about the full extent of heterogeneity of AD may be higher than what was analyzed. In general, we were able to demonstrate the presence of heterogeneity in AD given the ADNI dataset. However, to be able to elucidate disease heterogeneity and map the distinct pathological processes that drive it, a wider sampling of the patient population probed in a multiparametric fashion may be required.

Another limitation of this work is that the diseased population was studied by using either structural imaging data or genetic information. While this demonstrates the ability of the proposed framework to handle both imaging and non-imaging data, including additional information (*e.g.,* amyloid PET imaging, tau imaging, cerebrospinal fluid biomarkers, etc.) would be beneficial in better characterizing the dimensions and extent of heterogeneity. Nonetheless, HYDRA cannot currently handle multiple sources of information. This could be made possible by extending HYDRA through the adoption of multiple kernel techniques [7]. Different kernels could be employed to encode different sources of information, allowing for their seamless integration. This extension could make HYDRA even more general, allowing its application to other exploratory problems, such as characterization of the breast cancer heterogeneity and the analysis of abnormal tissue subtypes, without being limited to the clustering of brain images.

We should note that the estimation of the subpopulations may be influenced by confounding variations due to age and sex differences. In its current form, our method does not explicitly take into account this case. Instead, we circumvent this by performing univariate covariate correction before feeding the data to our method. To tackle this shortcoming, we are currently working on extending the proposed method by explicitly modeling the effect of covariates within a unified clustering framework. However, the effect of the covariates also renders prohibitive the usage of the classification model to interpret the weight vectors of the hyperplanes (as explained in [68]). We circumvent this by performing voxel-wise group analysis between the inferred patient clusters. However, the interpretation of the group comparison results should be made with care since the significance of the comparison may be biased due to the sample splitting. The voxel-based comparisons should serve only as a qualitative tool and not as a quantitative one. Furthermore, to avoid the circularity of assessing group differences using the same features that the groups are clustered by, we have assessed group differences using features that have not been used in the clustering. Namely, we have assessed the genetic and demographic differences between the anatomic subtypes of AD and the anatomic and demographic differences between the genetic subtypes of AD.

A possible extension of our method is towards handling regression and longitudinal studies. This could allow us to elucidate the complex nature of spatiotemporal disease dynamics as well as to reveal varying paths of normal progression. Lastly, it is straightforward to derive a one-class version of HYDRA, analogous to the work of [132], to detect and subtype outliers among controls. This could potentially shed light on the heterogenous nature of healthy phenotypes.

**Conclusion**

HYDRA aims to separate two groups by deriving a non-linear classification boundary that is constructed by using multiple linear hyperplanes. The constructed polytope allows for the revealing heterogeneity by assigning subgroups of patients to different hyperplanes. HYDRA is general; it can handle imaging and non-imaging data and can find applications in exploratory analyses other than the clustering of brain images. We evaluated the performance of the method in simulated data, providing insight into its workings. Furthermore, we applied HYDRA to structural imaging and genetic dataset from ADNI, revealing disease subtypes that are consistent with the existing picture of pathological neurodegeneration and the function of the related SNPs. These results demonstrate the potential of our approach in teasing out heterogeneity.

# Chapter 3

# Inference through optimal spatial filtering: MIDAS

## 3.1 Introduction

Voxel-wise statistical mapping is a widely used technique in neuroimaging within cross-sectional studies. Its overarching goal is to generate maps that represent structural or functional patterns associated with either group differences or with non-imaging variables. This is typically performed by spatially aligning imaging measurements from a set of images, smoothing them using a fixed-size Gaussian kernel, and comparing them using mass-univariate voxel-wise statistical tests. Depending on the type of imaging features, these techniques may fall under the category of voxel-based morphometry (VBM) [159, 60, 3, 33, 78, 90, 134, 13, 59, 77, 19, 107], deformation-based morphometry (DBM) [5, 25, 26], or tensor-based morphometry (TBM) [144, 51, 138, 98, 23, 72]. These methods do not require a priori definition of regions of interest and have the advantage of exam-

ining the brain as a whole. As a consequence, they offer an automated, data-driven, and unbiased way to assess brain structure and function comprehensively.

One major limitation of mass-univariate techniques is that they ignore multivariate relations in the data. Additionally, the commonly applied local smoothing may obscure the effects of interest. Smoothing the data is necessary to ensure that the assumptions underlying the theory of Gaussian random fields are met, and to account for registration errors. Perhaps most importantly, smoothing is used to amplify the signal and reduce the noise before performing statistical analyses and can lead to a dramatic increase in sensitivity to detecting effects of interest. However, smoothing is typically not adapted to the scale and shape of the signal of interest (e.g., activation, atrophy, neuropathology), which is necessary to achieve high sensitivity and specificity in group comparisons or regressions with non-imaging variables. If the smoothing kernel is too small, noise and limited pooling of regional signal can severely reduce the statistical power of the ensuing statistical maps. Conversely, if the kernel is too large, the spatial specificity of the maps is reduced, leading to false conclusions about the origin of the effect of interest. Additionally, a kernel that is too large may also decrease the statistical power for detecting effects of interest by smearing them out through the introduction of information from regions that display no effect of interest. As a consequence, selecting the appropriate kernel size is a challenging task [79, 164]. In practice, this is performed in an empirical, or ad hoc fashion.

Towards addressing these limitations, information-based brain mapping techniques have become increasingly popular in recent years. These techniques use pattern classifiers to harness the rich multivariate information present in the interactions across many voxels to obtain more powerful statistical maps. These approaches were popularized by

the introduction of the searchlight methods [89, 123, 1]. Searchlight commonly applies local discriminative classifiers and creates an information map by assigning each searchlight's classification accuracy to its center voxel. In some variants of searchlight, Monte-Carlo sampling and combining information across overlapping neighborhoods is used to increase stability [15]. Despite its appealing multivariate nature, this strategy does not appropriately encode the importance of each voxel as it effectively ignores its contribution to the discriminative pattern. This may lead to important interpretation errors in practice. [44] demonstrated that searchlight methods might fail to detect informative voxels, or could misclassify voxels as informative, unless the searchlight region sufficiently covers, or matches the underlying pattern. Specifically, it is possible for voxels in the searchlight map to be categorized as significant, not because they are informative, but because they are at the center of a searchlight that contains the informative voxels. It is also possible to detect weakly-informative voxels when they are sufficiently numerous.

Towards addressing this limitation, a more refined way to characterize each voxel's importance was proposed by [161, 162] in their framework for optimally-discriminative voxel-based analysis (ODVBA). In ODVBA, non-negative discriminative projection was employed regionally to estimate the direction that best discriminates between two groups. Given this direction, the statistic of each voxel was assessed by taking into account the discrimination power of the voxel in terms of the pattern seen in its neighborhood. However, ODVBA was limited only in group-comparison settings, not being able to address regression tasks. More importantly, to obtain a statistical parametric map of group differences, ODVBA requires computationally expensive permutations tests.

To tackle these shortcomings, we propose a novel statistical method for cross-sectional

studies, termed MIDAS, which originates from *regionally linear **m**ultivariate **d**iscriminative **s**tatistical mapping*. Our goal is to efficiently obtain highly specific and sensitive brain maps with applications in structural and functional imaging. MIDAS seeks to increase statistical power by combining the signal from all voxels that constitute the effect of interest. Towards this end, it aims to locally determine the shape and spatial extent of the effect/signal of interest by fitting least squares SVM to a large number of overlapping neighborhoods, which fully and redundantly cover the brain. In this way, the effect of interest is estimated as the pattern that best discriminates between two groups, or predicts the variable of interest in regression designs. This pattern is equivalent to local filtering by an optimal kernel whose coefficients define the optimally discriminative/predictive pattern. By combining information from all neighborhoods that contain a given voxel, we produce voxel-wise statistics. These statistics are calculated by summing the contributions of each voxel to the estimated local hyperplanes and normalizing them by the sum of the respective SVM margins. In other words, informative voxels are defined as ones that contribute significantly to the discriminative direction of SVMs, which in turn, discriminate between two groups, or predict a variable of interest with a margin as large as possible. Critically, motivated by recent advances in deriving statistical significance maps for SVM classification [54, 55], we derive an analytical approximation of the null distribution of the estimated statistics. This allows us to effectively estimate voxel-wise p-value maps at a dramatic speed-up compared to permutation tests.

We validated the proposed framework against mass univariate techniques, as well as multivariate pattern analysis methods including searchlight, ODVBA, and SVM-based statistical significance maps. We created simulated data by introducing synthetic atrophy to

structural brain scans of healthy subjects to quantitatively evaluate the performance of the method. Quantitative evaluations were performed by assessing the sensitivity and specificity of the statistical significance maps in relation to the ground-truth regions. Moreover, we used data from a task-based functional magnetic resonance imaging study to test MIDAS. This dataset consisted of brain activation maps of subjects who took part in a forced choice deception experiment, where the groups were defined by truth-telling versus lying tasks [34, 95]. Due to the absence of ground-truth, the methods were quantitatively analyzed by measuring split sample reproducibility. Our experimental results indicate that the proposed method outperforms the commonly used univariate and multivariate algorithms in terms of sensitivity and specificity, as well as reproducibility. Lastly, the regression ability of MIDAS was demonstrated using a structural magnetic resonance imaging study of the cognitive performance of mild cognitive impairment subjects [108, 146]. In this setting, MIDAS was also able to yield highly sensitive maps compared to other state of the art methods.

The remainder of this chapter is organized as follows. In Section 3.2, we detail the proposed approach. In Section 3.3, we first experimentally validate MIDAS using simulated data and then apply MIDAS to data from functional and structural neuroimaging studies. We discuss the results in Section 3.4, while Section 3.4 concludes the chapter with our final remarks.
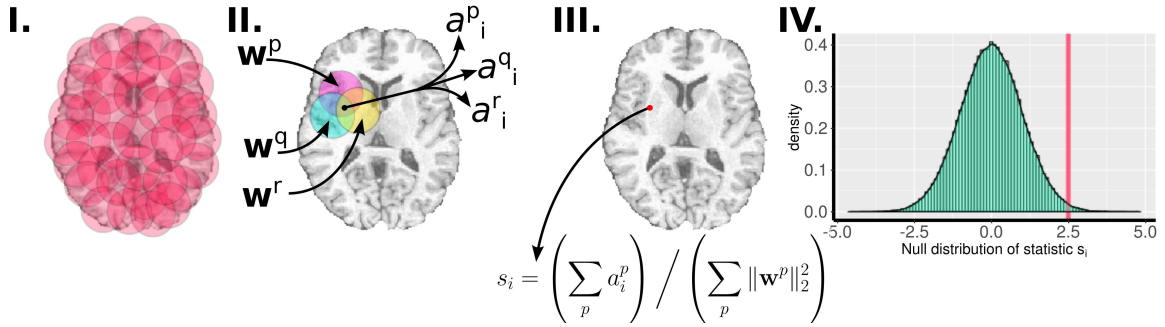
Figure 3.1: Overview of MIDAS: I) Neighborhoods are uniformly sampled such that the brain volume is sufficiently covered; II) Local discriminative analysis is performed on neighborhoods yielding weight vectors $\mathbf{w}$; III) A voxel-wise statistic is computed using the weight vectors; and IV) statistical significance is assessed through analytically approximating the null distribution of the voxel-wise statistic.

## 3.2 Method

### 3.2.1 Overview

Multivariate inference using discriminative adaptive smoothing (MIDAS) is a group analysis and regression framework that integrates a large number of regional discriminant, or regression, pattern analyses to obtain a voxel-wise statistical map analogous to those obtained via the general linear model (see Fig. 3.1). MIDAS scans the imaging volume using a sufficiently large set of overlapping neighborhoods (Fig. 3.1 I), and performs regional discriminative analysis that yields weight vectors (denoted by $\mathbf{w}$) (Fig. 3.1 II). The statistic for a particular voxel is computed by summing the weights corresponding to the voxel in all of the neighborhoods it resides in, and normalizing by the sum of the discriminative power of the respective neighborhoods (Fig. 3.1 III). Finally, the p-value corresponding to the voxel statistic is analytically obtained by approximating permutation tests (Fig. 3.1 IV).

### 3.2.2 Least squares support vector machine

MIDAS is based on the least squares support vector machine (LS-SVM) [141] to perform local discriminative analysis. LS-SVM is an ideal base learning method for the MIDAS framework as it can readily handle both classification and regression problems while admitting a closed form solution. Let $\mathbf{X} \in R^{n \times d}$ denote the $n$ by $d$ matrix that contains $d$-dimensional imaging features from $n$ independent subjects arranged row-wise. Likewise, let $\mathbf{y} \in R^n$ denote the vector that stores the clinical variables of the corresponding $n$ subjects. LS-SVM aims to relate the imaging features $\mathbf{X}$ with clinical variables $\mathbf{y}$ via a weight vector $\mathbf{w}$ and a bias term $b$ by optimizing the following objective:

$$\min_{\mathbf{w},b,\epsilon} \frac{\|\mathbf{w}\|_2^2}{2} + c\frac{\|\epsilon\|_2^2}{2}$$

$$\text{subject to } \mathbf{X}\mathbf{w} + \mathbf{1}b = \mathbf{y} + \epsilon. \tag{3.1}$$

This formulation describes a generalized fitting setting where the predictors captured in $\mathbf{X}$ can be used to predict the responses $\mathbf{y}$. The responses $\mathbf{y}$ can be either binary, yielding a group difference setting or they can be continuous, yielding a regression setting. Here, $\mathbf{w}$ is a d-dimensional vector that contains the weights given to each of the $d$ features for the fitting task, while $\epsilon$ is an n-dimensional vector providing slack for errors. Furthermore, $c$ is a hyper-parameter than controls the closeness of fit. The weight vector $\mathbf{w}$ can be solved in closed form by satisfying the Karesh-Kunh-Tucker (KKT) conditions, leading to a solution in the form:

$$\mathbf{w} = C\mathbf{y}, \tag{3.2}$$

where the solution for the **C** matrix is given in 3.2.3. Note that this is a linear solution in the **y** vector. Being able to express the solution vector in a closed linear form is important because it allows us also to express the null distribution of **w** analytically and without the need for very costly random permutations of the clinical variables **y**.

### 3.2.3 Optimization

The Lagrangian for LS-SVM is:

$$L(\mathbf{w}, b, \boldsymbol{\epsilon}, \boldsymbol{\lambda}) = \frac{\|\mathbf{w}\|_2^2}{2} + c\frac{\|\boldsymbol{\epsilon}\|_2^2}{2} + \boldsymbol{\lambda}^T (\mathbf{X}\mathbf{w} + \mathbf{1}b - \mathbf{y} - \boldsymbol{\epsilon}), \tag{3.3}$$

which leads to the following KKT conditions:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \mathbf{X}^T \boldsymbol{\lambda} = \mathbf{0}$$

$$\frac{\partial L}{\partial b} = \boldsymbol{\lambda}^T \mathbf{1} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\epsilon}} = c\boldsymbol{\epsilon} - \boldsymbol{\lambda} = \mathbf{0}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \mathbf{X}\mathbf{w} + \mathbf{1}b - \mathbf{y} - \boldsymbol{\epsilon} = \mathbf{0}. \tag{3.4}$$

These lead to the matrix equation:

$$\underbrace{\begin{bmatrix} \mathbf{I} & 0 & 0 & \mathbf{X}^T \\ 0 & 0 & 0 & \mathbf{1}^T \\ 0 & 0 & -\mathbf{I} & c\mathbf{I} \\ \mathbf{X} & \mathbf{1} & -\mathbf{I} & 0 \end{bmatrix}}_{M} \begin{bmatrix} \mathbf{w} \\ b \\ \boldsymbol{\epsilon} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \\ \mathbf{y} \end{bmatrix}, \tag{3.5}$$

72

which yields the solutions for $\mathbf{w}$ and $b$ as:

$$
\begin{bmatrix} \mathbf{w} \\ b \\ \epsilon \\ \lambda \end{bmatrix} = \boldsymbol{M}^{-1} \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \\ \mathbf{y} \end{bmatrix} \tag{3.6}
$$

Therefore, if $\boldsymbol{C} = \boldsymbol{M}^{-1}$ then $\mathbf{w}$ and $b$ can be recovered by taking into account the respective submatrices of $\boldsymbol{C}$:

$$
\mathbf{w} = \boldsymbol{C}[1:d, d+1+n+1:d+1+2n]\mathbf{y}
$$

$$
b = \boldsymbol{C}[d+1, d+1+n+1:d+1+2n]\mathbf{y}, \tag{3.7}
$$

which are linear solutions with respect to the clinical variables $\mathbf{y}$. Recall that $d$ is the dimensionality, and $n$ is the sample size of the data matrix $\mathbf{X}$.

## 3.2.4  Interpretability of weights through activations

[68] have cautioned against directly using discriminative model weights for interpretation in neuroimaging. This is because underlying noise patterns may skew the discriminative directions away from the true effect. Importantly, [68] showed that it is possible to proportionally recover the interpretable underlying effect, also known as the activation, $\mathbf{a}$, by rotating the estimated linear discriminative model $\mathbf{w}$ (*i.e.*, $\mathbf{Xw} = \mathbf{y}$) by left multiplying it

with the covariance matrix of the data:

$$\mathbf{a} \propto \text{Cov}(\mathbf{X})\mathbf{w} = \frac{1}{n}(\mathbf{X} - \overline{\mathbf{X}})^T(\mathbf{X} - \overline{\mathbf{X}})\mathbf{w}. \tag{3.8}$$

The covariance matrix can be estimated either empirically, or by using shrinkage esti-

mators [97].

In the case of LS-SVM, the multivariate discriminative pattern is estimated as $\mathbf{w} = C\mathbf{y}$.

Therefore, one can obtain the activation $\mathbf{a}$ through the following rotation:

$$\mathbf{a} \propto \underbrace{\frac{1}{n}(\mathbf{X} - \overline{\mathbf{X}})^T(\mathbf{X} - \overline{\mathbf{X}})C}_{M}\mathbf{y}. \tag{3.9}$$

One of the important advantages of activations $\mathbf{a}$ over discriminative weights $\mathbf{w}$ is that

activations allow the capture of multiple informative correlated features whereas the dis-

criminative weights $\mathbf{w}$ may only act on a subset of these features. Note that utilizing ac-

tivations over weights does not completely circumvent the issues of multicollinearity in

features [110]. However, covariance matrix multiplication does redistribute the signal cap-

tured in weights to correlated features. Furthermore, the sign of activations $\mathbf{a}$ is in parity

with their correlation with the responses $\mathbf{y}$. This allows the summation of correspond-

ing activations across multiple learners without cancellation, an issue that is present with

summing discriminative weights.

Hereafter, the activation $\mathbf{a}$ and its corresponding parametric matrix $M$ ($M = \text{Cov}(\mathbf{X})C$)

along with the weight vector $\mathbf{w}$ and its corresponding parametric matrix $C$ will be used to

construct the MIDAS statistic.

### 3.2.5 MIDAS Statistic

For all voxels in one volume, we estimate multiple multivariate discriminative patterns $\mathbf{w}^p$ and their corresponding activations $\mathbf{a}^p$ by applying the LS-SVM to different neighborhoods (indexed by $p$) that contain it. Thus, for the $i$th voxel, we obtain a set of values $\{w_i^p\}$ and $\{a_i^p\}$ corresponding to the coefficient values of the weight vectors and the activations at the respective location, as well as a set of squared decision margins $\{\frac{1}{\|\mathbf{w}^p\|_2^2}\}$. Our goal is to summarize these values by a single measure that represents the effect of interest (*e.g.*, group difference) at that spatial location, which will be used for statistical analysis.

We expect voxels that reflect effects of interest to have high absolute values of activations with the sign of the activation in correspondence with the direction of effect. The contribution of the $i$th voxel to the local activation at the $p$th neighborhood is given by $a_i^p$. Taking into account that a voxel belongs to multiple neighborhoods, its total activation contribution is given by the sum of the respective activations across these neighborhoods:

$$v_i = \sum_{p=1}^{P} a_i^p. \tag{3.10}$$

The above quantity should be high when a voxel is well localized in an area of significant effects of interest (*e.g.*, group difference), as it would contribute significantly to the activation patterns of multiple neighborhoods that contain it.

From a multivariate discrimination sense, in uninformative neighborhoods, we expect voxels to take low weight coefficient values. However, it is possible that some voxels take high absolute weight coefficient values due to overfitting. In such cases though, the decision margin of the neighborhood will be small, suggesting poor predictive power. As a

consequence, the predictive power of the learner provides us with a measure of reliability. If we denote the half squared margin for the LS-SVM applied to the $p$th neighborhood by $\frac{1}{\|\mathbf{w}^p\|_2^2}$, then the sum of the inverse predictive power of all learners, in which voxel $i$ participates, is given by:

$$m_i = \sum_{p=1}^{P} \|\mathbf{w}^p\|_2^2. \tag{3.11}$$

In designing the MIDAS statistic, we opt to emphasize contributions of voxels that are part of highly reliable machine learners, while limiting the importance of the ones that participate in regional learners of poor predictive power. Thus, we compute the per voxel statistic by modulating the total contribution of each voxel to the estimated local activation patterns with the total predictive power of the respective machine learners:

$$s_i = \frac{v_i}{m_i} = \frac{\sum_{p=1}^{P} a_i^p}{\sum_{p=1}^{P} \|\mathbf{w}^p\|_2^2}. \tag{3.12}$$

The above normalization enables higher scrutiny for voxels in non-discriminative neighborhoods, while further increasing the statistic of voxels in highly discriminative neighborhoods.

### 3.2.6  Moments calculation

Here, it is assumed that the data $\mathbf{X}$ and the clinical variables $\mathbf{y}$ remained fixed for a particular analysis. The randomness occurs from applying permutation operations on the clinical variables $\mathbf{y}$. Therefore, the expectation, variance and covariance operators, $\mathrm{E}(\cdot)$, $\mathrm{Var}(\cdot)$, $\mathrm{Cov}(\cdot)$ are with respect to the uniform distribution of permutations on $\mathbf{y}$.

Without loss of generality, it is assumed that the clinical variables are z-scored, such that under random permutation, $E(y_j) = 0$ and $\mathrm{Var}(y_j) = 1$. Otherwise, these can be z-scored prior to analysis.

The first moment is approximated, up to the first order term, using the delta method [20]:

$$
\begin{aligned}
E(s_i) = E\left(\frac{v_i}{m_i}\right) &\approx \frac{E(v_i)}{E(m_i)} \\
&= \frac{\sum_{p=1}^{P} E(a_i^p)}{\sum_{p=1}^{P} E(\|\mathbf{w}^p\|_2^2)} \\
&= \frac{\sum_{p=1}^{P} \sum_{j=1}^{n} M_{i,j}^p E(y_j)}{\sum_{p=1}^{P} E(\|\mathbf{w}^p\|_2^2)} \\
&= 0.
\end{aligned}
\tag{3.13}
$$

The second moment is also approximated, up to the first order term, using the delta method:

$$
\begin{aligned}
\mathrm{Var}(s_i) &\approx \frac{\mathrm{Var}(v_i)}{E(m_i)^2} \\
&= \frac{\mathrm{Var}\left(\sum_{p=1}^{P} a_i^p\right)}{E\left(\sum_{p=1}^{P} \|\mathbf{w}^p\|_2^2\right)^2} \\
&= \frac{\sum_{p=1}^{P} \sum_{q=1}^{P} \mathrm{Cov}(a_i^p, a_i^q)}{E\left(\sum_{p=1}^{P} \|\mathbf{w}^p\|_2^2\right)^2}.
\end{aligned}
\tag{3.14}
$$

Note that $\mathrm{Var}(a_i^p) = \sum_{i=1}^{n} M_{i,j}^p{}^2$, and $\mathrm{Cov}(a_i^p, a_i^q) = \sum_{i=1}^{n} M_{i,j}^p M_{i,j}^q$.

Also, since $\mathrm{E}(w_i^p) = \sum_{j=1}^n C_{i,j}^p \mathrm{E}(y_j) = 0$, then $\mathrm{E}(w_i^{p\,2}) = \mathrm{Var}(w_i^p) = \sum_{j=1}^n C_{i,j}^{p\,2}$. Therefore,

$$\mathrm{E}\left(\sum_{p=1}^P \|\mathbf{w}^p\|_2^2\right) = \sum_{p=1}^P \sum_{k=1}^d \mathrm{E}(w_i^{p\,2}) = \sum_{p=1}^P \sum_{i=1}^d \sum_{j=1}^n C_{i,j}^{p\,2}. \tag{3.15}$$

Taken together, the second moment is estimated as:

$$\mathrm{Var}(s_i) \approx \frac{\sum_{p=1}^P \sum_{q=1}^P \sum_{j=1}^n M_{i,j}^p M_{i,j}^q}{\left(\sum_{p=1}^P \sum_{i=1}^d \sum_{j=1}^n C_{i,j}^{p\,2}\right)^2}. \tag{3.16}$$

### 3.2.7   Statistical significance

Permutation tests, or exact tests, are a well known framework for hypothesis testing when the underlying distribution of the statistic of interest is either hard to compute, or unknown [117]. Permutation testing has been previously explored to assess the statistical significance of SVM weight vectors [54, 55]. Specifically, voxel-wise p-values can be obtained by comparing the estimated solution to a null distribution constructed by solving the LS-SVM problem using instances of target clinical variables $\mathbf{y}$ shuffled by random permutations. Such permutation procedures are computationally intensive. However, a statistic of the form $\frac{w_i}{\|\mathbf{w}\|^2}$ can be analytically approximated by a normal distribution, resulting in efficient inference strategies [55]. Using analogous analysis, one can show (see 3.2.6) that the MIDAS statistic (Eq. 3.12) is a sub-gaussian random variable whose tails can be approximated by a Gaussian distribution:

$$\frac{s_i}{\sqrt{Var(s_i)}} = \left(\frac{\sum_{p=1}^P \sum_{q=1}^P \sum_{j=1}^n M_{i,j}^p M_{i,j}^q}{\left(\sum_{p=1}^P \sum_{i=1}^d \sum_{j=1}^n C_{i,j}^{p\,2}\right)^2}\right)^{-1/2} \frac{\sum_{p=1}^P a_i^p}{\sum_{p=1}^P \|\mathbf{w}^p\|_2^2} \xrightarrow{D} \mathcal{N}(0,1). \tag{3.17}$$

### 3.2.8 Multiple clinical variables

The optimal weight vector in LS-SVM is a product of the aforementioned $C$ matrix, which solely depends on the data samples $\mathbf{X}$ and the non-imaging variables $\mathbf{y}$ (*e.g.*, clinical diagnosis). Therefore, multiple discriminative model weights, $W \in \mathbf{R}^{d \times r}$, can be obtained if multiple non-imaging variables, $Y \in \mathbf{R}^{n \times r}$ (*e.g.*, diagnosis, age, sex, etc.), are used for training:

$$W = CY. \tag{3.18}$$

As explored in [68], these models can be adjusted for the underlying noise patterns, as well as the interdependent effects between the non-imaging variables, by left and right multiplying the weight vectors $\mathbf{W}$ with the data covariance matrix and the inverse label covariance matrix, respectively. This results in activation patterns $A$, where the effect captured by each weight vector is independent of the underlying noise and possible imbalances in the non-imaging variable distributions:

$$A = \mathrm{Cov}(\mathbf{X})CY\mathrm{Cov}(Y)^{-1}$$
$$= \underbrace{\frac{1}{n}(\mathbf{X} - \overline{\mathbf{X}})^T(\mathbf{X} - \overline{\mathbf{X}})}_{M} CY \left( \frac{1}{n}(Y - \overline{Y})^T(Y - \overline{Y}) \right)^{-1}. \tag{3.19}$$

The expectation of the multiple activations is still zero, which results in the corresponding MIDAS statistic for the $q$th non-imaging variable, $s_i^q$ to also have an expectation of zero:

$$E(s_i^q) = 0. \tag{3.20}$$

79

Using the steps taken to estimate variance yields that for the $q$th weight vector,

$$\text{Var}(s_i^q) \approx H_q^T \text{Cov}(Y) H_q \frac{\sum_{p=1}^{P} \sum_{q=1}^{P} \sum_{j=1}^{n} M_{i,j}^p M_{i,j}^q}{\left(\sum_{p=1}^{P} \sum_{i=1}^{d} \sum_{j=1}^{n} C_{i,j}^{p \; 2}\right)^2}, \tag{3.21}$$

where $H = \left(\frac{1}{n}(Y - \overline{Y})^T (Y - \overline{Y})\right)^{-1}$, and $H_q$ is the $q$th column of $H$.

### 3.2.9 Parameters Selection and Implementation

There are two main parameters in MIDAS. The first parameter is the neighborhood radius, $r$, which controls the size of the local discriminative analysis window. The second parameter is the weight $c$ in the LS-SVM objective (Eq. 3.1). This parameter controls for the amount of slackness in the constraints of the LS-SVM objective, allowing for cases when the data points $\mathbf{X}$ are not linearly separable with respect to the labels $\mathbf{y}$. In other words, $c$ controls the degree to which $\mathbf{w}$ fits the data. One particular way by which the c and r parameters can be selected is by using the resulting significance maps for feature selection and assessing out of sample predictive performance through nested cross-validation [119].

One can also set the number of neighborhoods $P$, which are sampled such that the full brain volume is covered. The MIDAS statistic (Eq. (3.12)) is self-normalized to have zero mean and unit variance independent of the selection of $P$. In our experiments, $P$ is selected such that each voxel across the brain is covered at least 20 times for a given neighborhood radius. A practical suggestion for setting $P$ is to assess the reproducibility of resulting statistical maps over a range of candidate of a number of neighborhoods and choose the minimum value that attains stability.

Note that the topology of the regional neighborhoods need not be spherical nor com-

pact for the resulting statistic to be valid. Thus, neighborhoods that are discontinuous or anisotropic may be deployed in implementation. However, for simplicity, spherical neighborhoods were used in the implementation described within.

Lastly, to ensure that the coverage of the brain is relatively uniform, the MIDAS implementation accounts for the number of times each voxel has been covered to cover undersampled regions at each iteration adaptively.

## 3.3   Experimental Validation

### 3.3.1   Evaluated Methods

Towards evaluating the proposed method, we qualitatively and quantitatively compare MIDAS against commonly used brain mapping methods using both simulated and real neuroimaging data.

**Voxel-based morphometry (VBM) [60, 3, 4, 33]**   This has been one of the most widely used and established methods for voxel-based analysis in neuroimaging studies. The method entails segmentation of gray matter (GM) tissue and spatial normalization to a common template. Local intensities of GM maps are modulated by scaling with the amount of contraction. Differences are then detected by comparing modulated GM maps after Gaussian smoothing. Comparisons are performed by applying Student's t-test in a mass-univariate fashion.

**Permutation-based voxel-based morphometry (P-VBM) [117, 158]**   This is a non-parametric analog of the VBM method, where the voxel-wise significance is assessed by comparing

the test statistic against a null distribution formed by permuting the clinical variables. We performed 2000 permutations in our experiments.

**Optimally discriminative voxel-based analysis (ODVBA) [161, 162]** ODVBA is a technique that aims to determine the optimal spatially adaptive smoothing of images. It uses local non-negative discriminative projection (NDP) to estimate the direction that best discriminates between two groups. The local NDP vectors are then used to derive voxel-wise statistics, whose significance is assessed through permutation tests. The lack of a closed form solution to the NDP problem results in a significant computational burden.

**Searchlight [89, 123]** Searchlight aims to pull signal from all voxels in a spatial region through multivariate analysis. Specifically, a local classifier is applied to the neighborhood surrounding each voxel in k-fold cross-validation (CV) setting. In the following experiments, linear SVM is used as the base learner for searchlight analysis. Each voxel is characterized by the cross-validated classification accuracy. Statistical significance is assessed by permuting the group memberships and recalculating the k-fold CV accuracy for the null distribution.

**SVM-based statistical significance testing (P-SVM) [30, 54, 55]** SVM classification is performed to estimate the optimal hyperplane that separates two classes using all voxels as features and the group memberships as labels. The importance of the hyperplane coefficients is assessed through an analytic approximation of permutation testing. This method is very similar to MIDAS in its use of SVM weight vectors to assign significance to voxel-wise differences. However, the key difference is that MIDAS attempts to find re-

Figure 3.2: The frontal lobe regions that were subjected to simulated atrophy in the validation experiments is denoted by the red mask.

gionally optimal filters, while P-SVM takes into account the whole volume. Furthermore, P-SVM utilizes a hard margin variant of LS-SVM, which may lead to overfitting and false positive regions in high-dimensional settings.

### 3.3.2 Experiments Using Simulated Data

We first validated the proposed method using synthetic data. Specifically, we used a structural magnetic resonance imaging (sMRI) data set consisting of 1.5 Tesla T1-weighted MRI volumetric scans of 200 healthy control subjects. MRI scans were first pre-processed using previously validated and published techniques [60]. The preprocessing pipeline includes: (1) skull-stripping [36]; (2) N3 bias correction [137]; (3) tissue segmentation into gray matter (GM), white matter, cerebrospinal fluid, and ventricles [100]; (4) deformable mapping [121] to a standardized template space [81]; (5) calculation of regional volumetric maps called RAVENS maps [33]; (6) normalization of the resulting maps by the individual intracranial volume; and (7) resampling to 2mm$^3$. After pre-processing, the samples were split into equally sized groups, and group differences were induced by simulating atrophy

within a predefined regional mask (Fig. 3.2). Atrophy was introduced as a multiplicative reduction to the existing tissue volume to preserve the underlying covariance structure of the brain anatomy.

This synthetic setting allows for quantitative evaluation of the sensitivity and specificity of the proposed method in detecting the introduced atrophy. Moreover, it allows for quantitative comparisons against the methods above. In evaluating all methods, we simulated several scenarios that are commonly encountered in neuroimaging studies. First, we examined the robustness of the methods in detecting a fixed level of simulated atrophy under varying parameter settings. Next, we assessed the sensitivity of the methods by analyzing their ability to detect decreasing levels of simulated atrophy at a fixed parameter setting. Similarly, we tested the effectiveness of the methods in detecting differently shaped and sized atrophy patterns. Next, we evaluated how the sample size affects the performance of the methods. Lastly, we assessed the false positive rate of these methods.

**Analytical vs. Experimental Estimation of p-values**

MIDAS makes use of an efficient, analytic approximation to estimate p-values (Eq. 3.17). To assess the validity of the approximation, we compared the analytically approximated p-values of the MIDAS statistic to the ones that were empirically estimated through non-parametric testing based on 2000 permutations (see Fig. 3.3 Left). One can visually appraise the high alignment between the two estimations. Few inconsistencies were observed, which may be due to an insufficient number of permutations. This is further supported by the decreasing mean squared error of the analytically approximated p-values compared to the empirically estimated ones with increasing number of permutations (Fig. 3.3
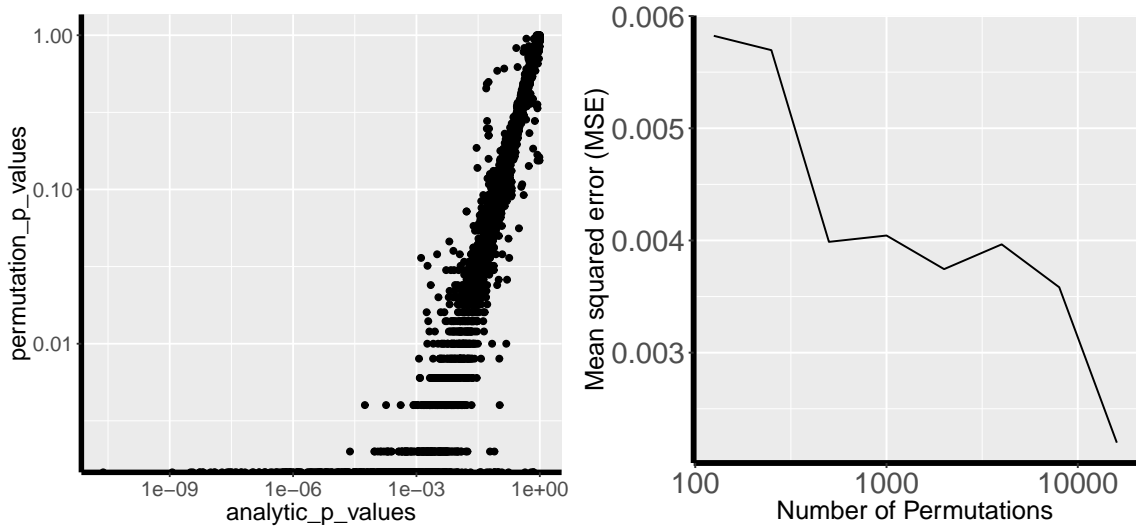
Figure 3.3: Left: Log-log scale plot of the distribution of analytic- vs. permutation-based estimation of MIDAS p-values. Right: Mean squared error of p-value estimation as a function of increasing number of permutations.

Right).

**Robustness to parameter variation**

In this experiment, we introduced 35% atrophy in the data, and evaluated how the performance of each method changes when varying its key parameters. For VBM, P-VBM, and P-SVM the full-width half maximum (FWHM) of the Gaussian smoothing kernel for the input images was varied from 4mm to 10mm, at 2mm intervals. For Searchlight, the searchlight radius was ranged from 2 voxels (4mm) to 5 voxels (10mm). For ODVBA and MIDAS, the neighborhood radius r was varied from 8 voxels (16mm) to 20 voxels (40mm). For MIDAS, the c parameter was varied from $10^{-1}$ to 100. The performance was assessed by thresholding detections at false discovery rate (FDR) [11] level $q < 0.05$, and then calculating the True Positive Rate (TPR) and the False Positive Rate (FPR).

Quantitative results for all methods are shown in Fig. 3.5. The TPR and FPR, as well as
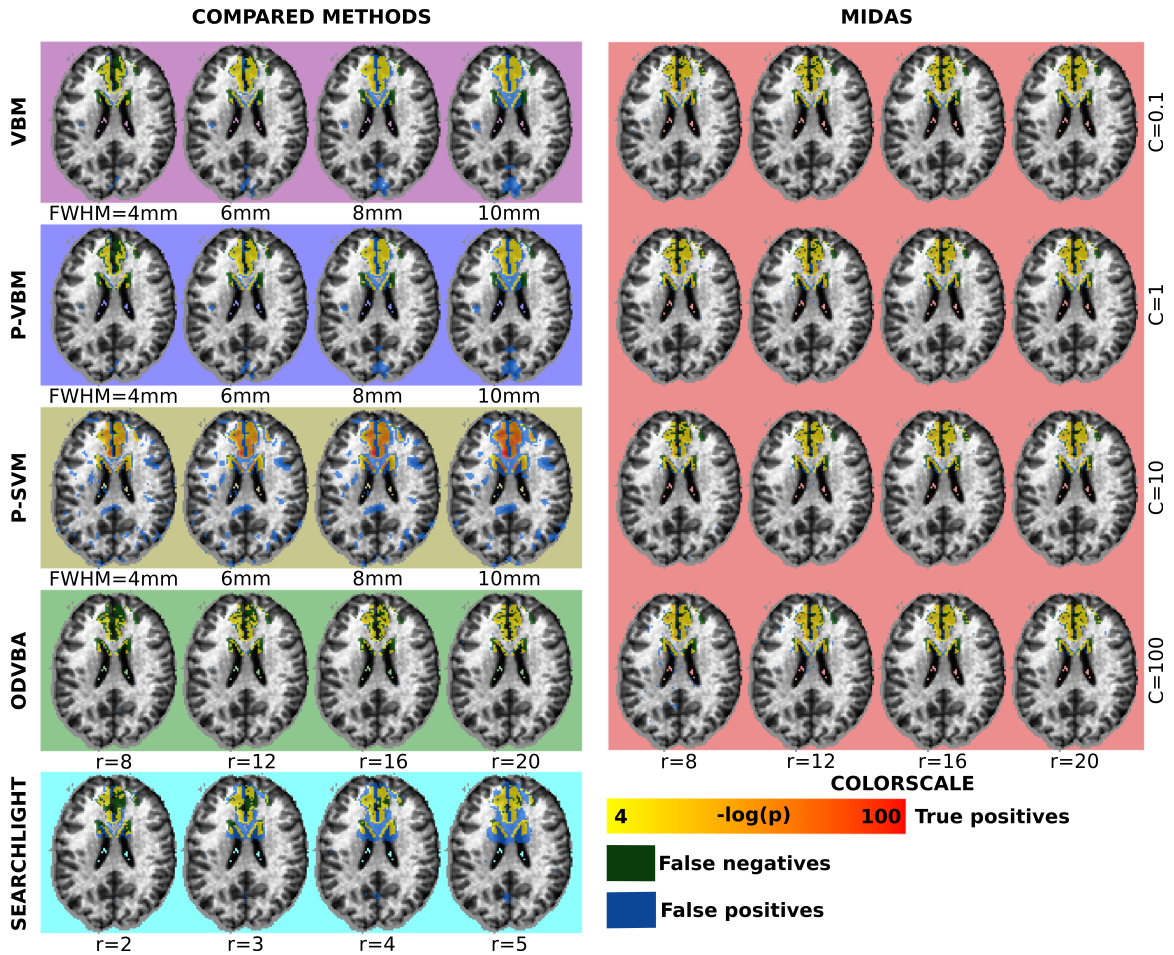
Figure 3.4: Detection results obtained by all methods using the dataset with 35% simulated atrophy. Regions were estimated by thresholding significance maps at FDR level of $q < 0.05$. The resulting masks were compared to the ground-truth: true positives for all methods are color-coded by yellow to red gradient, false positives are denoted by blue voxels, while false negatives are denoted by green voxels. Results by varying the Gaussian smoothing kernel size (in the case of VBM, P-VBM, P-SVM), the neighborhood radius size (in the case of ODVBA and Searchlight), as well as the neighborhood radius and c parameter (in the case of MIDAS), are shown.

the entire receiver operating curve (ROC) for all methods, are reported. MIDAS produced the fewest false positives for almost all parameter configurations. At the same time, MIDAS was able to obtain high TPR. Methods that depend on Gaussian smoothing, such as VBM, P-VBM, and P-SVM, were able to obtain high TPR at the cost of high FPR. Similarly, Searchlight was not able to attain high TPR without conceding high FPR. Converging con-
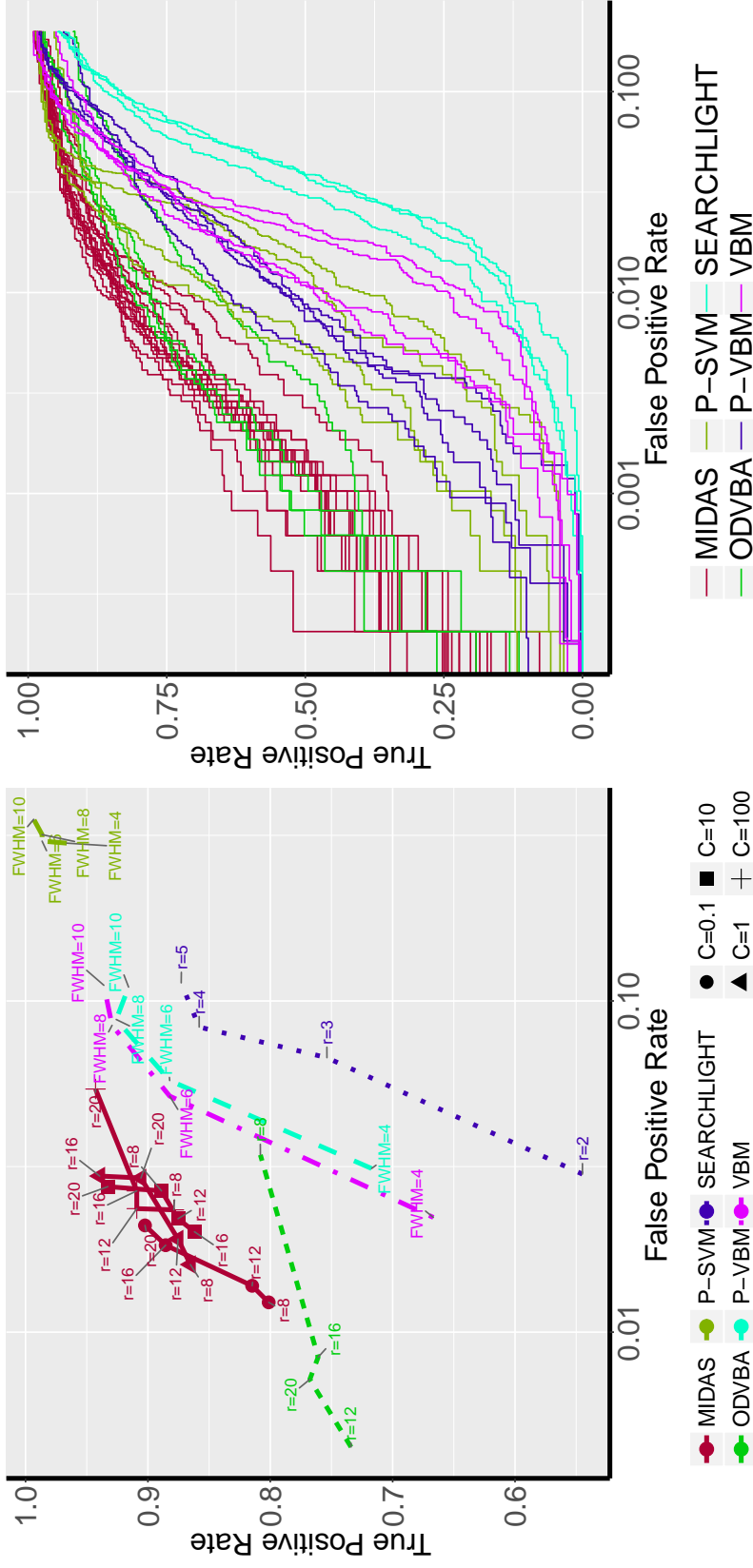
Figure 3.5: Left: Estimated FPR (x-axis) and TPR (y-axis) for all methods and parameters using data with 35% simulated atrophy. Detected regions were estimated by thresholding significance maps at FDR level $q < 0.05$. Right: Receiver operating curves for all methods at all considered parameter combinations. Results for different methods are color-coded as follows: MIDAS, red; VBM, magenta; P-VBM, blue; ODVBA, green; Searchlight, cyan; and P-SVM, yellow. Results were obtained by varying method parameters. For MIDAS, different values for neighborhood radius $r$ and loss penalty $c$ (indicated by different markers) were examined. For VBM, P-VBM, and P-SVM different levels of Gaussian smoothing were investigated. For ODVBA and searchlight, increasing neighborhood radii were tested. The proposed method maintained on average higher sensitivity and specificity than all other methods. Moreover, MIDAS exhibited relatively stable performance independent of the employed parameters. ODVBA showed similar specificity to MIDAS, but at the cost of relatively lower sensitivity and significantly increased computational time.

clusions can be drawn by visually inspecting the ROC curve. MIDAS achieved the highest area under the curve, followed by ODVBA. ODVBA and MIDAS are similar in spirit as they both perform local discriminative learning to tease out local signal patterns. However, MIDAS, on top of attaining a slightly higher TPR, is computationally more efficient than ODVBA. MIDAS makes use of efficient analytical approximations resulting in a computational time that is three magnitudes faster than that of ODVBA, which is based on computationally expensive permutation tests.

The regions that were detected as significant for all methods and parameter configurations are shown in Fig. 3.4. In agreement with the quantitative results, we note that VBM, P-VBM, were able to decrease the number of false negatives (shown in white) with increasing smoothing, albeit at the cost of increasing the number of false positives (shown in orange). A similar trend was observed in the case of Searchlight when increasing the neighborhood radius. P-SVM detects the effect of interest for all parameters, but produces false positives. ODVBA, on the contrary, did not produce false positives, but the number of false negatives depended on the size of the local neighborhood. MIDAS produced few false positives, while also achieving few false negatives. Importantly, the results were stable across all parameter settings.

**Sensitivity to the size of the simulated effect**

To further evaluate the capability of the compared methods to detect the simulated atrophy, we created additional datasets by varying the simulated atrophy in the frontal lobe mask (Fig. 3.2) from 15% to 35%. Detected regions were first determined by thresholding significance maps at FDR level $q < 0.05$, and then compared to the ground-truth. As
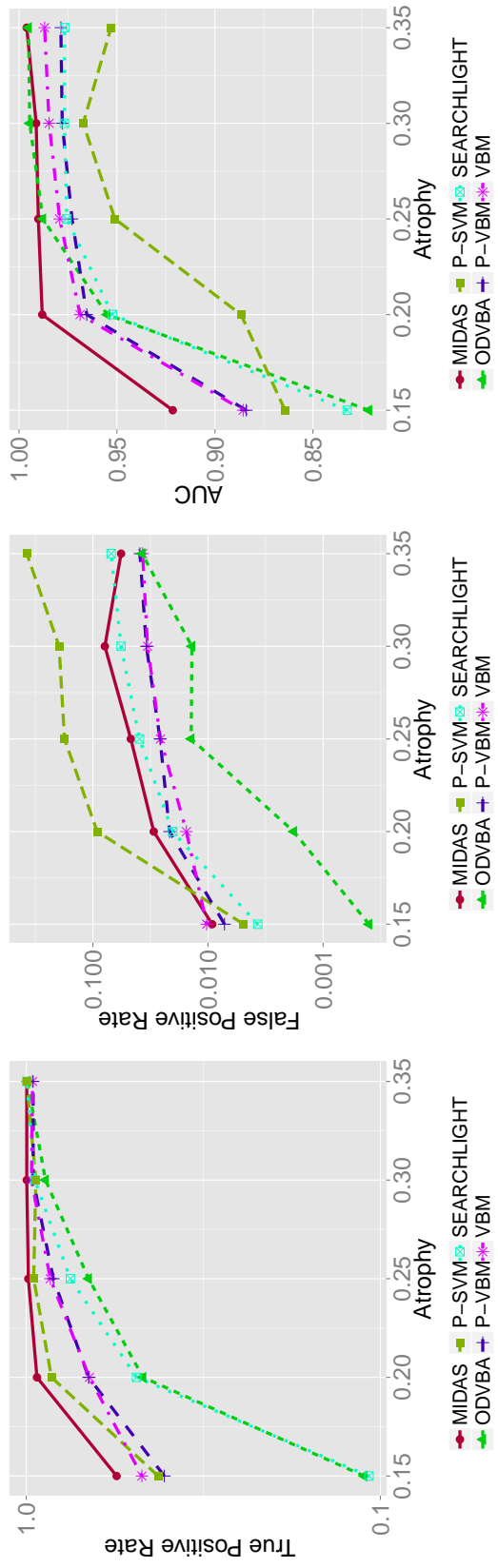
Figure 3.6: Performance as a function of the degree of simulated atrophy. Performance is quantified by estimating TPR (left) and FPR (center) at FDR level $q < 0.05$, as well as measuring the area under the receiver operating characteristic curve (AUC; right). Results for different methods are color-coded as follows: MIDAS, red; VBM, magenta; P-VBM, blue; ODVBA, green; Searchlight, cyan; and P-SVM, yellow. For each method, the parameters that yielded the highest TPR for the 35% simulated atrophy experiment were used. As a consequence, most methods achieved high TPR, with increased TPR being observed for higher degrees of atrophy. The methods differed with respect to the FPR they achieved. MIDAS attained lower FPR than all other methods, except for ODVBA, which also exhibited the lowest TPR.

Figure 3.7: Regions detected by all methods for different degrees of introduced atrophy. Regions were estimated by thresholding significance maps at FDR level of $q < 0.05$. True positives for all methods are color-coded by yellow to red gradient, false positives are denoted by blue voxels, while false negatives are denoted by green voxels.

previously, we evaluated the performance of the methods by calculating TPR and FPR, as well as measuring the area under the receiver operating characteristic curve (AUC). For each method, the parameters that yielded the highest TPR for the 35% simulated atrophy experiment were used. Specifically, for VBM, P-VBM, and P-SVM, the FWHM of the Gaussian smoothing kernel for the input images was set to 8mm. The Searchlight radius was also fixed to an 8mm radius. The neighborhood radius of ODVBA and MIDAS was set to 16mm, while the c parameter of MIDAS was set to the default value of 1.

As expected given the choice of parameters, all methods achieved high TPR, while increasing the degree of simulated atrophy resulted in increased TPR (see Fig. 3.6). MIDAS was able to reveal the true signal for varying levels of atrophy, and at a TPR comparable to VBM and P-VBM. Importantly, MIDAS was able to attain lower FPR than both VBM and P-VBM for all atrophy levels. Only ODVBA was able to attain slightly lower FPR than MIDAS for some atrophy levels, but that was achieved at the cost of much lower TPR. The above differences were also reflected in the AUC measurements. Increased atrophy resulted in increased AUC values for all methods, with VBM, P-VBM, and Searchlight converging in lower values than MIDAS and ODVBA. MIDAS and ODVBA achieved similar best performance for high levels of atrophy, while MIDAS retained high-quality performance for low levels of atrophy too.

The regions that were detected as significant for all methods and degrees of atrophy are shown in Fig. 3.7. In agreement with the quantitative results, we note that VBM, P-VBM, P-SVM, and Searchlight were able to identify increased portions of the underlying signal for increased degrees of simulated atrophy. However, they also resulted in an increasing number of false positives. ODVBA and MIDAS, on the contrary, were able to able to

recover increasing portions of the simulated signal, while introducing less false positives.

**Sensitivity to the shape of the simulated effect**

The goal of this experiment is to investigate how the shape and extent of the underlying pathology influence, in conjunction with the used parameters, the performance of the different methods. Towards this end, the simulated atrophy in the frontal lobe was broken into three subregions of different morphology (Fig. 3.8 A, B, C), and 35% atrophy was introduced successively to each subregion while leaving the rest intact. Moreover, each method was run by using multiple parameters. For VBM, P-VBM, and P-SVM, we varied the FWHM of the Gaussian smoothing kernel from 4mm to 10mm, with a step of 2mm. Similarly, the radius of the searchlight was ranged from 4mm to 10mm. The neighborhood radius of ODVBA and MIDAS was ranged from 16mm to 40mm. Lastly, the c parameter of MIDAS was set to 1.

The performance of all combinations of methods and parameters was assessed by measuring the AUC, which was calculated by comparing the ground-truth mask and the respective voxel-wise statistics (Fig. 3.8). We note that different levels of smoothing (as utilized by VBM, P-VBM, and P-SVM) were optimal for detecting different effects. For example, in the case of elongated and more focal simulated effects (Fig. 3.8A and C), less smoothing was optimal for VBM compared to the case of the larger simulated effect in Fig. 3.8B. This is because, to detect the underlying signal better, a matched filter is required. As a consequence, focal patterns require less smoothing than larger ones to yield specific brain maps. Similarly, in the case of Searchlight, different neighborhood sizes, containing sufficient informative voxels, were optimal for detecting different effects. Contrary to the

other methods, MIDAS was able to detect effects of different shape and extent with high accuracy regardless of the choice of parameters.

**Exploring the effect of the sample size**

This experiment aims to study the statistical power of each method as a function of the sample size. Towards this end, we generated multiple datasets by introducing 35% simulated atrophy in the frontal lobe mask (Fig. 3.2), and varying the sample size from 40 to 400. For every sample size, we applied each method ten times, and estimated the average AUC (Fig. 3.9).

Given enough samples, all methods were able to detect the strong simulated signal. With increasing available data, most methods converged to a high AUC value. However, important differences were observed for lower samples sizes. In these cases, MIDAS demonstrated advantageous statistical power in detecting the underlying signal compared to the other methods. Additionally, the comparable statistical power of ODVBA relative to MIDAS is offset by its high computational expense, which is at least two orders of magnitude higher than the runtime of MIDAS.

**Evaluating the family-wise error**

In this experiment, we evaluated the probability of making one or more false discoveries for each method. Towards this end, we used random subsets of healthy controls subjects without inducing any simulated atrophy. As a consequence, the null hypothesis of no group difference should be true. Group comparisons were performed ten times using each method, and the FPR at $p < 0.05$ level was computed for each method (Fig. 3.10 Left).

As expected, the FPR at $p < 0.05$ was within 5% for all methods. It should be noted that P-SVM yielded noticeably higher FPR than all other methods. This finding was also observed in experiments discussed in Secs. 3.3.2 and 3.3.2.

To further compare the behavior of the methods in the absence of any signal, p-value scatter plots between the different methods are shown in Fig. 3.10 Right. One observation is that the p-values of Searchlight were uncorrelated to the p-values of all other methods. Another interesting finding is that the p-values of P-SVM followed a sub-linear relationship with respect to VBM p-values. While higher p-values of P-SVM followed a linear trend with VBM p-values, the lower p-values of VBM were further lowered by P-SVM. This may explain why P-SVM generates a higher number of significant voxels, even in the absence of underlying signal.

**Simulated regression case study**

To demonstrate the regression ability of MIDAS, we created another validation dataset by continuously varying the simulated atrophy in bilateral temporal lobe regions (Fig. 3.11) as a function of age in 100 control subjects whose ages ranged from 55 to 90. For the mildest effect simulated, 55-year-olds experienced zero atrophy while 90-year-olds were simulated to have 15% atrophy. In the strongest case simulated, 55-year-olds again experienced zero atrophy, while 90-year-olds experienced 50% atrophy bilaterally in temporal lobe regions. To further render this simulation realistic and to decrease signal to noise ratio, 15 % label noise was added in the sense that the exhibited atrophy was randomly modulated to be within 15 % of the expected atrophy at the subject age.

For evaluation, in addition to MIDAS, only VBM and searchlight were suitable for

94
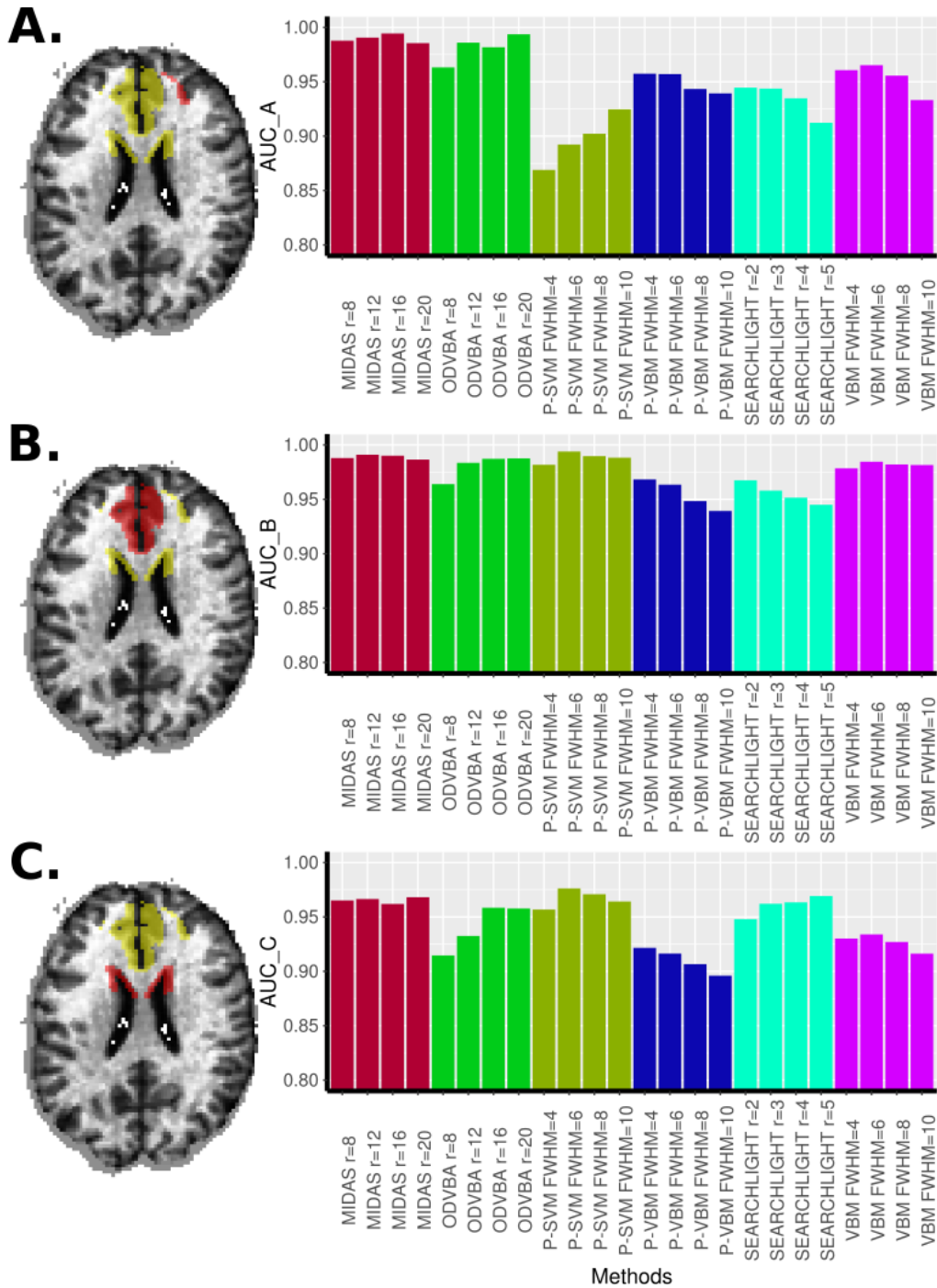
Figure 3.8: The AUC of compared methods for three simulated effect shapes. The red subregion of the atrophy mask (in yellow) was subjected to 35% atrophy. The resulting AUC of the compared methods under varying smoothing parameters (as in the case of VBM, P-VBM, and P-SVM), or radii (as in the case of MIDAS, ODVBA, and Searchlight), is shown on the right.
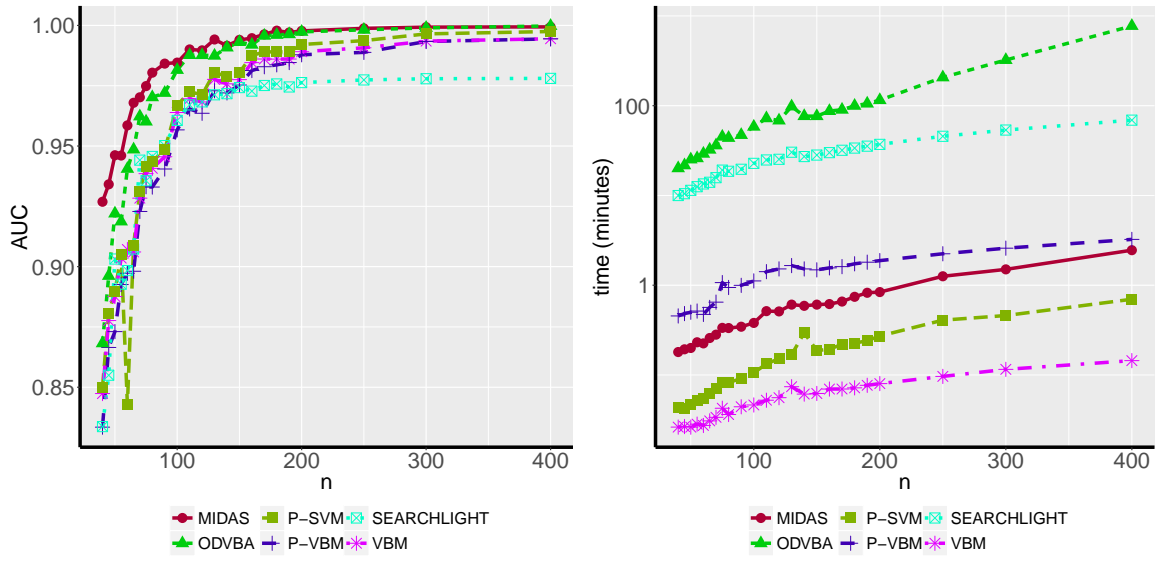
Figure 3.9: AUC versus sample size (n; left) and runtime versus sample size (n; right). While the AUC of all compared methods increased with sample size, MIDAS was more powerful than the compared methods at all sample sizes. Furthermore, while ODVBA approached the statistical power of MIDAS at larger sample sizes, this was at the cost of being several orders of magnitude more computationally expensive.
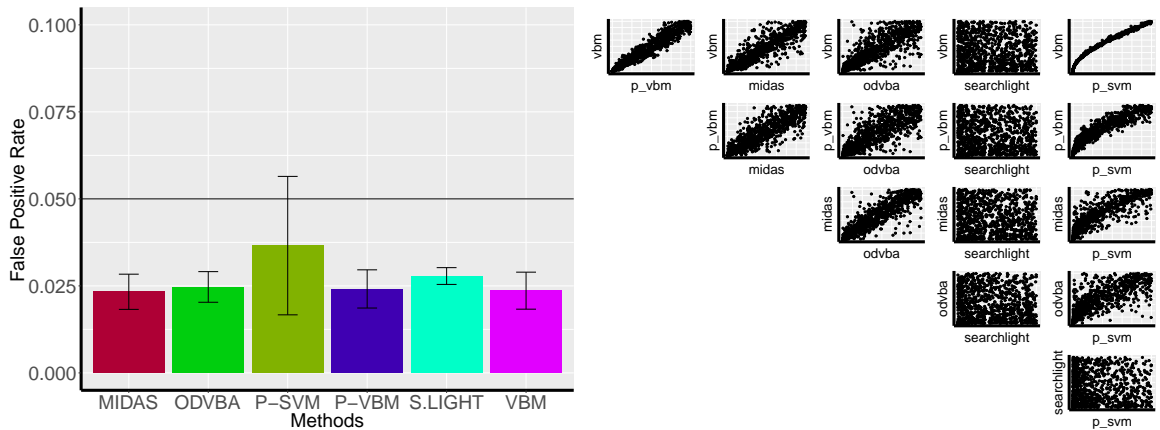


Figure 3.10: Left: The FPR at $p < 0.05$ level for all methods is shown. Right: Pairwise comparison of p-values obtained by all methods. Vertical and horizontal axes range from 0 to 1.

Figure 3.11: The temporal lobe regions that were subjected to simulated atrophy in the regression validation experiments is denoted by the red mask.
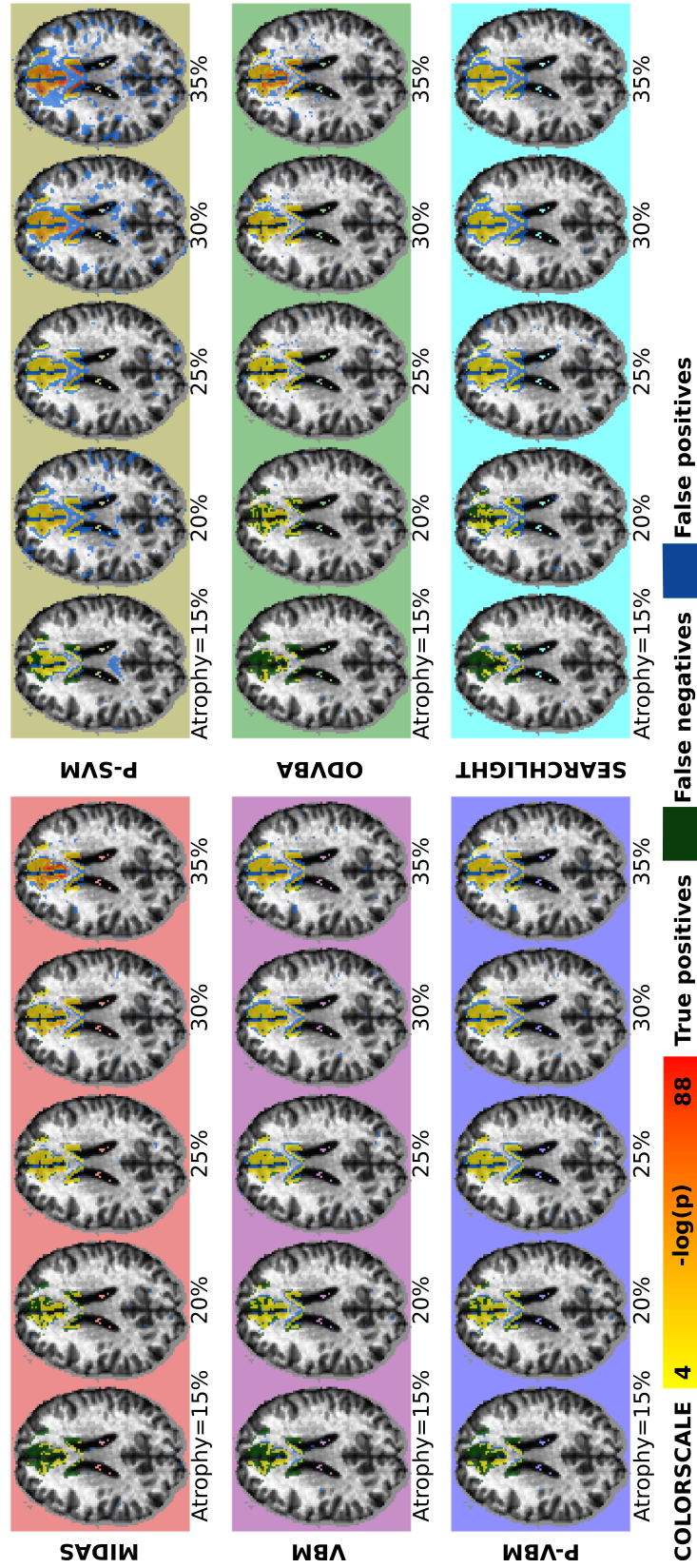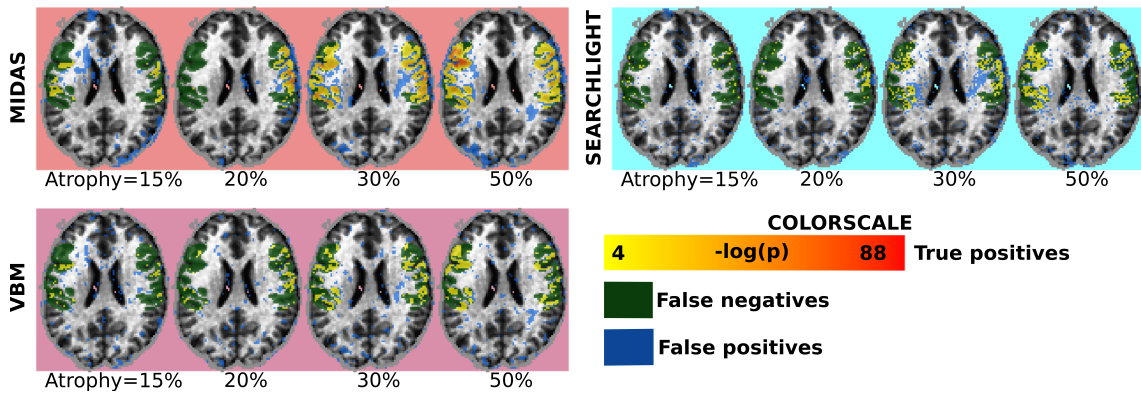


Figure 3.12: Regions detected by all methods for different degrees of introduced atrophy. Regions were estimated by thresholding significance maps at FDR level of $q < 0.05$. True positives for all methods are color-coded by yellow to red gradient, false positives are denoted by blue voxels, while false negatives are denoted by green voxels.
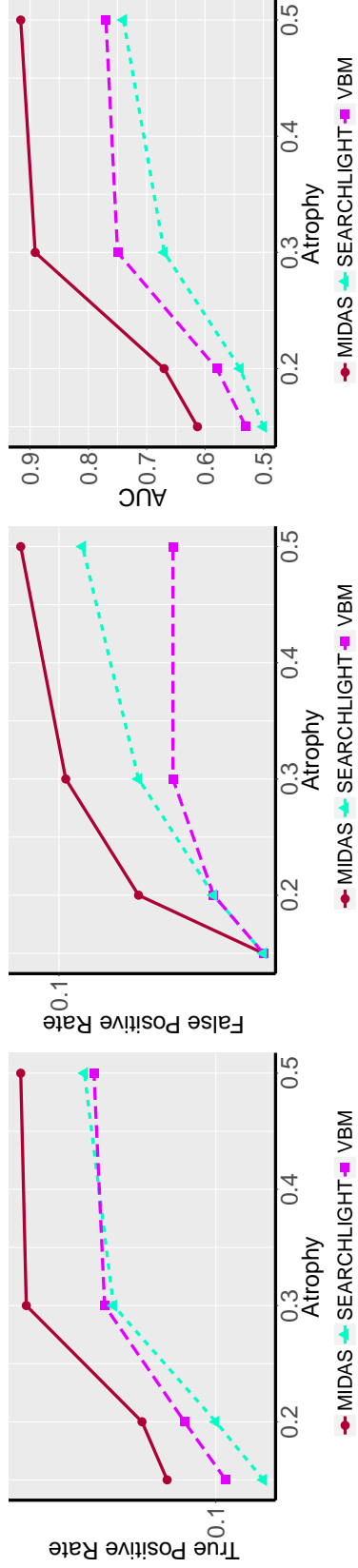
Figure 3.13: Performance as a function of the degree of simulated atrophy in the regression validation case. Performance is quantified by estimating TPR (left) and FPR (center) at FDR level $q < 0.05$, as well as measuring the area under the receiver operating characteristic curve (AUC; right). Results for different methods are color-coded as follows: MIDAS, red; VBM, magenta; and Searchlight, cyan. Increasing atrophy levels resulted in increases in both TPR and FPR for all methods. However, considering these two measures in conjunction with AUC revealed that MIDAS had the greatest TPR to FPR trade-off of all compared methods.

regression cases and were thus tested in this validation study.

Similar to previous evaluations, detected regions were first determined by threshold-ing significance maps at FDR level $q < 0.05$, and then compared to the ground-truth. We evaluated the performance of the methods by calculating TPR and FPR, as well as measur-ing the area under the receiver operating characteristic curve (AUC). For VBM, the FWHM of the Gaussian smoothing kernel for the input images was set to 8mm. The Searchlight radius was also fixed to an 8mm radius. The neighborhood radius of MIDAS was set to 16mm, while the c parameter of MIDAS was set to the default value of 1.

The TPR and FPR of MIDAS, VBM, and searchlight in the simulated regression case study are displayed in Figure 3.13. MIDAS was able to uncover the underlying atrophy pattern at much weaker level of signal than both VBM and searchlight. As observed in previous sections, all methods exhibited higher TPR at increasing signal strength but with increased levels of false positives. However, the AUC plot demonstrates that the increase in true positives is much greater in magnitude than false positives. In addition, VBM and searchlight exhibited similar levels of true positives. However, searchlight false positives were considerably greater which resulted in lower AUC.

To further visualize the results, the regions that were detected as significant for all methods and degrees of atrophy are shown in Fig. 3.12.

### 3.3.3 Functional Neuroimaging Data from a Lie Detection Study

We applied MIDAS along with the comparative methods to a dataset comprising func-tional MRI (fMRI) scans of individuals undertaking lying and truth-telling tasks in a forced choice deception experiment [95]. For the study, 52 right-handed males (mean age=19.36±0.5)

were recruited. Functional data were pre-processed to obtain parameter estimate images (PEIs) as described by [34], who also provided the data.

The parameters used for MIDAS were c=1 and r=16. For ODVBA, r=16 was used. For Searchlight, r=3 was used. For VBM, P-VBM, and P-SVM, FWHM=8mm was the smoothing kernel for the images.

The estimated PEIs were then given as input to the compared methods to locate the brain regions that were most distinctive between the two tasks. Specifically, two PEIs for each subject were obtained and formed two groups that included 52 PEIs corresponding to truth-telling and 52 PEIs corresponding to lying, disregarding the pairing present in the samples. Although this neglect reduces potential statistical power, this was done to compare all methods on equal footing. Statistically significant regions at FDR level $q < 0.05$ for all methods are shown in Fig. 3.14. We note that Searchlight and VBM approaches detected fewer regions. On the contrary, P-SVM, ODVBA and MIDAS found similar regions to be significantly different between the task-based groups, including cerebellum, insular cortex, cingulate, medial frontal gyrus, and postcentral gyrus. Detected regions align well with previously reported results [95]. P-SVM resulted in statistical maps exhibiting the largest spatial extent. However, this may be due to including false positive regions as was observed in the simulation experiments. MIDAS, on the contrary, demonstrated the highest significance in group differences within the identified voxels. Specifically, MIDAS was able to detect highly specific activation in the supramarginal gyrus, which is associated with truth-telling.

Due to the lack of ground-truth, we further quantitatively evaluated the compared methods in terms of split sample reproducibility. The study sample was randomly di-

vided into halves ten times, and for each split, the compared methods were applied. Reproducibility was calculated in two ways by measuring the Dice coefficient and the adjusted Rand index (ARI) [74] between the significant regions detected at each split after FDR correction ($q < 0.05$). While Dice coefficient is a common measure for assessing the overlap between sets, ARI provides a complementary view of set similarity that is adjusted for chance. This property of ARI enables a more fair comparison of the set of voxels that pass the significance threshold across sample splits when the regions of significance vary in spatial extent. Although there is no consensus on what is considered to be a good value of Dice coefficient and ARI, a Dice of over 0.50 is considered to be acceptable while ARI of over 0.75 is deemed excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor [48].

The average Dice coefficient and adjusted Rand index (ARI) across pairs of sample splits are reported in Fig. 3.15 for all methods. MIDAS demonstrated the highest average split sample reproducibility at $0.64 \pm 0.07$ (Dice) and $0.46 \pm 0.09$ (ARI). The second highest performing method in terms of Dice coefficient was P-SVM with an average Dice of $0.61 \pm 0.08$. On the other hand, VBM had the second highest ARI at an average of $0.31 \pm 0.08$. Searchlight had the lowest average split sample reproducibility at with an average Dice coefficient of $0.18 \pm 0.20$ and average ARI of $0.17 \pm 0.04$.

### 3.3.4 Structural Neuroimaging Data from a Cognitive performance study

To observe the regression performance in a clinical dataset, we applied MIDAS, VBM, and searchlight to a structural MRI (sMRI) dataset comprising of 100 mild cognitive impairment subjects from the Alzheimer's disease neuroinitiative (ADNI) study. The sMRI images were processed using the same steps as described in section 3.3.2 to yield gray mat-
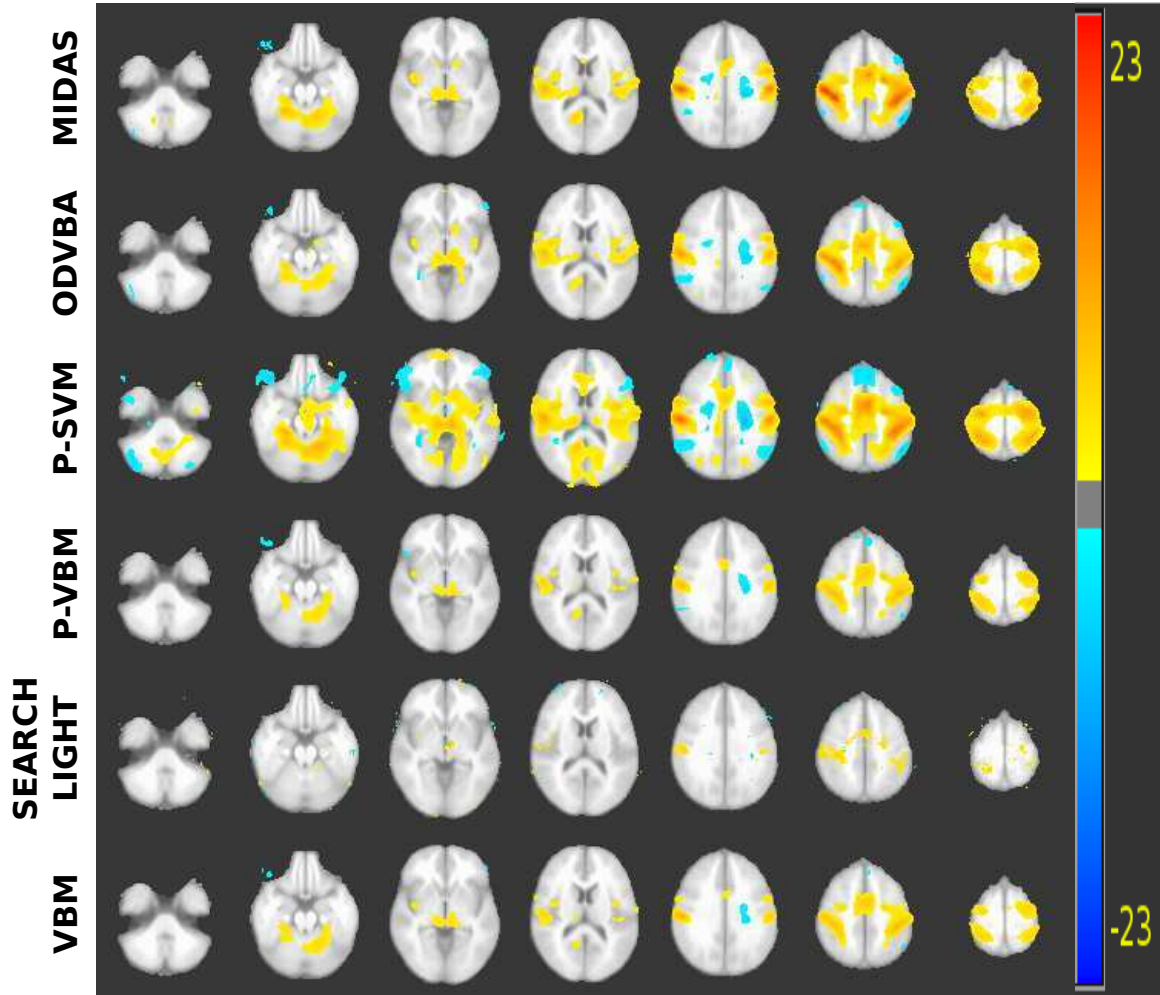
Figure 3.14: Significant regions detected after FDR correction ($q < 0.05$) by all methods using the functional MRI lie detection task dataset. The color intensity indicates -log p value. Warmer colors indicate increased activation during truth telling, while colder colors indicate increased activation during lying. The color scale is matched for all methods to facilitate comparisons.

ter volumetric tissue density maps. The continuous score that the imaging features were regressed against was *Alzheimer's Disease Assessment Scale Cognitive Behavior Section*(Adas-cog-13) which is a measure of cognitive performance that is widely used in Alzheimer's disease trials [108]. A higher Adas-cog-13 score indicates a greater level of cognitive dysfunction.

The parameters used for MIDAS were c=1, r=16. The searchlight radius was r=3 while

102
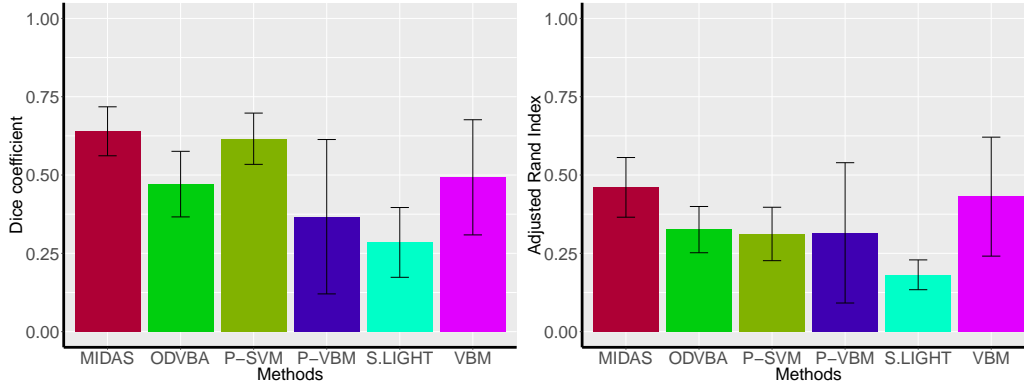
Figure 3.15: Average split sample reproducibility of the compared methods for the lie de-tection dataset measured by Dice coefficient and adjusted Rand index. The error bars de-note standard deviation. MIDAS demonstrated the highest reproducibility at 0.64 Dice coefficient and 0.46 adjusted Rand index on average. P-SVM had the second highest Dice coefficient at 0.61 on average while VBM had the second highest adjusted Rand index at 0.43 on average. Searchlight achieved the lowest reproducibility with an average Dice coefficient and adjusted Rand index of 0.18 and 0.17, respectively.

the VBM smoothing kernel was FWHM=8mm.

The gray matter tissue density maps, known as RAVENS maps [33], were given as in-put to the compared methods to locate the brain regions that were most associated with cognitive performance as quantified by Adas-cog-13 score. Statistically significant regions at FDR level $q < 0.05$ for all methods are shown in Fig. 3.16. For maps that are corrected for multiple comparisons, VBM yielded fewer regions than MIDAS while searchlight failed to yield any significant regions. Significance maps that are not corrected for multiple com-parisons are shown in Fig. 3.17 with voxels passing $p < 0.05$. Similarly, MIDAS yielded regions with more extreme p-values as well as a greater amount of them relative to VBM and searchlight. Importantly, MIDAS was able to accurately associate white matter hyper-intensities and medial temporal lobe atrophy with increased Adas-cog-13 scores which is corroborated by larger sample studies in past literature [146].
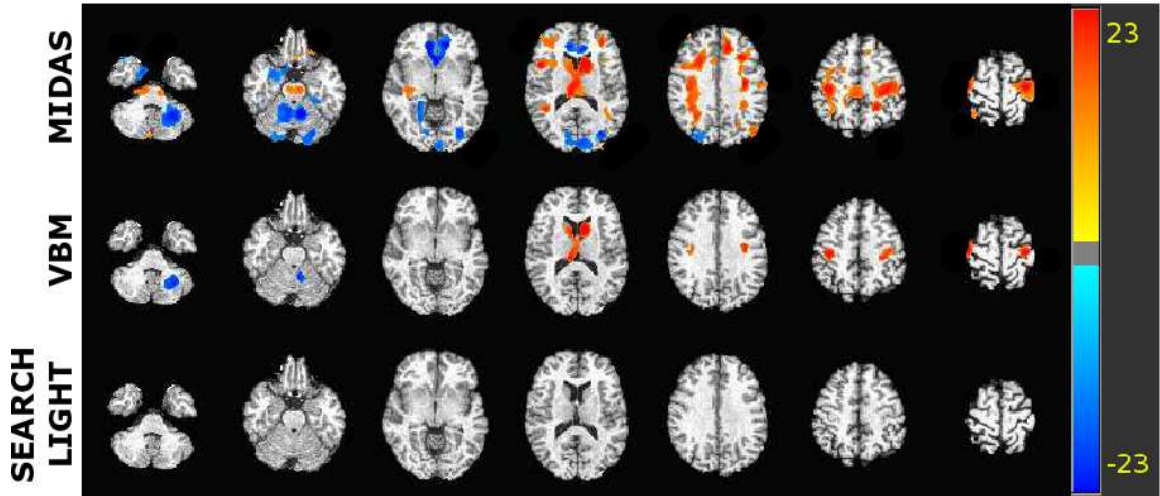
Figure 3.16: Significant regions detected after FDR correction ($q < 0.05$) by all methods using the structural MRI mild cognitive impairment dataset. The color intensity indicates -log p value. Warmer colors indicate increased tissue density correlated with Adas-cog-13 score, while colder colors indicate decreased tissue density correlated with Adas-cog-13 score. The color scale is matched for all methods to facilitate comparisons.
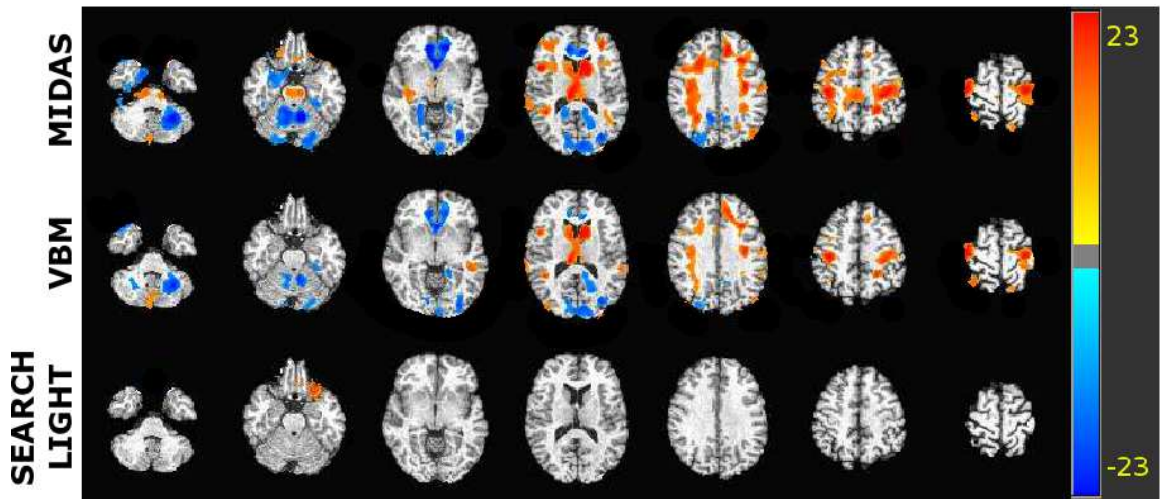


Figure 3.17: Significant regions *uncorrected for multiple comparisons* ($p < 0.05$) by all methods using the structural MRI mild cognitive impairment dataset. The color intensity indicates -log p value. Warmer colors indicate increased tissue density correlated with Adas-cog-13 score, while colder colors indicate decreased tissue density correlated with Adas-cog-13 score. The color scale is matched for all methods to facilitate comparisons.

## 3.4   Discussion & Conclusion

**Synopsis**

In this chapter, we have introduced a novel multivariate pattern analysis method, termed MIDAS, for statistical parametric mapping analysis of images. In the proposed framework, discriminative learning is applied to regional neighborhoods towards estimating the multivariate pattern that best reflects the effect of interest, such as a group difference or regression against a clinical variable. Information from regional discriminants derived from multiple neighborhoods is combined to estimate a statistic for each voxel that is associated with them. Intuitively speaking, this statistic assigns high values to voxels that contribute significantly to highly discriminative learners. Critically, an analytic approximation of the null distribution is employed towards efficiently estimating voxel-wise significance without the need for very costly permutation tests. The proposed framework was extensively validated using simulated data, and tested on real functional MRI data pertaining to lie detection and a structural MRI dataset of mild cognitive impairment. Compared to commonly used brain mapping techniques, the proposed framework demonstrated advantageous performance, underscoring its potential to efficiently map effects of interest in both structural and functional data.

**Comparison with voxel-based analysis methods**

Commonly applied voxel-based analysis techniques smooth the data spatially using kernels defined in an ad hoc or empirical way, thus imposing *a priori* assumptions regarding the shape and spatial extent of the effect of interest, which itself might be heterogeneous

105

throughout the brain. Such assumptions may lead to reduced statistical power and spatial specificity of the resulting maps as the applied smoothing is seldom adapted to the scale and shape of the signal of interest. In sharp contrast to them, the main premise of MIDAS is that it optimally detects effects of interest by effectively applying a form of matched filtering. Since the underlying effect to which the matched filter should adapt is not known in advance, regional discriminative analyses are used to combine information from the most informative voxels resulting in optimal regional filtering. Critically, this filtering does not blur or smear out the derived statistical parametric maps, since those are eventually formed at the voxel resolution by forming a voxel-wise statistic informed by all regional learners that include a particular voxel.

**Comparison with multivariate methods**

MIDAS is somewhat similar in spirit to Searchlight, ODVBA, and P-SVM. Nonetheless, it significantly deviates from them. First, MIDAS creates the information map by taking into account the contribution of each voxel to the classifiers that include it as well as the classifier's discriminative power. This is in contrast to Searchlight that assigns each searchlight's classification accuracy to its center voxel. This difference has two important implications: i) it allows for a more refined characterization of the importance of each voxel, and ii) it increases computational efficiency by relaxing the requirement of running regional classification for every voxel. The only requirement in the case of MIDAS is that the employed neighborhoods should cover sufficiently the whole image volume. By appropriately combining information from all neighborhoods, MIDAS can estimate the per-voxel statistics.

Second, MIDAS and ODVBA share the goal of estimating the optimal spatially adap-

tive filtering of the data. However, their design and implementation are significantly different. ODVBA is designed to tackle group comparison tasks, and cannot naturally handle regression tasks. Moreover, ODVBA is based on non-negative discriminative projection, which hinders an analytical approximation of the statistical significance map. As a consequence, computationally expensive permutation tests are required for the estimation of voxel-wise p-values. On the contrary, MIDAS is generic and can readily handle both group comparison and regression tasks. Additionally, MIDAS introduces an analytical approximation of the null distribution of the proposed statistic, achieving significant speed-up making it attractive for computational neuroanatomy applications using large neuroimaging data.

P-SVM is also based on an analytical estimation of voxel-wise significance maps. This estimation is founded on the assumption of a high dimensional low sample size setting. MIDAS does not make such an assumption when deriving its analytical approximation model. Interestingly, this model can be understood as a bootstrapping generalization of P-SVM [54, 55], endowed with a similar, yet different, null distribution. Theoretically, bootstrapping can be used to stabilize otherwise noisy statistics [43]. Empirically, we showed that MIDAS statistic yields a higher AUC than P-SVM for different degrees of atrophy (Fig. 3.5) and number of samples (Fig. 3.9). Lastly, in MIDAS, we further incorporated the correction procedure proposed by [68] to utilize interpretable activations rather than weight vectors.

In summary, the proposed framework addresses important limitations of alternative methods. MIDAS makes use of optimal spatially adaptive filtering to detect with improved sensitivity structural, or functional signal of interest. Specifically, MIDAS was

found in several experiments to consistently delineate effects of interest while being relatively invariant to the tuning parameters, facilitating its usage and favoring reproducible research. Additionally, it demonstrated increased sensitivity in detecting the signal of interest at various degrees of strength, without introducing false positives. Notably, increased sensitivity was observed for both small and large sample sizes. Moreover, we experimentally found that MIDAS is capable of revealing underlying effects of different shape and spatial extent across multiple parameter settings. The robustness of MIDAS with respect to varying shape and size of regions sought is primarily due to its inherent adaptive nature in estimating the regionally optimal way to filter the data. This optimal filtering does not smear out the underlying signal while allowing MIDAS to truly delineate sharp multivariate patterns rather than peri-voxel patterns mapped through searchlight [44]. Critically, using functional MRI data in a split sample setting, we showcased the high robustness of the proposed framework as quantified by the reproducibility of the obtained results. Reproducibility, being orthogonal to the measures of sensitivity and specificity, further assures the reliability and robustness of the proposed method.

**Comparison with multivariate feature selection methods**

The output of MIDAS is a spatial map reflecting significant group effects or correlations with non-imaging variables. As such, this map can be used to perform feature selection for a subsequent classification, or regression task using a properly nested cross-validation scheme. In that sense, MIDAS bears similarities to multivariate feature selection methods [66, 96, 126, 53, 129]. These methods are designed to identify a set of appropriate features for making predictions on unseen data. Towards this end, they are often based on elaborate

measures whose null distribution is difficult to estimate. Importantly, as these methods are particularly concerned with maximizing the accuracy of the predictions, they may be influenced by confounding variations in the data, rendering the features uninterpretable with respect to the processes under study [68]. Lastly, they often choose a small set of features, which may not fully reflect the true underlying variability despite their superior prediction performance. On the other hand, MIDAS may not result in improved predictive accuracy, but yields tractable, analytically solvable statistics for interpretable inferences.

**Importance of the choice of local learner**

The choice of the least squares support vector machine as the base learner for the local discriminative analysis is important for the computational efficiency of the proposed framework. The LS-SVM admits a closed form solution, which is estimated as a linear function of clinical variables. This allows for analytically estimating the solution vector's null distribution, which in turn enables the analytical estimation of the distribution of the MIDAS statistic. This is in contrast to several variants of the searchlight family of methods, as well as ODVBA, whose base learners cannot be solved in closed form, thereby requiring costly permutation testing procedures.

Importantly, the choice of the LS-SVM adds to the versatility of the proposed framework. The LS-SVM can tackle both classification and regression designs. Moreover, the LS-SVM can be readily modified to accommodate different regularization terms encoding distinct assumption regarding the nature (*e.g.*, smoothness, or spatial extent) of the underlying signal (see 3.4). Critically, this does not impact the closed form nature of the solution, thus maintaining the benefits of rapid analytical approximations of null distributions.

**Generalization to different study designs and data types**

The proposed framework is designed for cross-sectional studies where each subject provides a single image. However, the framework is general in terms of input imaging modality. In this chapter, we demonstrated its applicability to both structural and functional data. However, the underlying statistical model does not make any further assumptions regarding the nature of the input data and can be applied to a very broad family of statistical parametric mapping tasks. Moreover, while our validation setting was based on classification tasks, MIDAS is also applicable to regression tasks. Our formulation does not make any assumption about the domain of clinical variables, which are not constrained to be binary. As a consequence, one may readily apply MIDAS when aiming to capture effects of interest reflected by continuous variables, such as aging or development. This is an important advantage of MIDAS compared to ODVBA, which is designed for binary scenarios.

**A note about regularization**

MIDAS employs LS-SVM as the base local discriminative learner. In the described formulation, the LS-SVM makes use of the Euclidean norm $\|\mathbf{w}\|_2^2$ to enforce the smoothness of the estimated weights. Nonetheless, this choice does not preclude the use of different regularization terms, which could better encode the nature of the imaging data. Such a regularization term is $\mathbf{w}^T \Sigma \mathbf{w}$, which enforces nearby voxels to carry similar weights [12], potentially improving the quality of the resulting statistical brain maps.

## Accounting for covariate effects

A practical feature of the MIDAS framework is that the use of local linear learner admits explicit covariate effect corrections as derived in [68] and explained in detail in 3.2.8. This procedure enables the analysis of datasets with non-uniform distribution of covariates, whose effects would otherwise bias the resulting statistical parametric maps. This property of MIDAS is in stark contrast with ODVBA and searchlight family of methods, which necessitate the prior correction of the covariate effects. The latter may be problematic if the covariates to be corrected are not uniformly distributed with respect to the groups of interest.

## Limitations and extensions

MIDAS in its formulation assumes a linear relation between clinical variables and the imaging features where the statistical mapping is to be performed. While this assumption is mainly made to facilitate the use of the analytical estimation of the null distribution for fast computational speed, it is one of the limitations of MIDAS. Contrastingly, searchlight family of methods can admit non-linear learners such as Gaussian kernels for information mapping and may be more sensitive to non-linear relations between the clinical variables and imaging features.

It is possible to generalize MIDAS statistic to handle non-linear kernels such as Gaussian radial basis function (RBF) kernel and non-differentiable regularizations such as $\ell_1$-norm that induces a sparse prior on model weights. One possible extension of MIDAS is one that utilizes the Gaussian kernel in the local learner formulation. [27] have shown that the non-linear decision boundaries using Gaussian kernels can be locally linearly approx-

imated. While these approaches may have advantages in its ability to generalize a wider class of predictive situations, it incurs a high computational price as present in searchlight family of methods since the estimation of null distribution is not as straightforward as in the linear case and requires permutation testing.

One limitation of MIDAS is that in its current formulation it is only applicable to cross-sectional study designs where each subject provides a single image for analysis. However, it is possible to extend MIDAS to allow paired sample study designs as well as longitudinal studies by utilizing difference maps and longitudinal slopes as input features. Specifically, to emulate a paired statistical test, a group of difference images can be contrasted against a group of a commensurate number of empty images using the current MIDAS implementation. Furthermore, to emulate a longitudinal study where each subject provides a set of images over the course of time, the slopes and intercepts of subject trajectories can be input to MIDAS to result in corresponding slope and intercept statistical maps. To fully take into account paired and longitudinal study designs requires an alternative loss function in equation (3.1) and is an interesting future direction.

Lastly, another limitation of MIDAS is that it can handle only a single imaging modality in its analysis. A future direction of work is to incorporate multiple kernel methods [61] in the base learners of MIDAS to handle multi-modal datasets such as imaging and genetic or MRI and positron emission tomography (PET) imaging.

**Conclusion**

In conclusion, we have shown in this chapter that it is possible to efficiently obtain high-quality brain maps by exploiting locally linear discriminative analysis and analytic ap-

proximations of permutation tests. We experimentally demonstrated that MIDAS bears

important advantages compared to commonly used brain mapping techniques, underlin-

ing its potential value in neuroimaging studies.

# Chapter 4

# Inference in the presence of confounds: Generative discriminative machine (GDM)

## 4.1 Introduction

Voxel-based analysis [3] of imaging data has enabled the detailed mapping of regionally specific effects, which are associated with either group differences or continuous non-imaging variables, without the need to define *a priori* regions of interest. This is achieved by adopting a generative model that aims to explain signal variations as a function of categorical or continuous variables of clinical interest. Such a model is easy to interpret. However, it does not fully exploit the available data since it ignores correlations between different brain regions [31].

Conversely, supervised multivariate pattern analysis methods take advantage of de-

pendencies among image elements. Such methods typically adopt a discriminative setting to derive multivariate patterns that best distinguish the contrasted groups. This results in improved sensitivity and numerous approaches have been proposed to efficiently obtain meaningful multivariate brain patterns [89, 131, 126, 29, 64, 53]. However, such approaches suffer from certain limitations. Specifically, their high expressive power often results in overfitting due to modeling spurious distracter patterns in the data [68]. Confounding variations may thus limit the application of such models in multi-site studies [125] that are characterized by significant population or scanner differences, and at the same time hinder the interpretability of the models. This limitation is further emphasized by the lack of analytical techniques to estimate the null distribution of the model parameters, which makes statistical inference costly due to the requirement for permutation tests.

Hybrid generative discriminative models have been proposed to improve the interpretability of discriminative models [104, 9]. However, these models also do not have analytically obtainable null distribution, which makes challenging the assessment of the statistical significance of their model parameters. Last but not least, their solution is often obtained through non-convex optimization schemes, which reduces reproducibility and out-of-sample prediction performance.

To tackle the aforementioned challenges, we propose a novel framework termed *generative-discriminative machine* (GDM), which aims to obtain a multivariate model that is both accurate in prediction and whose parameters are interpretable. GDM combines ridge regression[70] and ordinary least squares (OLS) regression to obtain a model that is both discriminative, while at the same time being able to reconstruct the imaging features using a low-rank approximation that involves the group information. Importantly, the proposed model admits

a closed-form solution, which can be attained in dual space, reducing computational cost. The closed form solution of GDM further enables the analytic approximation of its null distribution, which makes statistical inference and p-value computation computationally efficient.

We validated the GDM framework on two large datasets. The first consists of Alzheimer's disease (AD) patients (n=415), while the second comprises Schizophrenia (SCZ) patients (n=853). Using the AD dataset, we demonstrated the robustness of GDM under varying confounding scenarios. Using the SCZ dataset, we effectively demonstrated that GDM could handle multi-site data without overfitting to spurious patterns, while at the same time achieving advantageous discriminative performance.

## 4.2 Method

### 4.2.1 Generative Discriminative Machine:

GDM aims to obtain a hybrid model that can both predict group differences and generate the underlying dataset. This is achieved by integrating a discriminative model (i.e., ridge regression [70]) along with a generative model (i.e., ordinary least squares regression (OLS)). Ridge and OLS are chosen because they can readily handle both classification and regression problems while admitting a closed form solution.

Let $\mathbf{X} \in R^{n \times d}$ denote the $n$ by $d$ matrix that contains the $d$ dimensional imaging features of $n$ independent subjects arranged row-wise. Likewise, let $\mathbf{Y} \in R^n$ denote the vector that stores the clinical variables of the corresponding $n$ subjects. GDM aims to relate the imaging features $\mathbf{X}$ with the clinical variables $\mathbf{Y}$ using the parameter vector $\mathbf{J} \in R^d$ by optimizing

116

the following objective:

$$\min_{\mathbf{J}} \underbrace{\|\mathbf{J}\|_2^2 + \lambda_1 \|\mathbf{Y} - \mathbf{XJ}\|_2^2}_{\text{ridge}} + \underbrace{\lambda_2 \|\mathbf{X}^T - \mathbf{JY}^T\|_2^2}_{\text{OLS}}. \tag{4.1}$$

If we now take into account information from k additional covariates (e.g., age, sex or other clinical markers) stored in $C \in R^{n \times k}$, we obtain the following GDM objective:

$$\min_{\mathbf{J}, \mathbf{W}_0, \mathbf{A}_0} \underbrace{\|\mathbf{J}\|_2^2 + \lambda_1 \|\mathbf{Y} - \mathbf{XJ} - C\mathbf{W}_0\|_2^2}_{\text{ridge}} + \underbrace{\lambda_2 \|\mathbf{X}^T - \mathbf{JY}^T - \mathbf{A}_0 C^T\|_2^2}_{\text{OLS}}, \tag{4.2}$$

where $\mathbf{W}_0 \in R^k$ contains the bias terms and $\mathbf{A}_0 \in R^{d \times k}$ the regression coefficients pertaining to their corresponding covariates. The inclusion of the bias terms in the ridge regression term allows us to preserve the direction of the parameter vector that imaging pattern that distinguishes between the groups, while at the same time achieving accurate subject-specific classification by taking into account each sample's demographic and other information. Similarly, the inclusion of additional coefficients in the OLS term allows for reconstructing each sample by additionally taking into account its demographic or other information. Lastly, the hyperparameters $\lambda_1$ and $\lambda_2$ control the trade-off between discriminative and generative models, respectively.

In figure 4.1, we demonstrate the trade-off between obtaining an interpretable generative model that captures the entirety of the underlying effects versus a discriminative model that captures the minimal number of features to achieve a prediction. The key concept of the GDM model is that through enforcing a discriminative model to be interpretable, we may avoid overfitting and obtain better predictive performance.
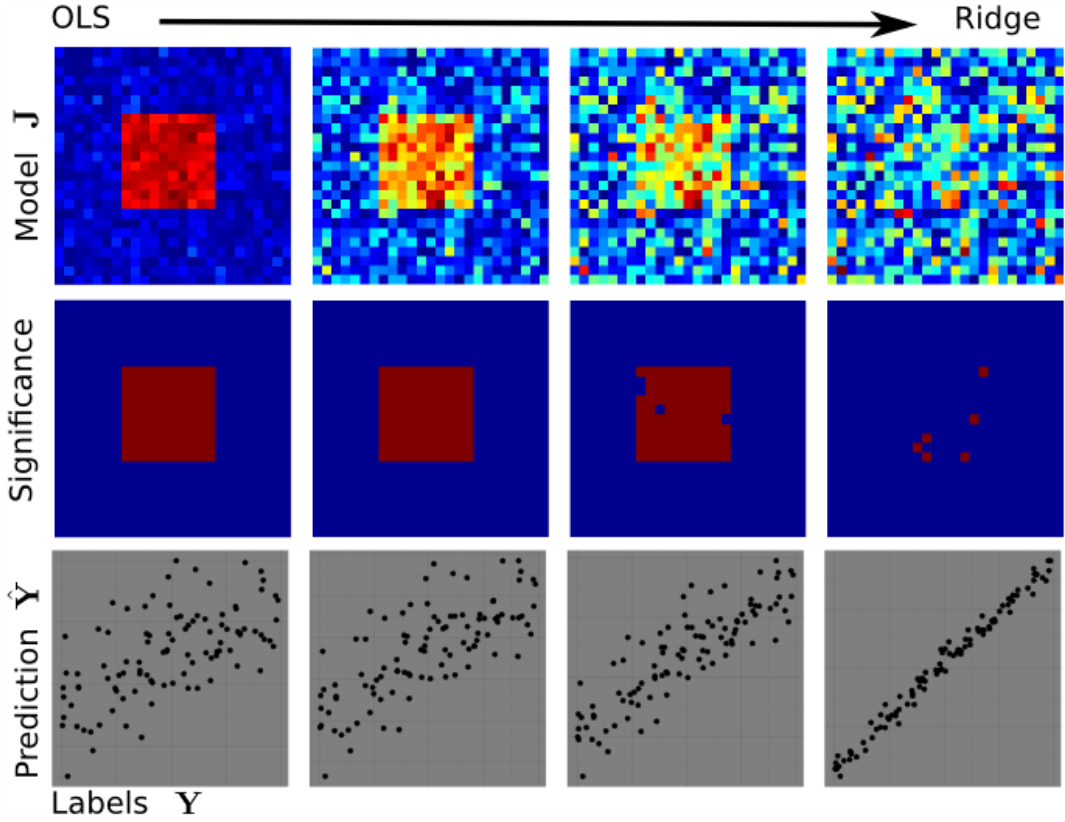
Figure 4.1: A demonstration of the GDM framework on a simulated dataset that comprises of a control group of uniform random data and a "patient" group that exhibits a square pattern of correlated features. Top row illustrates the GDM model weights **J**, while the middle row shows the spatial locations that pass statistical significance testing. The bottom row compares the group predictions $\hat{\mathbf{Y}}$ with the true groups **Y**. Left to right progression illustrates the effect of increasing the discriminative penalty $\lambda_1$ on both the interpretability of the GDM model and the prediction accuracy. Higher generative penalty $\lambda_2$ (towards left) yields a model that captures the underlying square effect while a higher discriminative penalty (towards right) yields a model that better predicts **Y**. The goal of GDM is to fine-tune the trade-off between interpretability and prediction accuracy.

### 4.2.2 Closed form solution:

The formulation in Eq. 4.2 is optimized by the following closed form solution:

$$\mathbf{J} = \left[\mathbf{I} + \lambda_1(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}) + \lambda_2(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y})\right]^{-1}$$

$$\times \left[(\lambda_1 + \lambda_2)(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y})\right], \tag{4.3}$$

which requires a $d \times d$ matrix inversion that can be costly in neuroimaging settings. To account for that, we solve Eq. 4.2 in the subject space using the following dual variables $\mathbf{\Lambda} \in R^n$:

$$\mathbf{\Lambda} = \mathbf{M}^{-1}_{[1:n,1:n]}\left(\mathbf{I} + \frac{\lambda_2 \mathbf{XX}^T \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1}\mathbf{C}^T - \lambda_2 \mathbf{XX}^T}{1 + \lambda_2(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y})}\right)\mathbf{Y}, \tag{4.4}$$

where $\mathbf{M}$ is the following $n + k \times n + k$ matrix:

$$\mathbf{M} = \begin{bmatrix} -\frac{\mathbf{XX}^T}{1+\lambda_2(\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y})} - \mathbf{I}/\lambda_1 & \mathbf{C} \\ \mathbf{C}^T & 0 \end{bmatrix}. \tag{4.5}$$

The dual variables $\mathbf{\Lambda}$ can be used to solve $\mathbf{J}$ using the following equation:

$$\mathbf{J} = \frac{\lambda_2 \mathbf{X}^T\mathbf{Y} - \lambda_2\mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y} - \mathbf{X}^T\mathbf{\Lambda}}{1 + \lambda_2(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y})}. \tag{4.6}$$

Once the solution for $\mathbf{J}$ has been obtained, the bias terms $\mathbf{W}_0$ and the regression coefficients $\mathbf{A}_0$ can be obtained using the following equations:

$$\mathbf{A}_0 = \mathbf{X}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} - \mathbf{J}\mathbf{Y}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1} \tag{4.7}$$

$$\mathbf{W}_0 = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y} - (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{XJ}$$

### 4.2.3 Analytic approximation of null distribution:

Using the dual formulation, the GDM parameters $\mathbf{J}$ can be shown to be a linear combination of the group labels $\mathbf{Y}$ and the following matrix $\mathbf{Q}$:

$$\mathbf{Q} = \frac{\lambda_2 \mathbf{X}^T - \lambda_2 \mathbf{X}^T C(C^T C)^{-1} C^T - \mathbf{X}^T M_{[1:n,1:n]}^{-1}\left(\mathbf{I} + \frac{\lambda_2 \mathbf{X}\mathbf{X}^T C(C^T C)^{-1} C^T - \lambda_2 \mathbf{X}\mathbf{X}^T}{1 + \lambda_2 (\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T C(C^T C)^{-1} C^T \mathbf{Y})}\right)}{1 + \lambda_2 (\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T C(C^T C)^{-1} C^T \mathbf{Y})},$$

(4.8)

such that $\mathbf{J} = \mathbf{Q}\mathbf{Y}$. It is shown in 4.2.4 that $\mathbf{Q}$ is approximately invariant to permutation operations on $\mathbf{Y}$. Assuming $\mathbf{Y}$ is zero mean, unit variance yields that $\mathrm{E}(J_i) = 0$ and $\mathrm{Var}(J_i) = \sum_j Q_{i,j}^2$ under random permutations of $\mathbf{Y}$ approximated by random draws. As shown in [151], asymptotically this yields that

$$J_i \xrightarrow{D} \mathcal{N}\left(0, \sqrt{\sum_j Q_{i,j}^2}\right),$$

(4.9)

which allows efficient statistical inference on the parameter values of $J_i$. Specifically, statistical significance of $J_i$ can be obtained by performing z-test on $\frac{J_i}{\sqrt{\sum_j Q_{i,j}^2}}$.

### 4.2.4 Permutation invariance of the parametric matrix

The only appearance of the permuted labels $\mathbf{Y}$ in the parametric $\mathbf{Q}$ matrix described in equation 4.8 comes in the form of $\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T C(C^T C)^{-1} C^T \mathbf{Y}$ where $C(C^T C)^{-1} C^T$ is a rank deficient projection matrix with rank of $k$. We show here that $\mathbf{Y}_r^T\mathbf{Y}_r - \mathbf{Y}_r^T C(C^T C)^{-1} C^T \mathbf{Y}_r$ where $\mathbf{Y}$ is a random variable under random draws of elements of $\mathbf{Y}$ is concentrated around $n - k$ with high probability, thus $\mathbf{Q}$ is approximately invariant to randomness in permutations

120

of $\mathbf{Y}$.

**Theorem 1.** *Let $\mathbf{Y} \in \mathbf{R}^n$ be a vector such that $E(Y_i) = 0$ and $Var(Y_i) = 1$ under random permuta-tions. Let $\mathbf{M} \in \mathbf{R}^{n \times n}$ be rank $n - k$ projection matrix where $k < n$ and $k$ is fixed. Then as $n \to \infty$,*

$$P\left( \left| \frac{\mathbf{Y}^T \mathbf{M} \mathbf{Y}}{n} - \frac{n-k}{n} \right| > \sqrt{\frac{2}{n}} \right) \le \frac{k}{n}$$

*Proof.* In [8], it is shown that $E(\mathbf{Y}^T \mathbf{M} \mathbf{Y}) = \text{tr}(\mathbf{M}\Sigma) + \boldsymbol{\mu}^T \mathbf{M} \boldsymbol{\mu}$ where $\Sigma = \text{Cov}(\mathbf{Y})$ and $\boldsymbol{\mu} = E(\mathbf{Y})$. Since $n$ is assumed to be large, $\Sigma \to \mathbf{I}_n$. Furthermore, $\boldsymbol{\mu} = 0$ as given. Therefore,

$$E(\mathbf{Y}^T \mathbf{M} \mathbf{Y}) = \text{tr}(\mathbf{M} \mathbf{I}_n) = n - k \tag{4.10}$$

since $\mathbf{M}$ is a projection matrix of rank $n - k$.

The variance of $\mathbf{Y}^T \mathbf{M} \mathbf{Y}$ can be analyzed by first decomposing $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$ where $\mathbf{H}$ is a projection matrix of rank $k$. Using this yields:

$$\text{Var}(\mathbf{Y}^T \mathbf{M} \mathbf{Y}) = \text{Var}(\mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{H} \mathbf{Y})$$

$$= \text{Var}(\mathbf{Y}^T \mathbf{H} \mathbf{Y}) \tag{4.11}$$

since $\mathbf{Y}^T \mathbf{Y}$ is constant under random permutations of $\mathbf{Y}$. Furthermore, the variance of $\mathbf{Y}^T \mathbf{H} \mathbf{Y}$ is upper bounded by the variance of $\mathbf{Y}_g^T \mathbf{H} \mathbf{Y}_g$ where $\mathbf{Y}_g$ are a multivariate Gaussian random variable with same mean and variance as $\mathbf{Y}$ since $\mathbf{Y}^T \mathbf{H} \mathbf{Y}$ is a subgaussian random variable due to finite support. As shown in [127],

$$\text{Var}(\mathbf{Y}_g^T \mathbf{H} \mathbf{Y}_g) = 2\text{tr}(\mathbf{H}\Sigma\mathbf{H}\Sigma) + 4\boldsymbol{\mu}^T \mathbf{H}\Sigma\mathbf{H}\boldsymbol{\mu}$$

$$= 2\mathrm{tr}(\boldsymbol{H}\mathbf{I}_n\boldsymbol{H}\mathbf{I}_n)$$

$$= 2\mathrm{tr}(\boldsymbol{H}\boldsymbol{H})$$

$$= 2\mathrm{tr}(\boldsymbol{H})$$

$$= 2k$$

$$\geq \mathrm{Var}(\mathbf{Y}^T \boldsymbol{H}\mathbf{Y}) \tag{4.12}$$

Note that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ since $\boldsymbol{H}$ as a projection matrix is idempotent. We can invoke Chebyshev's inequality to demonstrate the concentration of $\mathbf{Y}^T\boldsymbol{M}\mathbf{Y}$:

$$P\left(\left|\mathbf{Y}^T\boldsymbol{M}\mathbf{Y} - (n-k)\right| > \alpha\sqrt{2k}\right) \leq P\left(\left|\mathbf{Y}^T\boldsymbol{M}\mathbf{Y} - (n-k)\right| > \alpha\,\mathrm{Var}(\mathbf{Y}^T\boldsymbol{M}\mathbf{Y})\right)$$

$$\leq \frac{1}{\alpha^2} \tag{4.13}$$

If $\alpha$ is set as $\sqrt{\frac{n}{k}}$ then we get:

$$P\left(\left|\mathbf{Y}^T\boldsymbol{M}\mathbf{Y} - (n-k)\right| > \sqrt{2n}\right) \leq \frac{k}{n}$$

$$\Downarrow$$

$$P\left(\left|\frac{\mathbf{Y}^T\boldsymbol{M}\mathbf{Y}}{n} - \frac{n-k}{n}\right| > \sqrt{\frac{2}{n}}\right) \leq \frac{k}{n} \tag{4.14}$$

$\square$
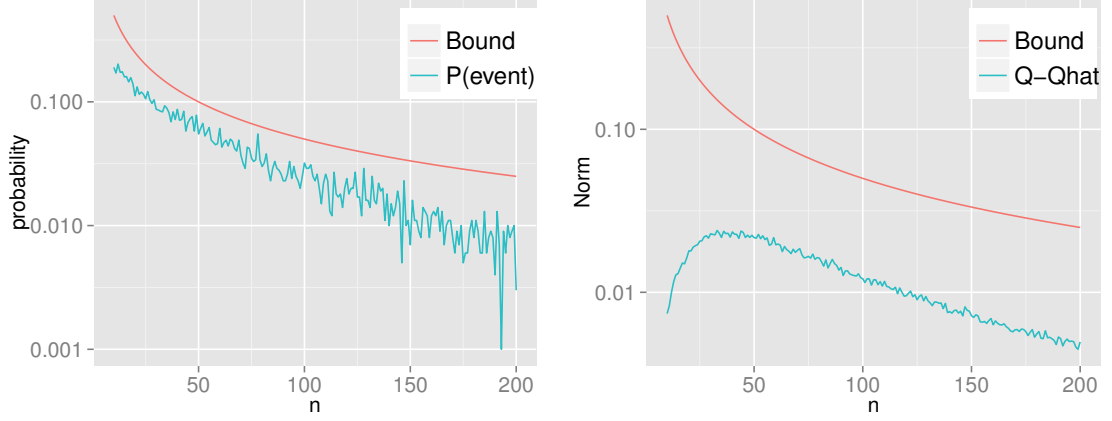
Figure 4.2: Left: The simulated probability of the event $\mathbf{I}\left(\left|\frac{\mathbf{Y}^T M \mathbf{Y}}{n} - \frac{n-k}{n}\right| > \sqrt{\frac{2}{n}}\right)$ for $Y \in R^n$ for $n = 10,\ldots,200$ and the upper bound $O(1/n)$. Right: The deviation of $Q$ from $\hat{Q}$: $\|Q - \hat{Q}\|_F$ and the upper bound $O(1/n)$

Therefore, as $n \to \infty$ and $k$ is fixed, $\mathbf{Q}$ matrix in Eq. 4.8 concentrates around $\hat{\mathbf{Q}}$ where

$$\hat{\mathbf{Q}} = \left[\mathbf{I}(1 + \lambda_2(n-k)) + \lambda_1(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T C(C^T C)^{-1}C^T\mathbf{X})\right]^{-1}\left[(\lambda_1 + \lambda_2)(\mathbf{X}^T - \mathbf{X}^T C(C^T C)^{-1}C^T)\right]$$

(4.15)

is invariant to permutations in $\mathbf{Y}$.

To confirm the validity of Eq. 4.14, we plot the expectation of the event $\mathbf{I}\left(\left|\frac{\mathbf{Y}^T M \mathbf{Y}}{n} - \frac{n-k}{n}\right| > \sqrt{\frac{2}{n}}\right)$ for $\mathbf{Y}$ drawn from a uniform distribution $U[-1,1]^n$ for $n = 10,\ldots,200$ for 1000 repetitions while $M = C(C^T C)^{-1}C^T$ is a rank 5 projection matrix i.e. $C \in R^{n\times 5}$. The expectation of the event $\mathbf{I}\left(\left|\frac{\mathbf{Y}^T M \mathbf{Y}}{n} - \frac{n-k}{n}\right| > \sqrt{\frac{2}{n}}\right)$ and the provided upper bound is plotted in figure 4.2.

## 4.3 Experimental validation

We compared GDM with a purely discriminative model, namely ridge regression [70], as well as with its generative counter-part, which was obtained through the procedure

outlined by Haufe et al. [68]. We chose these methods because their simple form allows the computation of their null distribution, which in turns enables the comparison of the statistical significance of their parameter maps.

We used two large datasets in two different settings. First, we used a subset of the ADNI study, consisting of 228 controls (CN) and 187 Alzheimer's disease (AD) patients, to evaluate out-of-sample prediction accuracy and reproducibility. Second, we used data from a multi-site Schizophrenia study, which consisted of 401 patients (SCZ) and 452 controls (CN) spanning three sites (USA n=236, China n=286, and Germany n=331), to evaluate the cross-site prediction and reproducibility of each method.

For all datasets, T1-weighted MRI volumetric scans were obtained at 1.5 Tesla. The images were pre-processed through a pipeline consisting of (1) skull-stripping; (2) N3 bias correction; and (3) deformable mapping to a standardized template space. Following these steps, a low-level representation of the tissue volumes was extracted by automatically partitioning the MRI volumes of all participants into 151 volumetric regions of interest (ROI). The ROI segmentation was performed by applying a multi-atlas label fusion method [37]. The derived ROIs were used as the input features for all methods.

### 4.3.1 Analytical approximation of p-values

To confirm that the analytical approximation of null distribution of GDM is correct, we estimated the p-values through the approximation technique as well as through permutation testing. A range of 10 to 10,000 permutations was applied to observe the error rate. This experiment was performed on the ADNI dataset. The results displayed in figure 4.3 demonstrate that the analytic approximation holds with approximately $O(1/\sqrt{\#\text{permutations}})$ er-

124
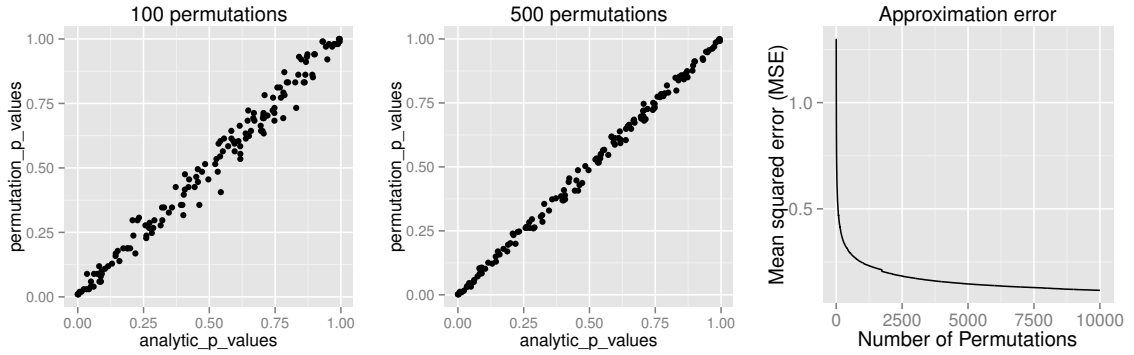
Figure 4.3: Comparison of permutation based p-values of GDM with their analytic approximations at varying permutation levels.

ror.

### 4.3.2 Out-of-sample prediction and reproducibility

To assess the discriminative performance and reproducibility of the compared methods under varying confounding scenarios, we used the ADNI dataset. We simulated four distinct training scenarios in increasing potential for confounding effects:

- Case 1: 50% AD + 50% CN subjects, mean age balanced

- Case 2: 75% CN + 25% AD, mean age balanced

- Case 3: 50% AD + 50% CN, oldest ADs, youngest CNs

- Case 4: 75% CN + 25% AD, oldest ADs, youngest CNs.

All models had their respective parameters cross-validated in an inner fold before performing out-of-sample prediction on a left out test set consisting of equal numbers of AD and CN subjects with balanced mean age. Furthermore, the inner product of training model parameters was compared between folds to assess the reproducibility of models. The trade-off between the reproducibility and prediction accuracy in training sets can be seen in figure 4.4. Training and testing folds were shuffled 100 times to yield a distribution.
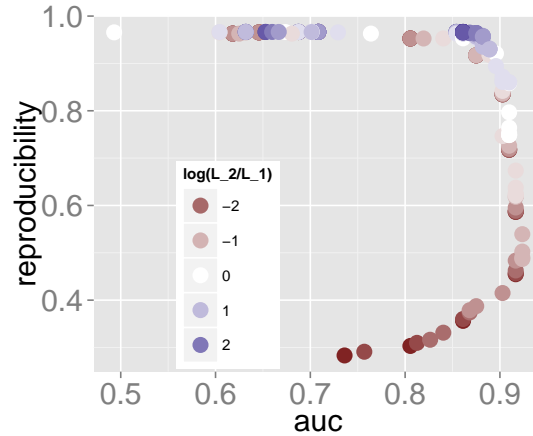
Figure 4.4: The trade-off between reproducibility and prediction accuracy in the GDM model under varying parameter combinations. The y-axis denotes the reproducibility of the GDM model computed by taking the average normalized inner-product of the vector $J$ across 10-fold cross-validation. The x-axis displays the prediction accuracy computed by the training AUC of the predictions made by using the $J$ vector obtained at a particular parameter combination $\lambda_1, \lambda_2$. The color scale denotes the ratio of $\lambda_2/\lambda_1$ where colder colors indicate a more generative model while warmer colors indicate a more discriminative model.

The prediction accuracies and the model reproducibility for the above cases are shown in figure 4.5. The results demonstrate that while GDM is not a purely discriminative model, its predictions outperformed ridge regression in all four cases. Regarding reproducibility, the Haufe et al. (2013) procedure yielded the most stable models since it yields a purely generative model. However, GDM was more reproducible than ridge regression.

**Multi-site study**

To assess the predictive performance of the compared methods in a multi-site setting, we used the Schizophrenia dataset that comprises data from three sites. All models had their respective parameters cross-validated while training in one site before making predictions in the other two sites. Each training involved using 90% of the site samples to allow for resampling the training sets 100 times to yield a distribution. The reproducibility across

126
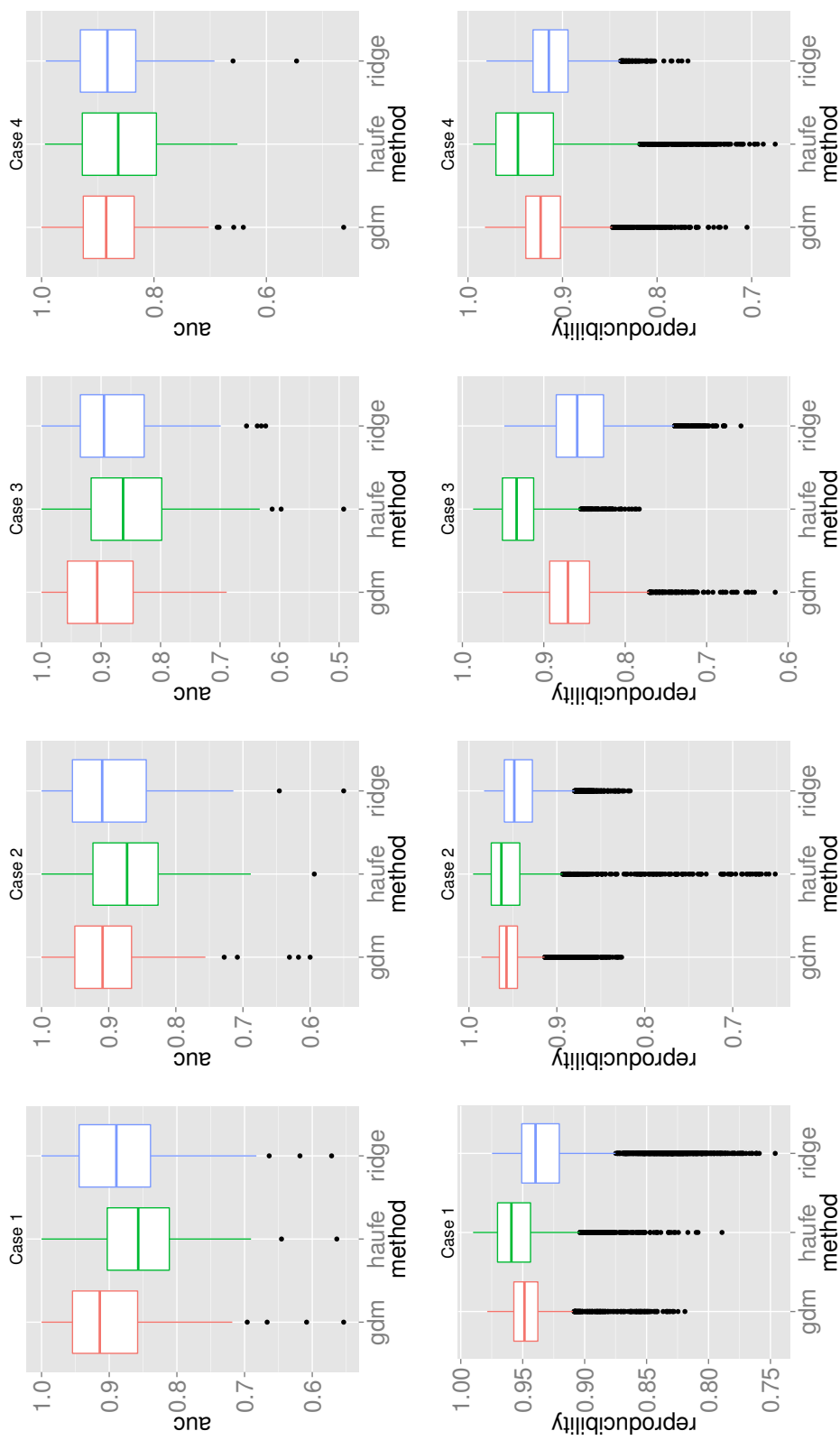
Figure 4.5: Cross validated out-of-sample AD vs. CN prediction accuracies (top row) and normalized inner-product reproducibility of training models (bottom row) for varying training scenarios and all compared methods.

the resampled sets was measured using the inner product between model parameters. The multi-site prediction and reproducibility results are visualized in figure 4.7.

In five out of six cross-site prediction settings, GDM outperformed all compared methods in terms accuracy. Also, GDM had higher reproducibility than ridge regression, while having slightly lower reproducibility than the generative procedure in Haufe et al. (2013).

**Statistical maps and p-values**

To qualitatively assess and explain the predictive performance of the compared methods for the AD vs. CN scenario, we computed the model parameter maps using full resolution gray matter tissue density maps for the ADNI dataset (Fig. 4.6 top). Furthermore, since the null distribution of GDM, as well as ridge regression, can be estimated analytically, we computed p-values for the model parameters and displayed the regions surviving false discovery rate (FDR) correction [11] at level $q < 0.05$ (Fig. 4.6 bottom).

The statistical maps demonstrated that both GDM and Haufe procedure yield patterns that accurately delineate the regions associated with AD, namely the widespread atrophy present in the temporal lobe, amygdala, and hippocampus. This is in contrast with the patterns found in ridge regression that resemble a hard to interpret speckle pattern with meaningful weights only on the hippocampus. This once again confirmed the tendency of purely discriminative models to capture spurious patterns. Furthermore, the p-value maps of the Haufe method and ridge regression demonstrate the wide difference between features selected by generative and discriminative methods and how GDM strikes a balance between the two to achieve superior predictive performance.
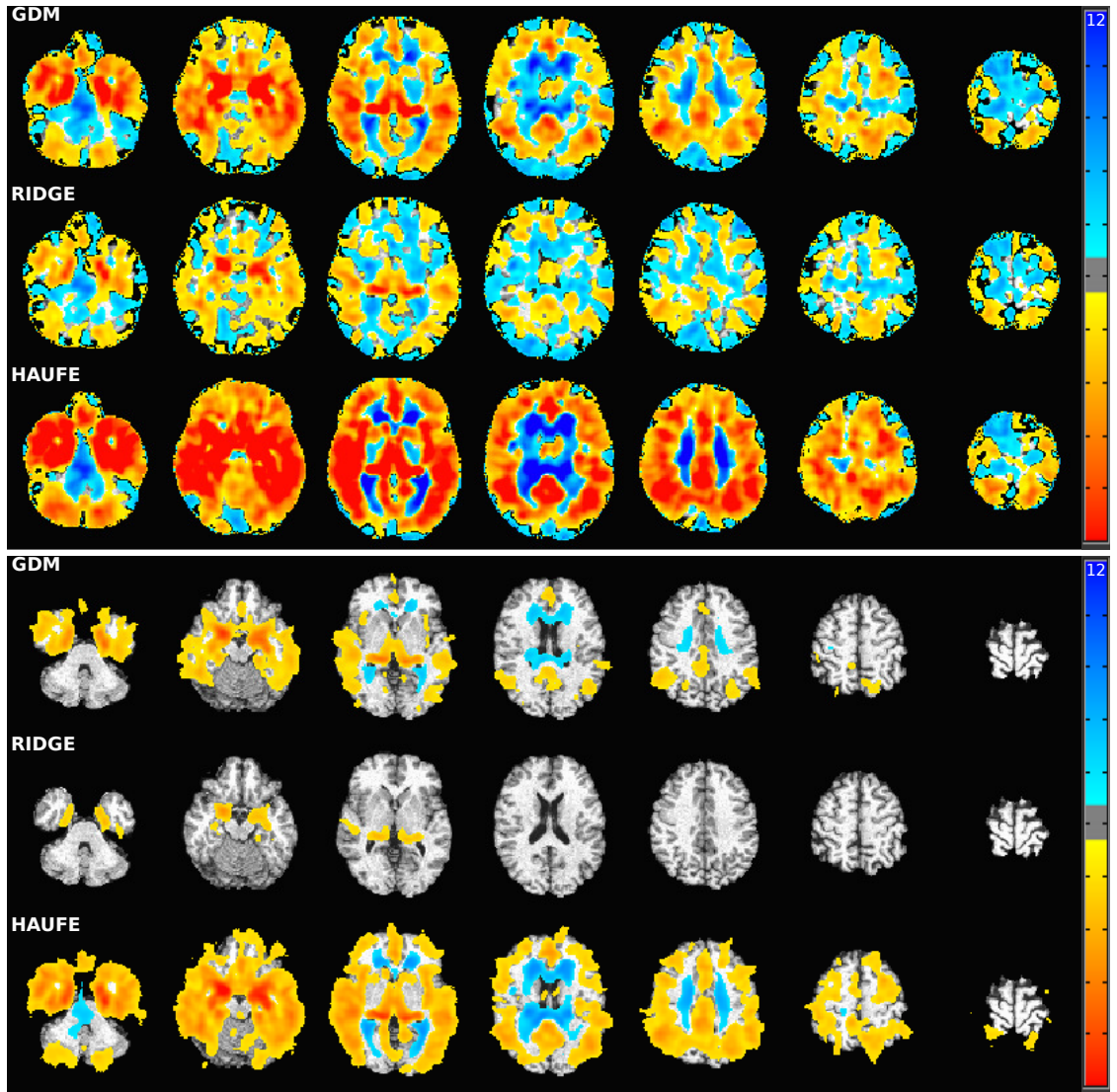
Figure 4.6: Top: Normalized parameter maps of compared methods for discerning group differences between AD patients and controls. Bottom: Parameter $\log_{10}$ p-value maps of the compared methods for discerning group differences between AD patients and controls after FDR correction at level $q < 0.05$. Warmer colors indicate decreasing volume with AD, while colder colors indicate increasing volume with AD.

## 4.4   Discussion & Conclusion

The interpretable patterns captured by GDM coupled with its ability to outperform discriminative models in terms of prediction underline its potential for neuroimaging analysis. We demonstrated that GDM obtains highly reproducible models through gener-
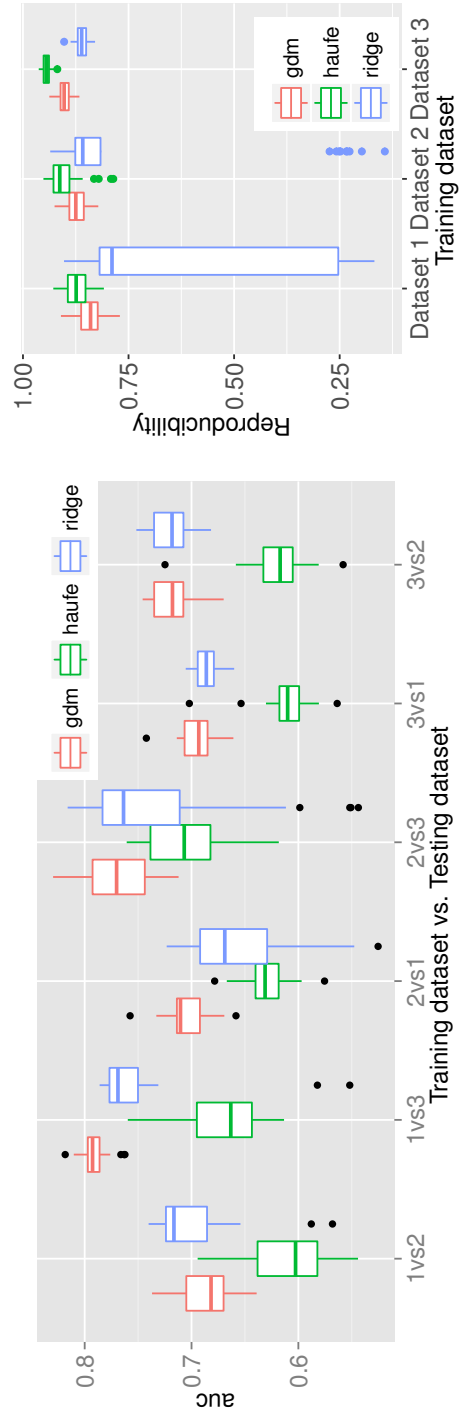
Figure 4.7: Cross validated multi-site SCZ vs. CN prediction accuracies (left) and normalized inner-product reproducibility of training models (right) for all compared methods.

ative modeling, thus avoiding overfitting that is commonly observed in neuroimaging settings. Overfitting is especially evident in multi-site situations, where discriminative models might subtly model spurious dataset effects and perform poorly in an out-of-site setting. Furthermore, by using a formulation that yields a closed form solution, we additionally demonstrated that is possible to assess the statistical significance of the model parameters efficiently.

While the methodology presented herein is analogous to generatively regularizing ridge regression with ordinary least squares regression, the framework proposed can be generalized to include generative regularization in other commonly used discriminative learning methods. Namely, it is possible to augment linear discriminant analysis (LDA), support vector machine (SVM), artificial neural network (ANN) objective with a similar generative term to yield an alternative generative discriminative model of learning. However, the latter two cases would not permit a closed form solution, making it impossible to estimate a null distribution analytically.

# Chapter 5

# Summary and future work

## Synopsis

Group studies in neuroimaging have the potential to elucidate the workings of the human brain in health and disorder. However, both univariate and multivariate analysis techniques, which are used to extract the differentiating patterns between control and patient groups, are limited by ad-hoc assumptions regarding the homogeneity and spatial uniformity of disease effects or confounds in the study sample. In this thesis, we have introduced three novel pattern analysis methods that allow us to move beyond these assumptions and limitations, thus allowing us to make use of the rich imaging data fully and enabling more powerful inferences.

In Chapter 2, we proposed a method that allows us to move beyond the assumption that there is a single imaging pattern of brain differences that discriminates patients from controls. To disentangle heterogeneous disease patterns, we developed a method termed *HYDRA* that optimizes a piecewise linear decision boundary between the control popu-

lation and the patient population. The piecewise linear boundary was shown to induce a subgrouping of the patient population that is informed by the differences of the patients from the controls, effectively yielding a supervised clustering solution. This feature contrasts HYDRA with unsupervised clustering methods that group patients based on their similarities to each other rather than their differences with respect to controls. Furthermore, we provided optimization routines for HYDRA in the dual domain to handle high dimensional neuroimaging settings. We validated the HYDRA algorithm on synthetic data with the known ground truth. Then, we applied HYDRA to an imaging and genetic study of Alzheimer's disease, which revealed novel data-driven anatomical and genetic subtypes of Alzheimer's disease.

In Chapter 3, we refuted the assumption that the spatial extent and the shape of the underlying disease effect is uniform across the brain, which justified the use of a single bandwidth filter to smooth the imaging data. We demonstrated that the traditional method of uniformly smoothing the entire anatomy is suboptimal in extracting the underlying signal in a highly specific and sensitive manner. To address this, we introduced the MIDAS algorithm, which optimizes for local adaptive filters that cover the brain volume. The local filters are formulated as linear discriminative models that are designed to maximize the differences between groups. We additionally showed that the coefficients of local adaptive filters could be used to construct a statistic whose null distribution can be analytically estimated, which enables efficient statistical significance testing. Furthermore, MIDAS was validated extensively using simulated atrophy experiments and was applied to both structural and functional MRI studies. In comparative scenarios against standard pattern analysis tools, MIDAS was shown to have higher sensitivity and specificity in un-

covering the underlying patterns. Furthermore, we showed that MIDAS could be utilized in both binary group comparison settings as well as regression settings with continuous clinical labels, making it pertinent for a wide range of neuroimaging studies.

Lastly, in Chapter 4, we identified that the current pattern analysis tools rely on the limitation that the control and disease groups are matched for covariates or the assumption that the unmatched covariates do not confound the group differences. To move beyond these limitations and assumptions, we introduced a methodology termed *generative discriminative machine* (GDM) that can generatively model the effects of confounds, while providing a discriminative model that is invariant to confounding effects. Our derivations showed that the model coefficients of GDM follow a null distribution that can be analytically estimated, which enables efficient statistical significance testing. Furthermore, the optimization of GDM was shown to be viable in the dual domain, which allows us to take advantage of the low sample size high dimensionality setting to improve computational speed. To demonstrate the utility of GDM, we have applied it to two large structural MRI studies. First, we applied it to a study of Alzheimer's disease patients and controls. There we artificially introduced confounds by resampling the dataset to reflect large covariate differences between the control group and the patients. In this setting, we showed that GDM was able to counteract the effects of confounds and more accurately discriminate between the controls and patients than traditional supervised learning methods. In addition, GDM was able to yield qualitatively more interpretable statistical maps compared to standard discriminative MVPA methods, accurately capturing meaningful anatomical structures. Next, we applied GDM to a study of Schizophrenia that spans multiple datasets from various countries. The confounding variations in multi-site studies is a significant

obstacle in the way of advancing big neuroimaging data analysis. In this setting, we also demonstrated that GDM was able to effectively handle the difficult confound due to site differences and to yield accurately discriminative and interpretable patterns.

## Future Work

The work presented in this thesis has advanced our capability to analyze neuroimaging studies by allowing us to move beyond the common assumptions made by standard analytical tools. However, there are several avenues for improving and extending the presented tools, which are left for future work. We detail below some of these directions.

### Regional multivariate inference for connectivity studies

The general optimization objective of MIDAS and corresponding statistic are suitable for any type data. Applying MIDAS to functional connectivity studies would potentially help amplify the detection of underlying patterns, which are relatively less pronounced and more difficult to detect than the structural patterns observed in neurodegenerative diseases. The input for this application would be the vectorized lower-triangular component of the connectivity graph matrices. The main challenge would be to determine the topology of the graph neighborhoods that would be sampled to replicate the overlapping MIDAS neighborhood sampling routine. These can be based on a seed node and the $\epsilon$-graph neighborhood around that node. Once graph neighborhoods have been determined, the node wise statistic can be derived using the same procedure as in equation (3.12).

## Non-linear regional multivariate inference

MIDAS is ultimately based on linear models to obtain analytically approximated null distributions. It would be of interest to utilize non-linear models as the base learner in regional discriminative analysis to handle a higher level complexity of features. However, non-linear models seldom permit closed-form solutions and thus obtaining an analytic approximation of the null distribution of model coefficients is not straightforward. Inspired by [103], it may be possible to obtain an approximate null distribution for statistics derived from models, such as lasso, that does not have a closed-form solution.

## Globally optimal maximum margin convex polytope using softmax functions

The HYDRA algorithm in Chapter 2 involves an iterative optimization routine that includes an assignment step followed by a hyperplane solution step. The optimization procedure is iterative due to the non-differentiability of the $\max(\cdot)$ function, which is used to assign each sample to the hyperplane that separates it from the control population with a maximum margin. Namely, the assignment variable of a subject is the signifier for the hyperplane that is maximally correlated with the imaging features of that subject. However, if the $\max(\cdot)$ function is relaxed by a function such as $\text{softmax}_\alpha(\mathbf{x})_i = \frac{e^{\alpha x_i}}{\sum_j e^{\alpha x_j}}$ that is differentiable, one can derive a reformulation of equation (2.1) that is globally optimizable without the need for additionally solving for the assignment step.

$$\underset{\{\mathbf{w}_j, b_j\}_{j=1}^K}{\text{minimize}} \sum_{j=1}^K \frac{\|\mathbf{w}_j\|_2^2}{2} + C \sum_i \text{softmax}_\alpha\{0, 1 - y_i(\mathbf{w}_j^T \mathbf{x}_i + b_j)\} \tag{5.1}$$

This formulation, which is convex and differentiable, yields a globally optimal solution, **W**′, which can also be used to induce a clustering on the patient population by using equation (3.3).

**Maximum margin convex regression polytope**

While the current formulation of HYDRA can utilize the dichotomy between patients and controls to cluster the patients, many neuroanatomical processes involve continuous changes rather than binary groupings. Therefore, it would be valuable to extract heterogeneous trajectories of change for processes such as aging or development. To be able to distinguish heterogeneity using continuous scores requires a regression reformulation of HYDRA. This is readily available since HYDRA is a generalization of support vector machines and thus a regression version of HYDRA would be analogously obtained by generalizing support vector regression [38].

**Statistical inference a posteriori to clustering**

One of the challenges encountered in Chapter 2 was the validity of obtaining p-values for statistical significance testing between the control group and the subgroups of patients clustered by HYDRA. While this problem is not unique to the setting of HYDRA, the concept of statistical inference following clustering remains an open problem. The main difficulty is to determine the null distribution of the statistic that is obtained after clustering. A permutation testing procedure can be followed to estimate the null distribution of the resulting statistics.

**Synthesis of GDM, HYDRA, and MIDAS frameworks**

Each of the methods described in this thesis allows us to independently address certain assumptions about the data that limit the statistical inferences we can draw from it. However, in almost all neuroimaging studies, limitations that we address in this thesis appear in synchrony. For example, the imaging patterns of Alzheimer's disease may be both heterogeneous and exhibit spatial non-uniformity. In addition, factors such as scanner differences and survivorship bias may further confound the ADNI dataset. Thus, it is imperative to take into account all of these factors and perform statistical inference using tools that can simultaneously be impervious to their confounding effects. Such a tool requires the combination of the frameworks of GDM, HYDRA, and MIDAS in a principled manner and is an important but challenging direction for future work.

# Appendix A

# Image preprocessing techniques

Neuroimaging data obtained from the scanner cannot be used for our analysis directly. In this section, we describe the common preprocessing steps that were used in all of our experiments.

**Region of interest (ROI) volumetry**

The high dimensionality of MR images hinders their analysis and interpretation. Extracting region of interests (ROI) effectively reduces the dimensionality of the data in an interpretable and anatomically meaningful way. We employed a multi-atlas segmentation algorithm [37] which uses a consensus labeling framework to fuse/integrate segmentation hypotheses generated by warping a broad ensemble of labeled atlases to the target space via the use of several warping algorithms, regularization parameters, and atlases. The label fusion integrates two complementary sources of information: a local similarity ranking to select locally optimal atlases and a boundary modulation term to refine the segmentation consistently with the target images intensity profile. The flowchart of the ROI

algorithm is presented in Figure A.1. In our analyses, we used this algorithm to partition the brain into approximately one hundred disjoint ROIs generated and obtained the volume of each ROI as a feature representation of the brain.
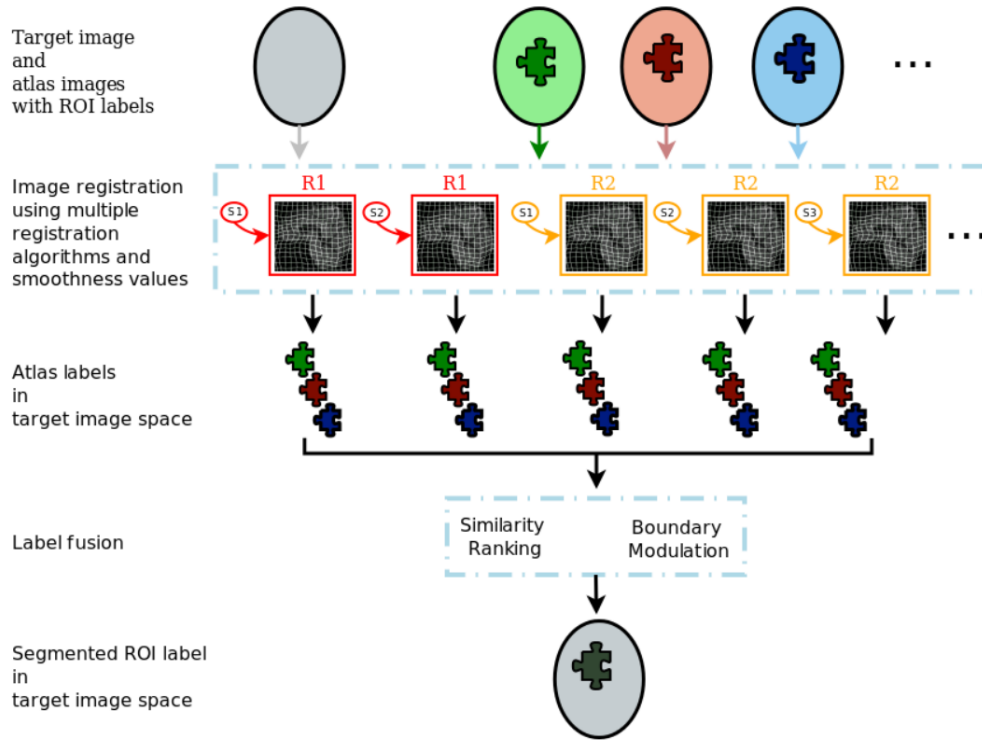


Figure A.1: Multi-atlas region of interest segmentation flowchart.

**Tissue density maps**

ROIs provide us with data in a dimension that we can easily handle in order to parse disease heterogeneity. However, in order to characterize disease processes in greater spatial detail, we employed tissue density maps for subsequent subgroup-analyses. Towards this end, we employed tissue density maps that allow us to characterize disease processes in greater spatial detail. Specifically, we employed a previously published volumetric approach to generate tissue density map for group comparisons [31], termed RAVENS (re-

gional analysis of volumes examined in normalized space) map. The RAVENS maps are obtained with the following procedures. An established deformable registration method [121] is used for warping individual images to a single subject brain template. The brain image scans are segmented into three tissue types: gray matter, white matter and cerebrospinal fluid [100]. RAVENS maps encode, locally and separately for each tissue type, the volumetric changes (local expansion or shrinkage) observed during the registration. They hence have the advantage of accounting for imperfect registration by taking the residual (error) of the imperfect registration into account.

# Appendix B

# Software

All three of the methodologies presented in this thesis have been supplemented with open-source MATLAB implementations shared on MathWorks File Exchange at `https://www.mathworks.com/matlabcentral/fileexchange/`.

**HYDRA**

HYDRA takes a comma separated values (csv) file as input. This file contains the imaging features of the subjects, covariate information, as well as group information. The program yields a clustering of the patient group that is informed by their differences to the control group. A snapshot of the HYDRA command-line is shown in Figure A.1.

**MIDAS**

MIDAS takes a comma separated values (csv) file as input. This file contains the image paths of the subjects, covariate information, as well as group information. The program yields a statistical map of the differences between groups that have undergone optimal

discriminative filtering. A snapshot of the MIDAS command-line is shown in Figure A.2.

**Generative discriminative machine (GDM)**

GDM takes a comma separated values (csv) file as input. This file contains the image paths of the subjects, covariate information, as well as group information (i.e., whether the subject is a control or a patient) or continuous variables for regression. The program yields a discriminative map that distinguishes between groups while providing a low-rank generative approximation of the data. A snapshot of the GDM command-line is shown in Figure A.3.

```
>> hydra
 function returns estimated subgroups by hydra for clustering
 configurations ranging from K=1 to K=10, or another specified range of
 values. The function returns also the Adjusted Rand Index that was
 calculated across the cross-validation experiments and comparing
 respective clustering solutions.

 INPUT

 REQUIRED
 [--input, -i] : .csv file containing the input features. (REQUIRED)
                 every column of the file contains values for a feature, with
                 the exception of the first and last columns. We assume that
                 the first column contains subject identifying information
                 while the last column contains label information. First line
                 of the file should contain header information. Label
                 convention: -1 -> control group - 1 -> pathological group
                 that will be partioned to subgroups
 [--outputDir, -o] : directory where the output from all folds will be saved (REQUIRED)

 OPTIONAL

 [--covCSV, -z] : .csv file containing values for different covariates, which
                 will be used to correct the data accordingly (OPTIONAL). Every
                 column of the file contains values for a covariate, with the
                 exception of the first column, which contains subject
                 identifying information. Correction is performed by solving a
                 solving a least square problem to estimate the respective
                 coefficients and then removing their effect from the data. The
                 effect of ALL provided covariates is removed. If no file is
                 specified, no correction is performed.

 NOTE: featureCSV and covCSV files are assumed to have the subjects given
       in the same order in their rows

 [--c, -c] : regularization parameter (positive scalar). smaller values produce
     sparser models (OPTIONAL - Default 0.25)
 [--reg_type, -r] : determines regularization type. 1 -> promotes sparsity in the
             estimated hyperplanes - 2 -> L2 norm (OPTIONAL - Default 1)
 [--balance, -b] : takes into account differences in the number between the two
             classes. 1-> in case there is mismatch between the number of
             controls and patient - 0-> otherwise (OPTIONAL - Default 1)
 [--init, -g] : initialization strategy. 0 : assignment by random hyperplanes
         (not supported for regression), 1 : pure random assignment, 2:
         k-means assignment, 3: assignment by DPP random
         hyperplanes (default)
 [--iter, -t] : number of iterations between estimating hyperplanes, and cluster
         estimation. Default is 50. Increase if algorithms fails to
         converge
 [--numconsensus, -n] : number of clustering consensus steps. Default is 20.
                 Increase if algorithm gives unstable clustering results.
 [--kmin, -m] : determines the range of clustering solutions to evaluate
             (i.e., kmin to kmax). Default  value is 1.
 [--kmax, -k] : determines the range of clustering solutions to evaluate
             (i.e., kmin to kmax). Default  value is 10.
 [--kstep, -s] : determines the range of clustering solutions to evaluate
             (i.e., kmin to kmax, with step kstep). Default  value is 1.
 [--cvfold, -f]: number of folds for cross validation. Default value is 10.
 [--vo, -j] : verbose output (i.e., also saves input data to verify that all were
     read correctly. Default value is 0
 [--usage, -u]  Prints basic usage message.
 [--help, -h]  Prints help information.
 [--version, -v]  Prints information about software version.

 OUTPUT:
 CIDX: sub-clustering assignments of the disease population (positive
       class).
 ARI: adjusted rand index measuring the overlap/reproducibility of
     clustering solutions across folds

 NOTE: to compile this function do
 mcc -m  hydra.m


 EXAMPLE USE (in matlab)
 hydra('-i','test.csv','-o','.','-k',3,'-f',3);
 EXAMPLE USE (in command line)
 hydra -i test.csv -o . -k 3 -f 3
================================================================
Contact: software@cbica.upenn.edu

Copyright (c) 2018 University of Pennsylvania. All rights reserved.
See COPYING file or http://www.med.upenn.edu/sbia/software/license.html
================================================================
```

Figure B.1: Command line interface of HYDRA

```
>> midas
 function returns MIDAS statistical maps and associated p-values computed
 by analytically estimating permutation testing

 INPUT

 REQUIRED
 [--input, -i] : .csv file containing full paths to input images. (REQUIRED)
               We assume that the first column contains subject identifying
               information; the second column contains the path to the
               images, while the last column contains label information.
               First line of the file should contain header information.

 [--outputDir, -o] : directory where the output from all folds will be saved (REQUIRED)

 OPTIONAL


 [--c, -c] : regularization parameter (positive scalar) (default C=0.1)
 [--radius, -r] : neighborhood radius in voxels (positive scalar) (default R=15)
 [--num, -p] : number of neighborhoods (positive scalar) (default P = 200 )
 [--usage, -u]  Prints basic usage message.
 [--help, -h]  Prints help information.
 [--version, -v]  Prints information about software version.


 OUTPUT:
 map = structure that stores MIDAS statistics and p-values for every group/covariate
 provided with the following structure
 map.stat = cell that stores MIDAS statistic for groups/covariate in same
 order (output as .nii.gz file as well)
 map.p = cell that stores MIDAS statistic for groups/covariate in same
 order (output as .nii.gz file as well)
 map.N = neighborhood coverage amounts (output as .nii.gz file as well)


 NOTE: to compile this function do
 mcc -m  midas -A [NIFTI toolbox directory]
 EXAMPLE USE (in matlab)
 midas('-i','test.csv','-o','.','-r',15,'-p',200,'-c',0.1)
 EXAMPLE USE (in command line)
 midas -i test.csv -o . -r 15 -p 200 -c 0.1
 =================================================================
Contact: software@cbica.upenn.edu

Copyright (c) 2018 University of Pennsylvania. All rights reserved.
See COPYING file or http://www.med.upenn.edu/sbia/software/license.html
 =================================================================
```

Figure B.2: Command line interface of MIDAS

```
>> run_gdm_experiment_nii
 function returns GDM statistical maps and associated p-values computed
 by analytically estimating permutation testing

 INPUT

 REQUIRED
[--input, -i] : .csv file containing full paths to input images. (REQUIRED)
               We assume that the first column contains subject identifying
               information; the second column contains the path to the
               images, while the last column contains label information.
               First line of the file should contain header information.

[--outputDir, -o] : directory where the output from all folds will be saved (REQUIRED)

 OPTIONAL


[--a, -a] : discriminative regularization parameter (positive scalar) (default a=0.1)
[--b, -b] : generative regularization parameter (positive scalar) (default b=0.1)
[--usage, -u]  Prints basic usage message.
[--help, -h]  Prints help information.
[--version, -v]  Prints information about software version.


 OUTPUT:
 map = structure that stores GDM statistics and p-values for every group/covariate
 provided with the following structure
 map.stat = cell that stores GDM statistic for groups/covariate in same
 order (output as .nii.gz file as well)
 map.p = cell that stores GDM statistic for groups/covariate in same
 order (output as .nii.gz file as well)


 NOTE: to compile this function do
 mcc -m  run_gdm_experiment_nii -A [NIFTI toolbox directory]
 EXAMPLE USE (in matlab)
 run_gdm_experiment_nii('-i','test.csv','-o','.','-a',1,'-b',1)
 EXAMPLE USE (in command line)
 run_gdm_experiment_nii -i test.csv -o . -a 1 -b 1
```

Figure B.3: Command line interface of GDM

# Appendix C

# List of genetics features for heterogeneity of Alzheimer's Disease

The SNPs used as features is given in table C.1. Two features were extracted from each subject for each SNP: the presence of the major-major and the major-minor alleles. Minor allele frequency (MAF) column in table C.1 denotes the likelihood of observing the rare minor allele in the population.

| Genetic features used for Control vs. AD Classification/Clustering using HYDRA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNPs associated with cognitive decline identified in [135]. | | | | | | | | | |
| [a]SNP | [b]Chr. | [c]Position | [d]Gene | [e]MAF | [a]SNP | [b]Chr. | [c]Position | [d]Gene | [e]MAF |
| rs2421847 | 1 | 171557600 | PRRC2C | 0.04 | rs4836694 | 9 | 132939792 | NCS1 | 0.11 |
| rs12091371 | 1 | 240605052 | FMN2 | 0.07 | rs118048115 | 10 | 122279476 | PPAPDC1A | 0.04 |
| rs6738962 | 2 | 80281173 | CTNNA2 | 0.04 | rs11023139 | 11 | 14224346 | SPON1 | 0.05 |
| rs78022502 | 2 | 128396167 | LIMS2 | 0.06 | rs61883963 | 11 | 14338703 | RRAS2 | 0.06 |
| rs538867 | 3 | 39513278 | MOBP | 0.03 | rs34162548 | 11 | 14556220 | PSMA1 | 0.05 |
| rs9857727 | 3 | 51095028 | DOCK3 | 0.1 | rs326946 | 11 | 110499253 | ARHGAP20 | 0.17 |
| rs2668205 | 3 | 165493136 | BCHE | 0.03 | rs147845115 | 12 | 51878760 | SLC4A8 | 0.03 |
| rs78647349 | 4 | 5237153 | STK32B | 0.04 | rs61144803 | 12 | 94235165 | CRADD | 0.04 |
| rs340635 | 4 | 87931404 | AFF1 | 0.03 | rs1399439 | 12 | 101221239 | ANO4 | 0.04 |
| rs113689198 | 5 | 109111327 | MAN2A1 | 0.03 | rs143258881 | 13 | 93945858 | GPC6 | 0.03 |
| rs112724034 | 5 | 109221026 | PGAM5P1 | 0.03 | rs17393344 | 13 | 109473946 | MYO16 | 0.06 |
| rs77636885 | 5 | 110719187 | CAMK4 | 0.03 | rs115102486 | 14 | 95764564 | CLMN | 0.03 |
| rs116348108 | 5 | 118435127 | DMXL1 | 0.04 | rs74006954 | 15 | 27712644 | GABRG3 | 0.03 |
| rs143954261 | 5 | 126729450 | MEGF10 | 0.04 | rs17301739 | 15 | 58730639 | LIPC | 0.07 |
| rs146579248 | 5 | 127382302 | FLJ33630 | 0.04 | rs8045064 | 16 | 24675589 | FLJ45256 | 0.05 |
| rs148763909 | 5 | 153837106 | SAP30L | 0.03 | rs9934540 | 16 | 77876763 | VAT1L | 0.03 |
| rs117780815 | 6 | 124326227 | NKAIN2 | 0.03 | rs62076103 | 17 | 45888374 | OSBPL7 | 0.07 |
| rs9494429 | 6 | 136288895 | PDE7B | 0.03 | rs62076130 | 17 | 45905622 | MRPL10 | 0.06 |
| rs75253868 | 6 | 151102830 | PLEKHG1 | 0.04 | rs4794202 | 17 | 45930539 | SP6 | 0.08 |
| rs58370486 | 7 | 16707861 | BZW2 | 0.03 | rs117964204 | 17 | 48692082 | CACNA1G | 0.04 |
| rs73071801 | 7 | 16811139 | TSPAN13 | 0.04 | rs72832584 | 17 | 59292436 | BCAS3 | 0.05 |
| rs1861525 | 7 | 25161602 | CYCS | 0.03 | rs7245858 | 19 | 51430596 | LOC390956 | 0.04 |
| rs17172199 | 7 | 43377276 | HECW1 | 0.08 | rs34972666 | 20 | 2384972 | TGM6 | 0.11 |
| rs73660619 | 8 | 3088173 | CSMD1 | 0.06 | rs75617873 | 22 | 44526105 | PARVB | 0.03 |
| SNPs associated with AD identified in [93] | | | | | | | | | |
| [a]SNP | [b]Chr. | [f]Position | [d]Gene | MAF | [a]SNP | [b]Chr. | [f]Position | [d]Gene | [e]MAF |
| rs6656401 | 1 | 207692049 | CR1 | 0.197 | rs11218343 | 11 | 121435587 | SORL1 | 0.039 |
| rs35349669 | 2 | 234068476 | INPP5D | 0.488 | rs983392 | 11 | 59923508 | MS4A6A | 0.403 |
| rs6733839 | 2 | 127892810 | BIN1 | 0.409 | rs10498633 | 14 | 92926952 | SLC24A4 - RIN3 | 0.217 |
| rs10948363 | 6 | 47487762 | CD2AP | 0.266 | rs17125944 | 14 | 53400629 | FERMT2 | 0.092 |
| rs11771145 | 7 | 143110762 | EPHA1 | 0.338 | rs3865444 | 19 | 51727962 | CD33 | 0.307 |
| rs28834970 | 8 | 27195121 | PTK2B | 0.366 | rs4147929 | 19 | 1063443 | ABCA7 | 0.19 |
| rs9331896 | 8 | 27467686 | CLU | 0.379 | rs429358 | 19 | 44908684 | APOE | 0.1492 |
| rs10792832 | 11 | 85867875 | PICALM | 0.358 | rs7412 | 19 | 44908822 | APOE | 0.07392 |
| rs10838725 | 11 | 47557871 | CELF1 | 0.316 | rs7274581 | 20 | 55018260 | CASS4 | 0.083 |

Table C.1: Genetic features used in HYDRA to classify AD from Controls and discover subtypes of AD. Abbreviations: [a]SNP — Single nucleotide polymorphism [b]Chr. — Chromosome, [c]Position — indicates base pair location in release 19, build 135 of the human genome in the dbSNP database, [d]Gene — Genes located ±100 kb of the top SNP, [e]MAF — minor allele frequency. [f]Position — indicates base pair location in release 19, build 37 of the human genome in the dbSNP database.

# Appendix D

# Related Published Work

The frameworks presented in this thesis and their applications have materialized in publications over the several years. Below is a chronological listing of the peer-reviewed journals and conference proceedings where the methods presented in this thesis have appeared:

## Journal Articles

1. Dong, Aoyan, Jon B. Toledo, Nicolas Honnorat, Jimit Doshi, Erdem Varol, Aristeidis Sotiras, David Wolk, John Q. Trojanowski, Christos Davatzikos, and Alzheimers Disease Neuroimaging Initiative. "Heterogeneity of neuroanatomical patterns in prodromal Alzheimers disease: links to cognition, progression and biomarkers." Brain (2016). [35]

2. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework." NeuroImage 145 (2017): 346-364. [152]

3. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "MIDAS: regionally linear multivariate discriminative statistical mapping." NeuroImage 174 (2018): 111-126. [151]

4. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Generative discriminative machines for multivariate inference and statistical mapping in medical imaging" NeuroImage, (In Preparation)

## Conference Articles

1. Varol, Erdem, and Christos Davatzikos. "Supervised block sparse dictionary learning for simultaneous clustering and classification in computational anatomy." MICCAI, 2014. [148]

2. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Disentangling disease heterogeneity with max-margin multiple hyperplane classifier." MICCAI, 2015.[149]

3. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Structured Outlier Detection in Neuroimaging Studies with Minimal Convex Polytopes." MICCAI, 2015.[150]

4. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Brain mapping through regional multivariate pattern analysis and discriminative adaptive smoothing." OHBM (2017)

5. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Regionally discriminative multivariate statistical mapping." ISBI 2018

6. Varol, Erdem, Aristeidis Sotiras, and Christos Davatzikos. "Generative discriminative models for multivariate inference and statistical mapping in medical imaging" MICCAI 2018

# Bibliography

[1] ALLEFELD, C., AND HAYNES, J. D. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. NeuroImage 89 (2014), 345–357.

[2] ASHBURNER, J. Computational anatomy with the SPM software. Magnetic resonance imaging 27, 8 (Oct. 2009), 1163–74.

[3] ASHBURNER, J., AND FRISTON, K. J. Voxel-based morphometry-the methods. Neuroimage 11, 6 (2000), 805–821.

[4] ASHBURNER, J., AND FRISTON, K. J. Why voxel-based morphometry should be used. Neuroimage 14, 6 (2001), 1238–1243.

[5] ASHBURNER, J., HUTTON, C., FRACKOWIAK, R., JOHNSRUDE, I., PRICE, C., AND FRISTON, K. Identifying global anatomical differences: deformation-based morphometry. Human Brain Mapping 6, 5-6 (Jan. 1998), 348–57.

[6] ASHBURNER, J., AND KLÖPPEL, S. Multivariate models of inter-subject anatomical variability. Neuroimage 56, 2 (2011), 422–439.

[7] BACH, F. R., LANCKRIET, G. R., AND JORDAN, M. I. Multiple Kernel Learning and the SMO Algorithm. In International Conference on Machine Learning (2004), p. 6.

[8] BATES, D. Quadratic forms of random variables: Stat 849 lectures, 2010.

[9] BATMANGHELICH, N. K., TASKAR, B., AND DAVATZIKOS, C. Generative-discriminative basis learning for medical imaging. IEEE transactions on medical imaging 31, 1 (2012), 51–69.

[10] BEN-HUR, A., ELISSEEFF, A., AND GUYON, I. A stability based method for discovering structure in clustered data. In Pacific Symposium on Biocomputing. (2002), vol. 17, pp. 6–17.

[11] BENJAMINI, Y., AND HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. of the royal stat. society. (1995).

[12] BERNAL-RUSIEL, J. L., REUTER, M., GREVE, D. N., FISCHL, B., SABUNCU, M. R., INITIATIVE, A. D. N., ET AL. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81 (2013), 358–370.

[13] BERNASCONI, N., DUCHESNE, S., JANKE, A., LERCH, J., COLLINS, D. L., AND BERNASCONI, A. Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy. NeuroImage 23, 2 (Oct. 2004), 717–23.

[14] BIFFI, A., SHULMAN, J., JAGIELLA, J., CORTELLINI, L., AYRES, A., SCHWAB, K., BROWN, D., SILLIMAN, S., SELIM, M., WORRALL, B., ET AL. Genetic variation at cr1 increases risk of cerebral amyloid angiopathy. Neurology 78, 5 (2012), 334–341.

[15] BJÖRNSDOTTER, M., RYLANDER, K., AND WESSBERG, J. A monte carlo method for locally multivariate brain mapping. NeuroImage 56, 2 (2011), 508–516.

[16] BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition 30, 7 (1997), 1145–1159.

[17] BRALTEN, J., FRANKE, B., ARIAS-VÁSQUEZ, A., HEISTER, A., BRUNNER, H. G., FERNÁNDEZ, G., AND RIJPKEMA, M. Cr1 genotype is associated with entorhinal cortex volume in young healthy adults. Neurobiology of aging 32, 11 (2011), 2106–e7.

[18] BUCHANAN, R. W., AND CARPENTER, W. T. Domains of psychopathology: an approach to the reduction of heterogeneity in schizophrenia. The Journal of nervous and mental disease 182, 4 (1994), 193–204.

[19] CASANOVA, R., SRIKANTH, R., BAER, A., LAURIENTI, P. J., BURDETTE, J. H., HAYASAKA, S., FLOWERS, L., WOOD, F., AND MALDJIAN, J. A. Biological parametric mapping: a statistical toolbox for multimodality brain image analysis. Neuroimage 34, 1 (2007), 137–143.

[20] CASELLA, G., AND BERGER, R. L. Statistical inference, vol. 2. Duxbury Pacific Grove, CA, 2002.

[21] CHANG, C.-C., AND LIN, C.-J. Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 3 (2011), 27.

[22] CHAPUIS, J., HANSMANNEL, F., GISTELINCK, M., MOUNIER, A., VAN CAUWEN-BERGHE, C., KOLEN, K., GELLER, F., SOTTEJEAU, Y., HAROLD, D., DOURLEN, P., ET AL. Increased expression of bin1 mediates alzheimer genetic risk by modulating tau pathology. Molecular psychiatry 18, 11 (2013), 1225–1234.

[23] CHIANG, M. C., REISS, A. L., LEE, A. D., BELLUGI, U., GALABURDA, A. M., KO-
RENBERG, J. R., MILLS, D. L., TOGA, A. W., AND THOMPSON, P. M. 3D pattern of
brain abnormalities in Williams syndrome visualized using tensor-based morphom-
etry. NeuroImage 36, 4 (2007), 1096–1109.

[24] CHUNG, M., WORSLEY, K., PAUS, T., CHERIF, C., COLLINS, D., GIEDD, J.,
RAPOPORT, J., AND EVANS, A. A unified statistical approach to deformation-based
morphometry. NeuroImage 14, 3 (2001), 595–606.

[25] CHUNG, M. K., WORSLEY, K. J., PAUS, T., CHERIF, C., COLLINS, D. L., GIEDD,
J. N., RAPOPORT, J. L., AND EVANS, A. C. A unified statistical approach to
deformation-based morphometry. NeuroImage 14, 3 (2001), 595–606.

[26] CHUNG, M. K., WORSLEY, K. J., ROBBINS, S., PAUS, T., TAYLOR, J., GIEDD, J. N.,
RAPOPORT, J. L., AND EVANS, A. C. Deformation-based surface morphometry ap-
plied to gray matter deformation. NeuroImage 18, 2 (2003), 198–213.

[27] COTTER, A., KESHET, J., AND SREBRO, N. Explicit approximations of the gaussian
kernel. arXiv preprint arXiv:1109.4603 (2011).

[28] CUINGNET, R., GERARDIN, E., TESSIERAS, J., AUZIAS, G., LEHÉRICY, S., HABERT,
M.-O., CHUPIN, M., BENALI, H., COLLIOT, O., INITIATIVE, A. D. N., ET AL. Auto-
matic classification of patients with alzheimer's disease from structural mri: a com-
parison of ten methods using the adni database. neuroimage 56, 2 (2011), 766–781.

[29] CUINGNET, R., GLAUNÈS, J. A., CHUPIN, M., BENALI, H., AND COLLIOT, O. Spa-
tial and anatomical regularization of svm: a general framework for neuroimaging

data. IEEE transactions on pattern analysis and machine intelligence 35, 3 (2013), 682–696.

[30] CUINGNET, R., ROSSO, C., CHUPIN, M., LEHÉRICY, S., DORMONT, D., BENALI, H., SAMSON, Y., AND COLLIOT, O. Spatial regularization of svm for the detection of diffusion alterations associated with stroke outcome. Medical image analysis 15, 5 (2011), 729–737.

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

[32] DAVATZIKOS, C., FAN, Y., WU, X., SHEN, D., AND RESNICK, S. M. Detection of pro-dromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiology of Aging 29, 4 (2008), 514–523.

[33] DAVATZIKOS, C., GENC, A., XU, D., AND RESNICK, S. M. Voxel-based morphome-try using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14, 6 (2001), 1361–1369.

[34] DAVATZIKOS, C., RUPAREL, K., FAN, Y., SHEN, D., ACHARYYA, M., LOUGHEAD, J., GUR, R., AND LANGLEBEN, D. D. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. Neuroimage 28, 3 (2005), 663–668.

[35] DONG, A., TOLEDO, J. B., HONNORAT, N., DOSHI, J., VAROL, E., SOTIRAS, A., WOLK, D., TROJANOWSKI, J. Q., DAVATZIKOS, C., AND INITIATIVE, A. D. N. Het-

erogeneity of neuroanatomical patterns in prodromal alzheimer?s disease: links to cognition, progression and biomarkers. Brain 140, 3 (2016), 735–747.

[36] DOSHI, J., ERUS, G., OU, Y., GAONKAR, B., AND DAVATZIKOS, C. Multi-atlas skull-stripping. Academic radiology 20, 12 (2013), 1566–1576.

[37] DOSHI, J., ERUS, G., OU, Y., RESNICK, S. M., GUR, R. C., GUR, R. E., SATTERTH-WAITE, T. D., FURTH, S., AND DAVATZIKOS, C. Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. NeuroImage (2015).

[38] DRUCKER, H., BURGES, C. J., KAUFMAN, L., SMOLA, A. J., AND VAPNIK, V. Support vector regression machines. In Advances in neural information processing systems (1997), pp. 155–161.

[39] DUCHESNE, S., CAROLI, A., GEROLDI, C., BARILLOT, C., FRISONI, G. B., AND COLLINS, D. L. MRI-based automated computer classification of probable AD versus normal controls. IEEE transactions on medical imaging 27, 4 (Apr. 2008), 509–20.

[40] DUKART, J., SCHROETER, M. L., MUELLER, K., INITIATIVE, A. D. N., ET AL. Age correction in dementia–matching to a healthy brain. PloS one 6, 7 (2011), e22193.

[41] DUSTIN, M. L., OLSZOWY, M. W., HOLDORF, A. D., LI, J., BROMLEY, S., DESAI, N., WIDDER, P., ROSENBERGER, F., VAN DER MERWE, P., ALLEN, P. M., AND SHAW, A. S. A Novel Adaptor Protein Orchestrates Receptor Patterning and Cytoskeletal Polarity in T-Cell Contacts. Cell 94, 5 (Sept. 1998), 667–677.

[42] Ecker, C., Marquand, A., Mourão Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., Maltezos, S., Murphy, C. M., Robertson, D., Williams, S. C., and Murphy, D. G. M. Describing the brain in autism in five dimensions–magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. The Journal of Neuroscience 30, 32 (2010), 10612–10623.

[43] Efron, B., and Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science 1, 1 (1986), 54–75.

[44] Etzel, J. A., Zacks, J. M., and Braver, T. S. Searchlight analysis: promise, pitfalls, and potential. Neuroimage 78 (2013), 261–269.

[45] Fennema-Notestine, C., Panizzon, M. S., Thompson, W. R., Chen, C.-H., Eyler, L. T., Fischl, B., Franz, C. E., Grant, M. D., Jak, A. J., Jernigan, T. L., et al. Presence of apoe $\varepsilon$4 allele associated with thinner frontal cortex in middle age. Journal of Alzheimer's Disease 26, Suppl 3 (2011), 49.

[46] Fenton, W. S., McGlashan, T. H., Victor, B. J., and Blyler, C. R. Symptoms, subtype, and suicidality in patients with schizophrenia spectrum disorders. American journal of psychiatry 154, 2 (1997), 199–204.

[47] Filipovych, R., Resnick, S. M., and Davatzikos, C. Jointmmcc: Joint maximum-margin classification and clustering of imaging data. Medical Imaging, IEEE Transactions on 31, 5 (2012), 1124–1140.

[48] FLEISS, J. L., LEVIN, B., AND PAIK, M. C. Statistical methods for rates and proportions. John Wiley & Sons, 2013.

[49] FOLSTEIN, M. F., FOLSTEIN, S. E., AND MCHUGH, P. R. "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of psychiatric research 12, 3 (1975), 189–198.

[50] FOX, N. C., COUSENS, S., SCAHILL, R., HARVEY, R. J., AND ROSSOR, M. N. Using serial registered brain magnetic resonance imaging to measure disease progression in alzheimer disease: power calculations and estimates of sample size to detect treatment effects. Archives of Neurology 57, 3 (2000), 339–344.

[51] FOX, N. C., CRUM, W. R., SCAHILL, R. I., STEVENS, J. M., JENSSEN, J. C., AND ROSSOR, M. N. Imaging of onset and progression of Alzheimer's disease with voxel-compression mapping of serial magnetic resonance images. The Lancet 358 (2001), 201–5.

[52] FU, Z., ROBLES-KELLY, A., AND ZHOU, J. Mixing linear svms for nonlinear classification. Neural Networks, IEEE Transactions on 21, 12 (2010), 1963–1975.

[53] GANZ, M., GREVE, D. N., FISCHL, B., KONUKOGLU, E., INITIATIVE, A. D. N., ET AL. Relevant feature set estimation with a knock-out strategy and random forests. Neuroimage 122 (2015), 131–148.

[54] GAONKAR, B., AND DAVATZIKOS, C. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 78 (2013), 270–283.

[55] GAONKAR, B., AND OTHER. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Medical image analysis 24, 1 (2015), 190–204.

[56] GASER, C., VOLZ, H.-P., KIEBEL, S., RIEHEMANN, S., AND SAUER, H. Detecting structural changes in whole brain based on nonlinear deformations?application to schizophrenia research. Neuroimage 10, 2 (1999), 107–113.

[57] GESCHWIND, D. H., AND LEVITT, P. Autism spectrum disorders: developmental disconnection syndromes. Current opinion in neurobiology 17, 1 (Feb. 2007), 103–11.

[58] GIEDD, J. N., BLUMENTHAL, J., JEFFRIES, N. O., CASTELLANOS, F. X., LIU, H., ZIJDENBOS, A., PAUS, T., EVANS, A. C., AND RAPOPORT, J. L. Brain development during childhood and adolescence: a longitudinal mri study. Nature neuroscience 2, 10 (1999), 861–863.

[59] GIULIANI, N. R., CALHOUN, V. D., PEARLSON, G. D., FRANCIS, A., AND BUCHANAN, R. W. Voxel-based morphometry versus region of interest: A comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophrenia Research 74, 2-3 (2005), 135–147.

[60] GOLDSZAL, A. F., DAVATZIKOS, C., PHAM, D. L., YAN, M. X., BRYAN, R. N., AND RESNICK, S. M. An image-processing system for qualitative and quantitative volumetric analysis of brain images. Journal of Computer Assisted Tomography 22, 5 (1998), 827—-837.

[61] GÖNEN, M., AND ALPAYDIN, E. Multiple kernel learning algorithms. Journal of machine learning research 12, Jul (2011), 2211–2268.

[62] GOOD, C. D., JOHNSRUDE, I. S., ASHBURNER, J., HENSON, R. N., FRISTEN, K., AND FRACKOWIAK, R. S. A voxel-based morphometric study of ageing in 465 normal adult human brains. In Biomedical Imaging, 2002. 5th IEEE EMBS International Summer School on (2002), IEEE, pp. 16–pp.

[63] GRAHAM, J. M., AND SAGAR, H. J. A data-driven approach to the study of heterogeneity in idiopathic parkinson's disease: identification of three distinct subtypes. Movement Disorders 14, 1 (1999), 10–20.

[64] GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B., AND TAYLOR, J. E. Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72 (2013), 304–321.

[65] GU, Q., AND HAN, J. Clustered support vector machines. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (2013), pp. 307–315.

[66] GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. Gene selection for cancer classification using support vector machines. Machine learning 46, 1-3 (2002), 389–422.

[67] HASHIMOTO, M., YASUDA, M., TANIMUKAI, S., MATSUI, M., HIRONO, N., KAZUI, H., AND MORI, E. Apolipoprotein e $\varepsilon 4$ and the pattern of regional brain atrophy in alzheimer's disease. Neurology 57, 8 (2001), 1461–1466.

[68] HAUFE, S., ET AL. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87 (2014), 96–110.

[69] HINRICHS, C., SINGH, V., XU, G., AND JOHNSON, S. C. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage 55, 2 (Mar. 2011), 574–89.

[70] HOERL, A. E., AND KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 1 (1970), 55–67.

[71] HONEA, R. A., VIDONI, E., HARSHA, A., AND BURNS, J. M. Impact of apoe on the healthy aging brain: a voxel-based mri and dti study. Journal of Alzheimer's disease: JAD 18, 3 (2009), 553.

[72] HUA, X., LEOW, A. D., PARIKSHAK, N., LEE, S., CHIANG, M. C., TOGA, A. W., JACK, C. R., WEINER, M. W., AND THOMPSON, P. M. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: An MRI study of 676 AD, MCI, and normal subjects. NeuroImage 43, 3 (2008), 458–469.

[73] HUANG, C., WAHLUND, L. O., ALMKVIST, O., ELEHU, D., SVENSSON, L., JONSSON, T., WINBLAD, B., AND JULIN, P. Voxel- and VOI-based analysis of SPECT CBF in relation to clinical and psychological heterogeneity of mild cognitive impairment. NeuroImage 19, 3 (2003), 1137–1144.

[74] HUBERT, L., AND ARABIE, P. Comparing partitions. Journal of classification 2, 1 (1985), 193–218.

[75] JESTE, S. S., AND GESCHWIND, D. H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nature Reviews Neurology 10, 2 (2014), 74.

[76] JESTE, S. S., AND GESCHWIND, D. H. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nature reviews. Neurology 10, 2 (Feb. 2014), 74–81.

[77] JOB, D. E., WHALLEY, H. C., JOHNSTONE, E. C., AND LAWRIE, S. M. Grey matter changes over time in high risk subjects developing schizophrenia. NeuroImage 25, 4 (2005), 1023–1030.

[78] JOB, D. E., WHALLEY, H. C., MCCONNELL, S., GLABUS, M., JOHNSTONE, E. C., AND LAWRIE, S. M. Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry. NeuroImage 17, 2 (2002), 880–889.

[79] JONES, D. K., SYMMS, M. R., CERCIGNANI, M., AND HOWARD, R. J. The effect of filter size on VBM analyses of DT-MRI data. NeuroImage 26, 2 (2005), 546–554.

[80] JUOTTONEN, K., LEHTOVIRTA, M., HELISALMI, S., RIEKKINEN SR, P., AND SOINI-NEN, H. Major decrease in the volume of the entorhinal cortex in patients with alzheimer's disease carrying the apolipoprotein e $\varepsilon$4 allele. Journal of Neurology, Neurosurgery & Psychiatry 65, 3 (1998), 322–327.

[81] KABANI, N. J., MACDONALD, D. J., HOLMES, C. J., AND EVANS, A. C. 3D anatomical atlas of the human brain. NeuroImage 7, 4 PART II (1998), S717.

[82] KAK, A. C., AND ROSENFELD, A. Digital picture processing. New York (1982).

[83] KANTCHELIAN, A., TSCHANTZ, M. C., HUANG, L., BARTLETT, P. L., JOSEPH, A. D., AND TYGAR, J. Large-margin convex polytope machine. In Advances in Neural Information Processing Systems (2014), pp. 3248–3256.

[84] KLÖPPEL, S., STONNINGTON, C. M., CHU, C., DRAGANSKI, B., SCAHILL, R. I., ROHRER, J. D., FOX, N. C., JACK, C. R., ASHBURNER, J., AND FRACKOWIAK, R. S. J. Automatic classification of MR scans in Alzheimer's disease. Brain 131, 3 (2008), 681–689.

[85] KLÖPPEL, S., STONNINGTON, C. M., CHU, C., DRAGANSKI, B., SCAHILL, R. I., ROHRER, J. D., FOX, N. C., JACK JR, C. R., ASHBURNER, J., AND FRACKOWIAK, R. S. Automatic classification of mr scans in alzheimer's disease. Brain 131, 3 (2008), 681–689.

[86] KOUTSOULERIS, N., GASER, C., JÄGER, M., BOTTLENDER, R., FRODL, T., HOLZINGER, S., SCHMITT, G. J., ZETZSCHE, T., BURGERMEISTER, B., SCHEUERECKER, J., ET AL. Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study. Neuroimage 39, 4 (2008), 1600–1612.

[87] KOUTSOULERIS, N., GASER, C., JÄGER, M., BOTTLENDER, R., FRODL, T., HOLZINGER, S., SCHMITT, G. J. E., ZETZSCHE, T., BURGERMEISTER, B., SCHEUERECKER, J., BORN, C., REISER, M., MÖLLER, H.-J., AND MEISEN-ZAHL, E. M. Structural correlates of psychopathological symptom dimensions in

schizophrenia: a voxel-based morphometric study. NeuroImage 39, 4 (Feb. 2008), 1600–12.

[88] KOUTSOULERIS, N., MEISENZAHL, E. M., DAVATZIKOS, C., BOTTLENDER, R., FRODL, T., SCHEUERECKER, J., SCHMITT, G., ZETZSCHE, T., DECKER, P., REISER, M., 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 (2009), 700–712.

[89] KRIEGESKORTE, N., GOEBEL, R., AND BANDETTINI, P. Information-based functional brain mapping. PNAS 103, 10 (2006), 3863–3868.

[90] KUBICKI, M., SHENTON, M. E., SALISBURY, D. F., HIRAYASU, Y., KASAI, K., KIKINIS, R., JOLESZ, F. A., AND MCCARLEY, R. W. Voxel-based morphometric analysis of gray matter in first episode schizophrenia. NeuroImage 17, 4 (2002), 1711–1719.

[91] KULESZA, A., AND TASKAR, B. Determinantal point processes for machine learning. arXiv preprint arXiv:1207.6083 (2012).

[92] LAM, B., MASELLIS, M., FREEDMAN, M., STUSS, D. T., AND BLACK, S. E. Clinical, imaging, and pathological heterogeneity of the alzheimer's disease syndrome. Alzheimer's research & therapy 5, 1 (2013), 1.

[93] LAMBERT, J.-C., IBRAHIM-VERBAAS, C. A., HAROLD, D., NAJ, A. C., SIMS, R., BELLENGUEZ, C., JUN, G., DESTEFANO, A. L., BIS, J. C., BEECHAM, G. W., ET AL. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics 45, 12 (2013), 1452–1458.

[94] LANGE, T., ROTH, V., BRAUN, M. L., AND BUHMANN, J. M. Stability-based validation of clustering solutions. Neural computation 16, 6 (2004), 1299–1323.

[95] LANGLEBEN, D. D., LOUGHEAD, J. W., BILKER, W. B., RUPAREL, K., CHILDRESS, A. R., BUSCH, S. I., AND GUR, R. C. Telling truth from lie in individual subjects with fast event-related fmri. Human brain mapping 26, 4 (2005), 262–272.

[96] LANGS, G., MENZE, B. H., LASHKARI, D., AND GOLLAND, P. Detecting stable distributed patterns of brain activation using Gini contrast. NeuroImage 56, 2 (may 2011), 497–507.

[97] LEDOIT, O., AND WOLF, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of empirical finance 10, 5 (2003), 603–621.

[98] LEPORE, N., BRUN, C., CHOU, Y.-Y., CHIANG, M.-C., DUTTON, R. A., HAYASHI, K. M., LU, A., LOPEZ, O. L., AIZENSTEIN, H. J., TOGA, A. W., BECKER, J. T., AND THOMPSON, P. M. Generalized Tensor-Based Morphometry of HIV / AIDS Using Multivariate Statistics on Strain Matrices. IEEE Transactions on Medial Imaging 27, 1 (2006), 129–141.

[99] LEWIS, S. J. G., FOLTYNIE, T., BLACKWELL, A. D., ROBBINS, T. W., OWEN, A. M., AND BARKER, R. A. Heterogeneity of Parkinson's disease in the early clinical stages using a data driven approach. Journal of neurology, neurosurgery, and psychiatry 76, 3 (Mar. 2005), 343–8.

[100] LI, C., GORE, J. C., AND DAVATZIKOS, C. Multiplicative intrinsic component optimization (mico) for mri bias field estimation and tissue segmentation. Magnetic resonance imaging 32, 7 (2014), 913–923.

[101] LINN, K. A., GAONKAR, B., DOSHI, J., DAVATZIKOS, C., AND SHINOHARA, R. T. Addressing confounding in predictive models with an application to neuroimaging. The international journal of biostatistics 12, 1 (2016), 31–44.

[102] LLOYD, S. P. Least squares quantization in pcm. Information Theory, IEEE Transactions on 28, 2 (1982), 129–137.

[103] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J., AND TIBSHIRANI, R. A significance test for the lasso. Annals of statistics 42, 2 (2014), 413.

[104] MAIRAL, J., BACH, F., AND PONCE, J. Task-driven dictionary learning. IEEE transactions on pattern analysis and machine intelligence 34, 4 (2012), 791–804.

[105] MCEVOY, L. K., FENNEMA-NOTESTINE, C., RODDEY, J. C., HAGLER, D. J., HOLLAND, D., KAROW, D. S., PUNG, C. J., BREWER, J. B., AND DALE, A. M. Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment. Radiology 251, 1 (Apr. 2009), 195–205.

[106] MCINTOSH, A. R., AND MIŠIĆ, B. Multivariate statistical analyses for neuroimaging data. Annual review of psychology 64 (2013), 499–525.

[107] MEDA, S. A., GIULIANI, N. R., CALHOUN, V. D., JAGANNATHAN, K., SCHRETLEN, D. J., PULVER, A., CASCELLA, N., KESHAVAN, M., KATES, W., BUCHANAN, R.,

SHARMA, T., AND PEARLSON, G. D. A large scale (N = 400) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry. Schizophrenia Research 101, 1-3 (2008), 95–105.

[108] MOHS, R. The alzheimer's disease assessment scale: an instrument for assessing treatment efficacy. Psychopharmacol. Bull. 2 (1983), 448–450.

[109] MOURÃO MIRANDA, J., BOKDE, A. L. W., BORN, C., HAMPEL, H., AND STETTER, M. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage 28, 4 (Dec. 2005), 980–95.

[110] MUMFORD, J. A., POLINE, J.-B., AND POLDRACK, R. A. Orthogonalization of regressors in fmri models. PLoS One 10, 4 (2015), e0126255.

[111] MURRAY, M. E., GRAFF-RADFORD, N. R., ROSS, O. A., PETERSEN, R. C., DUARA, R., AND DICKSON, D. W. Neuropathologically defined subtypes of alzheimer's disease with distinct clinical characteristics: a retrospective study. The Lancet Neurology 10, 9 (2011), 785–796.

[112] MURRAY, M. E., GRAFF-RADFORD, N. R., ROSS, O. a., PETERSEN, R. C., DUARA, R., AND DICKSON, D. W. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. The Lancet. Neurology 10, 9 (Sept. 2011), 785–96.

[113] NAJ, A. C., JUN, G., BEECHAM, G. W., WANG, L.-S., VARDARAJAN, B. N., BUROS, J., GALLINS, P. J., BUXBAUM, J. D., JARVIK, G. P., CRANE, P. K., ET AL. Common

variants at ms4a4/ms4a6e, cd2ap, cd33 and epha1 are associated with late-onset alzheimer's disease. Nature genetics 43, 5 (2011), 436–441.

[114] NENADIC, I., GASER, C., AND SAUER, H. Heterogeneity of brain structural variation and the structural imaging endophenotypes in schizophrenia. Neuropsychobiology 66, 1 (2012), 44–49.

[115] NENADIC, I., SAUER, H., AND GASER, C. Distinct pattern of brain structural deficits in subsyndromes of schizophrenia delineated by psychopathology. NeuroImage 49, 2 (Jan. 2010), 1153–60.

[116] NG, A. Y., JORDAN, M. I., WEISS, Y., ET AL. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems 2 (2002), 849–856.

[117] NICHOLS, T. E., AND HOLMES, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping 15, 1 (2002), 1–25.

[118] NOH, Y., JEON, S., LEE, J. M., SEO, S. W., KIM, G. H., CHO, H., YE, B. S., YOON, C. W., KIM, H. J., CHIN, J., ET AL. Anatomical heterogeneity of alzheimer disease based on cortical thickness on mris. Neurology 83, 21 (2014), 1936–1944.

[119] OLIVETTI, E., MOGNON, A., GREINER, S., AND AVESANI, P. Brain decoding: biases in error estimation. In Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), 2010 First Workshop on (2010), IEEE, pp. 40–43.

[120] OSADCHY, M., HAZAN, T., AND KEREN, D. K-hyperplane hinge-minimax classifier. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)* (2015), pp. 1558–1566.

[121] OU, Y., SOTIRAS, A., PARAGIOS, N., AND DAVATZIKOS, C. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis 15*, 4 (2011), 622–639.

[122] PAYAMI, H., ZAREPARSI, S., MONTEE, K. R., SEXTON, G. J., KAYE, J. A., BIRD, T. D., YU, C. E., WIJSMAN, E. M., HESTON, L. L., LITT, M., AND SCHELLENBERG, G. D. Gender difference in apolipoprotein E-associated risk for familial Alzheimer disease: a possible clue to the higher incidence of Alzheimer disease in women. *American journal of human genetics 58*, 4 (1996), 803–811.

[123] PEREIRA, F., AND BOTVINICK, M. Information mapping with pattern classifiers: a comparative study. *Neuroimage 56*, 2 (2011), 476–496.

[124] PRINCE, J., ZETTERBERG, H., ANDREASEN, N., MARCUSSON, J., AND BLENNOW, K. APOE epsilon4 allele is associated with reduced cerebrospinal fluid levels of Abeta42. *Neurology 62*, 11 (2004), 2116–2118.

[125] RAO, A., MONTEIRO, J. M., MOURAO-MIRANDA, J., INITIATIVE, A. D., ET AL. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage 150* (2017), 23–49.

[126] RASMUSSEN, P. M., HANSEN, L. K., MADSEN, K. H., CHURCHILL, N. W., AND STROTHER, S. C. Model sparsity and brain pattern interpretation of classification

models in neuroimaging. Pattern Recognition 45, 6 (2012), 2085–2100.

[127] RENCHER, A. C., AND SCHAALJE, G. B. Linear models in statistics. John Wiley & Sons, 2008.

[128] RIDDLE, W. R., LI, R., FITZPATRICK, J. M., DONLEVY, S. C., DAWANT, B. M., AND PRICE, R. R. Characterizing changes in mr images with color-coded jacobians. Magnetic resonance imaging 22, 6 (2004), 769–777.

[129] RONDINA, J. M., HAHN, T., DE OLIVEIRA, L., MARQUAND, A. F., DRESLER, T., LEITNER, T., FALLGATTER, A. J., SHAWE-TAYLOR, J., AND MOURAO-MIRANDA, J. Scors?a method based on stability for feature selection and mapping in neuroimaging. IEEE transactions on medical imaging 33, 1 (2014), 85–98.

[130] SABUNCU, M. R., BALCI, S. K., SHENTON, M. E., AND GOLLAND, P. Image-driven population analysis through mixture modeling. Medical Imaging, IEEE Transactions on 28, 9 (2009), 1473–1487.

[131] SABUNCU, M. R., AND VAN LEEMPUT, K. The relevance voxel machine (rvoxm): a self-tuning bayesian model for informative image-based prediction. IEEE transactions on medical imaging 31, 12 (2012), 2290–2306.

[132] SATO, J. R., DA GRAÇA MORAIS MARTIN, M., FUJITA, A., MOURÃO-MIRANDA, J., BRAMMER, M. J., AND AMARO, E. An fmri normative database for connectivity networks using one-class support vector machines. Human brain mapping 30, 4 (2009), 1068–1076.

[133] SAYKIN, A. J., SHEN, L., FOROUD, T. M., POTKIN, S. G., SWAMINATHAN, S., KIM, S., RISACHER, S. L., NHO, K., HUENTELMAN, M. J., CRAIG, D. W., THOMPSON, P. M., STEIN, J. L., MOORE, J. H., FARRER, L. A., GREEN, R. C., BERTRAM, L., JACK, C. R., AND WEINER, M. W. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. Alzheimer's and Dementia 6, 3 (2010), 265–273.

[134] SHEN, D., AND DAVATZIKOS, C. Very high-resolution morphometry using mass-preserving deformations and HAMMER elastic registration. NeuroImage 18, 1 (2003), 28–41.

[135] SHERVA, R., TRIPODIS, Y., BENNETT, D. A., CHIBNIK, L. B., CRANE, P. K., DE JAGER, P. L., FARRER, L. A., SAYKIN, A. J., SHULMAN, J. M., NAJ, A., ET AL. Genome-wide association study of the rate of cognitive decline in alzheimer's disease. Alzheimer's & Dementia 10, 1 (2014), 45–52.

[136] SHIMODAIRA, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference 90, 2 (2000), 227–244.

[137] SLED, J. G., ZIJDENBOS, A. P., AND EVANS, A. C. A nonparametric method for automatic correction of intensity nonuniformity in mri data. Medical Imaging, IEEE Transactions on 17, 1 (1998), 87–97.

[138] STUDHOLME, C., CARDENAS, V., BLUMENFELD, R., SCHUFF, N., ROSEN, H. J., MILLER, B., AND WEINER, M. Deformation tensor morphometry of semantic dementia with quantitative validation. NeuroImage 21, 4 (May 2004), 1387–98.

[139] SUGIYAMA, M., KRAULEDAT, M., AND MÃŽLLER, K.-R. Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research 8, May (2007), 985–1005.

[140] SUNDERLAND, T., MIRZA, N., PUTNAM, K. T., LINKER, G., BHUPALI, D., DURHAM, R., SOARES, H., KIMMEL, L., FRIEDMAN, D., BERGESON, J., CSAKO, G., LEVY, J. A., BARTKO, J. J., AND COHEN, R. M. Cerebrospinal fluid beta-amyloid1-42 and tau in control subjects at risk for Alzheimer's disease: the effect of APOE epsilon4 allele. Biological psychiatry 56, 9 (2004), 670–6.

[141] SUYKENS, J. A., AND VANDEWALLE, J. Least squares support vector machine classifiers. Neural processing letters 9, 3 (1999), 293–300.

[142] TAGER-FLUSBERG, H., AND JOSEPH, R. M. Identifying neurocognitive phenotypes in autism. Philosophical Transactions of the Royal Society of London B: Biological Sciences 358, 1430 (2003), 303–314.

[143] TAKÁCS, G. Convex polyhedron learning and its applications. PhD thesis, Citeseer, 2009.

[144] THOMPSON, P. M., GIEDD, J. N., WOODS, R. P., MACDONALD, D., EVANS, A. C., AND TOGA, A. W. Growth patterns in the developing brain detected by using continuum mechanical tensor maps. Nature 404, 6774 (2000), 190–193.

[145] TREUSCH, S., HAMAMICHI, S., GOODMAN, J. L., MATLACK, K. E. S., CHUNG, C. Y., BARU, V., SHULMAN, J. M., PARRADO, A., BEVIS, B. J., VALASTYAN, J. S., HAN, H., LINDHAGEN-PERSSON, M., REIMAN, E. M., EVANS, D. A., BENNETT,

D. a., Olofsson, A., DeJager, P. L., Tanzi, R. E., Caldwell, K. a., Caldwell, G. a., and Lindquist, S. Functional Links Between A Toxicity, Endocytic Trafficking, and Alzheimer's Disease Risk Factors in Yeast. Science 334, 6060 (Dec. 2011), 1241–1245.

[146] van de Pol, L. A., Korf, E. S., van der Flier, W. M., Brashear, H. R., Fox, N. C., Barkhof, F., and Scheltens, P. Magnetic resonance imaging predictors of cognition in mild cognitive impairment. Archives of neurology 64, 7 (2007), 1023–1028.

[147] Vapnik, V. The nature of statistical learning theory. springer, 2000.

[148] Varol, E., and Davatzikos, C. Supervised block sparse dictionary learning for simultaneous clustering and classification in computational anatomy. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, 2014, pp. 446–453.

[149] Varol, E., Sotiras, A., and Davatzikos, C. Disentangling disease heterogeneity with max-margin multiple hyperplane classifier. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, 2015, pp. 702–709.

[150] Varol, E., Sotiras, A., and Davatzikos, C. Structured outlier detection in neuroimaging studies with minimal convex polytopes. In International Conference on Medical Image Computing and Computer-Assisted Intervention (2016), Springer, pp. 300–307.

[151] VAROL, E., SOTIRAS, A., AND DAVATZIKOS, C. Midas: Regionally linear multivariate discriminative statistical mapping. NeuroImage (2018).

[152] VAROL, E., SOTIRAS, A., DAVATZIKOS, C., INITIATIVE, A. D. N., ET AL. Hydra: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework. NeuroImage 145 (2017), 346–364.

[153] VEMURI, P., GUNTER, J. L., SENJEM, M. L., WHITWELL, J. L., KANTARCI, K., KNOPMAN, D. S., BOEVE, B. F., PETERSEN, R. C., AND JACK, C. R. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39, 3 (Feb. 2008), 1186–97.

[154] VEMURI, P., GUNTER, J. L., SENJEM, M. L., WHITWELL, J. L., KANTARCI, K., KNOPMAN, D. S., BOEVE, B. F., PETERSEN, R. C., AND JACK JR, C. R. Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies. Neuroimage 39, 3 (2008), 1186–1197.

[155] WÅHLSTEDT, C., THORELL, L. B., AND BOHLIN, G. Heterogeneity in adhd: Neuropsychological pathways, comorbidity and symptom domains. Journal of abnormal child psychology 37, 4 (2009), 551–564.

[156] WHITWELL, J. L., DICKSON, D. W., MURRAY, M. E., WEIGAND, S. D., TOSAKULWONG, N., SENJEM, M. L., KNOPMAN, D. S., BOEVE, B. F., PARISI, J. E., PETERSEN, R. C., JACK, C. R., AND JOSEPHS, K. A. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. The Lancet. Neurology 11, 10 (Oct. 2012), 868–77.

[157] WHITWELL, J. L., PETERSEN, R. C., NEGASH, S., WEIGAND, S. D., KANTARCI, K., IVNIK, R. J., KNOPMAN, D. S., BOEVE, B. F., SMITH, G. E., AND JACK, C. R. Patterns of atrophy differ among specific subtypes of mild cognitive impairment. Archives of neurology 64, 8 (Aug. 2007), 1130–8.

[158] WINKLER, A. M., RIDGWAY, G. R., WEBSTER, M. A., SMITH, S. M., AND NICHOLS, T. E. Permutation inference for the general linear model. Neuroimage 92 (2014), 381–397.

[159] WRIGHT, I., MCGUIRE, P., POLINE, J.-B., TRAVERE, J., MURRAY, R., FRITH, C., FRACKOWIAK, R., AND FRISTON, K. A Voxel-Based Method for the Statistical Analysis of Gray and White Matter Density Applied to Schizophrenia. NeuroImage 2, 4 (dec 1995), 244–252.

[160] YAMASUE, H., KASAI, K., IWANAMI, A., OHTANI, T., YAMADA, H., ABE, O., KUROKI, N., FUKUDA, R., TOCHIGI, M., FURUKAWA, S., ET AL. Voxel-based analysis of mri reveals anterior cingulate gray-matter volume reduction in posttraumatic stress disorder due to terrorism. Proceedings of the National Academy of Sciences 100, 15 (2003), 9039–9043.

[161] ZHANG, T., AND DAVATZIKOS, C. Odvba: optimally-discriminative voxel-based analysis. IEEE transactions on medical imaging 30, 8 (2011), 1441–1454.

[162] ZHANG, T., AND DAVATZIKOS, C. Optimally-discriminative voxel-based morphometry significantly increases the ability to detect group differences in schizophrenia, mild cognitive impairment, and alzheimer's disease. Neuroimage 79 (2013), 94–110.

[163] ZHANG, T., KOUTSOULERIS, N., MEISENZAHL, E., AND DAVATZIKOS, C. Heterogeneity of Structural Brain Changes in Subtypes of Schizophrenia Revealed Using Magnetic Resonance Imaging Pattern Analysis. Schizophrenia Bulletin 41, 1 (Jan. 2015), 74–84.

[164] ZHANG, Y., ZOU, P., MULHERN, R. K., BUTLER, R. W., LANINGHAM, F. H., AND OGG, R. J. Brain structural abnormalities in survivors of pediatric posterior fossa brain tumors: A voxel-based morphometry study using free-form deformation. NeuroImage 42, 1 (2008), 218–229.