1-1-2012

Generative-Discriminative Low Rank Decomposition for Medical Imaging Applications

Nematollah Kayhan Batmanghelich
University of Pennsylvania, batmangh@seas.upenn.edu

Generative-Discriminative Low Rank Decomposition for Medical Imaging Applications

Abstract

In this thesis, we propose a method that can be used to extract biomarkers from medical images toward early diagnosis of abnormalities. Surge of demand for biomarkers and availability of medical images in the recent years call for accurate, repeatable, and interpretable approaches for extracting meaningful imaging features. However, extracting such information from medical images is a challenging task because the number of pixels (voxels) in a typical image is in order of millions while even a large sample-size in medical image dataset does not usually exceed a few hundred. Nevertheless, depending on the nature of an abnormality, only a parsimonious subset of voxels is typically relevant to the disease; therefore various notions of sparsity are exploited in this thesis to improve the generalization performance of the prediction task.

We propose a novel discriminative dimensionality reduction method that yields good classification performance on various datasets without compromising the clinical interpretability of the results. This is achieved by combining the modelling strength of generative learning framework and the classification performance of discriminative learning paradigm. Clinical interpretability can be viewed as an additional measure of evaluation and is also helpful in designing methods that account for the clinical prior such as association of certain areas in a brain to a particular cognitive task or connectivity of some brain regions via neural fibres.

We formulate our method as a large-scale optimization problem to solve a constrained matrix factorization. Finding an optimal solution of the large-scale matrix factorization renders off-the-shelf solver computationally prohibitive; therefore, we designed an efficient algorithm based on the proximal method to address the computational bottle-neck of the optimization problem. Our formulation is readily extended for different scenarios such as cases where a large cohort of subjects has uncertain or no class labels (semi-supervised learning) or a case where each subject has a battery of imaging channels (multi-channel), \etc. We show that by using various notions of sparsity as feasible sets of the optimization problem, we can encode different forms of prior knowledge ranging from brain parcellation to brain connectivity.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Electrical & Systems Engineering

First Advisor
Christos Davatzikos

Second Advisor
Ben Taskar

Keywords
Alzhiemer's disease, Biomarker, Generative-Discriminative Learning, Machine Learning, Medical Imagining, Optimization

Subject Categories
Biomedical | Computer Sciences | Electrical and Electronics

GENERATIVE-DISCRIMINATIVE LOW RANK DECOMPOSITION FOR MEDICAL IMAGING APPLICATIONS

Nematollah Kayhan Batmanghelich

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

2012

Christos Davatzikos, Professor, Radiology, UPenn
Supervisor of Dissertation

Ben Taskar, Professor, CIS, UPenn
Co-Supervisor of Dissertation

Dr. Saswati Sarkar
Graduate Group Chairperson

DISSERTATION COMMITTEE

Christos Davatzikos, Professor, Radiology, UPenn
Ben Taskar, Professor, CIS, UPenn
Ali Jadbabaie, Professor, ESE, UPenn
Daniel Lee, Professor, ESE, UPenn
Polina Golland, Professor, CSAIL, MIT

Dedication

To Shohreh for Lighting Up My Life.

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisors, Dr Christos Davatzikos and Dr Ben Taskar. This piece of work could not have come to existence without their unremitting supports. Ben and Christos taught me how to think independently while providing continuous advice in all aspects of research. I will be indefinitely indebted to them due to their encouragements which they supplied throughout all these years.

I am thankful to the other members of my committee: Professors Ali Jadbabaie, Daniel Lee, and Polina Golland. Their valuable suggestions and considerable advice have greatly improved the quality of this thesis. I would also like to thank Professors Ragini Verma and Kilian Pohl for the insightful discussions that I had with them in our lab meeting all these years.

An important role was played by my many friends at Penn and SBIA lab: Alex, Ali, Bilwaj, Ben, Dong, Hamed, Luke, Madhura, and Yasser. Special thanks to Aris and Guray for helping me to prepare this manuscript by proof-reading the chapters. I should thank Mark and Paraskevi with whom I had many nice conversations during my graduate life in SBIA.

I would like to manifest my deepest gratitude and appreciation to my parents, Farzaneh and Faramarz, and sisters Zahra and Fatemeh, for their relentless support and love. My sisters' friendship and my parents' love and support are the most precious possession that I hold and it has been a great source of the strength and inspiration for me.

Last but not least, I would like to thank my wife Shohreh for her unconditional love throughout my educational career. I owe her much more than I would ever be able to express.

ABSTRACT

GENERATIVE-DISCRIMINATIVE LOW RANK DECOMPOSITION FOR MEDICAL IMAGING APPLICATIONS

Nematollah Kayhan Batmanghelich

Christos Davatzikos

Ben Taskar

In this thesis, we propose a method that can be used to extract biomarkers from medical images toward early diagnosis of abnormalities. Surge of demand for biomarkers and availability of medical images in the recent years call for accurate, repeatable, and interpretable approaches for extracting meaningful imaging features. However, extracting such information from medical images is a challenging task because the number of pixels (voxels) in a typical image is in order of millions while even a large sample-size in medical image dataset does not usually exceed a few hundred. Nevertheless, depending on the nature of an abnormality, only a parsimonious subset of voxels is typically relevant to the disease; therefore various notions of sparsity are exploited in this thesis to improve the generalization performance of the prediction task.

We propose a novel discriminative dimensionality reduction method that yields good classification performance on various datasets without compromising the clinical interpretability of the results. This is achieved by combining the modelling strength of generative learning framework and the classification performance of discriminative learning paradigm. Clinical interpretability can be viewed as an additional measure of evaluation and is also helpful in designing methods that account for the clinical prior such as association of certain areas in a brain to a particular cognitive task or connectivity of some brain regions via neural fibres.

We formulate our method as a large-scale optimization problem to solve a constrained matrix factorization. Finding an optimal solution of the large-scale matrix factorization renders off-the-shelf solver computationally prohibitive; therefore, we designed an efficient algorithm based on the proximal method to address the computational bottle-neck of the optimization problem. Our formulation is readily extended for different scenarios such as cases where a large cohort of subjects has uncertain or no class labels (semi-supervised learning) or a case where each subject has a battery of imaging channels (multi-channel), *etc.*. We show that by using various notions of sparsity as feasible sets of the optimization problem, we can encode different forms of prior knowledge ranging from brain parcellation to brain connectivity.

Contents

List of Tables

List of Figures

DW-MRI	diffusion weighted MRI
WM	White Matter
GM	Grey Matter
CSF	Cerebrospinal Fluid
ROI	Region of Interest
VBA	Voxel-Based Analysis
rs-FC	Resting State Functional Connectivity
RSH	Real Spherical Harmonic
DTI	Diffusion Tenor Imaging
FA	Fractional Anistophy
fMRI	functional Magnetic Resonance Imaging
rs-fMRI	resting-state fMRI
FLAIR	Fluid Attenuated Inversion Recovery
T1WI	T1 Weighted Imaging
PET	Positron emission tomography
SPECT	Single-photon Emission Computed Tomography
BOLD	Blood Oxygen Level-Dependent signal
NMF	non-Negative Matrix Factorization
SVD	Singular Value Decomposition
BB	Barzilai Borwein

MP	Matching Persuit
OMP	Orthogonal Matching Persuit
QP	Quadratic Programming
SOCP	Second-Order Cone Programming
SDP	Semi-Definite Programming
LP	Linear Programming
DAG	Directed Acyclic Graph
HMM	Hidden Markov Model
BN	Bayesian Network
SVM	Support Vector Machine
GLM	General Linear Model
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
ICA	Independent Component Analysis
RFE-SVM	Recursive Feature Elimination Support Vector Machine
COMPARE	Classification Of Morphological Patterns using Adaptive Regional Elements [80]
TV	Total Variation

Chapter 1

Overview

Over recent years, there has been an increase in using medical imaging data for extracting biomarkers[1] used in several pathologies. Imaging biomarkers are also useful in clinical trials because they can detect subtle changes in physiology and anatomy very early, therefore assisting in the guided evaluation of a treatment's efficiency. Biomarkers are also important early diagnostic tools; for example, brain degeneration occurs years before clinical symptoms can be observed. In some diseases such as the Alzheimer disease (AD), 33 percent of patients with mild signs may not be diagnosed during their life spans and the diagnosis may not be confirmed completely without a direct examination of brain tissue at autopsy after the person has died [3]. Therefore noninvasive biomarkers can potentially improve early diagnosis of AD and early diagnosis can make treatment more effective.

Surge of demand for biomarkers and increasing amount of medical image available today call for accurate, repeatable, and interpretable approaches for extracting useful and meaningful imaging biomarkers. In this thesis, we developed a general computerized framework that can be used for variety of applications in imaging biomarker extraction. The method showed promising

[1] An imaging biomarker is a feature derived from image that represents a particular aspect of the anatomy or physiology of the organ (*e.g.*, brain) being imaged.

results on various scenarios such as supervised, semi-supervised (in presence of unlabelled data), and unsupervised tasks for uni- and multi-modal imaging, on different datasets such as neurodegenerative brain diseases such as Alzheimer's, mental skill degradation (*e.g.,* verbal skill), *etc.*. We showed that not only it yields accurate predictions but also produces clinically interpretable results that corroborates with what is reported in clinical literature.

Extracting biomarkers from medical imaging is a challenging task because the number of pixels (voxels) in a typical medical image is on the order of millions while even a large sample-size in medical image datasets does not usually exceed a few hundred or at most thousands (curse of dimensionality). Therefore, dimensionality reduction is required to improve the generalization of the classification task. Nevertheless, there are a lot of correlations between voxels and only a parsimonious subset of voxels is typically relevant to the abnormality. We have used different notions of sparsity through this thesis which are inspired by recent literature in Compressed Sensing [15], [69] and machine learning [175], [224]. We have shown that various notions of sparsity can be used to encode different types of prior knowledge about images.

One of the aims of the proposed method is to classify subjects as normal or patient (or perhaps into sub-categories of a disease); this problem falls into the discriminative learning paradigm in machine learning literature. In addition to achieving good generalization performance in term of classification, we desire a method that is clinically interpretable. Clinical interpretability serves two goals:

1. **Extra Validation:** If the areas delineated by the method corroborate clinical findings about the disease, it can provide additional level of qualitative confidence in addition to the quantitative measure (*i.e.,* classification accuracy). For example, for some abnormalities such as Alzheimer's, areas related to memory are usually affected; thus this qualitative measure can be used in tandem with the quantitative measure (*e.g.,* classification between normal and patient subjects).

2. **Incorporating Clinical Prior:** If the clinical interpretability is also considered in the design

of the algorithm, it allows clinical knowledge to be incorporated into the model as a prior. For example, a pathology may only affect gray-matter parts of brain; this prior knowledge can be instrumental to alleviate the curse of dimensionality of the original problem.

A generative framework (*e.g.*, Bayesian) is more appropriate to satisfy the "clinical interpretability" criterion.

In this thesis, we combine those two learning paradigms, generative and discriminative, and address related challenges for medical image classification applications. The proposed method is formulated as a large-scale matrix factorization problem. We use the matrix factorization framework for both modelling our assumptions and for dimensionality reduction in a discriminative way. Large dimensionality of the problem is rooted in the fact that medical images have usually very large dimension. Large-scale matrix factorization also received a lot of attention over the recent years due to its application in learning optimal dictionaries in Compressed-Sensing community [11] or recommendation systems such as Netflix in machine learning community [133], [175]. Finding optimal parameters is usually cast as an optimization problem which can be challenging for large-scale applications.

Our formulation has a few blocks of parameters; some of them are small- to mid-size blocks of variables that can be found via generic or specialized second-order solvers. However, there are also large-size blocks of variables that cannot be found via off-the-shelf solvers; an efficient fast solver is proposed to address this problem which is one of the contributions of this thesis. The optimization method proposed here is an instance of Forward-Backward schemes which have been re-discovered from optimization literature of 80's [158] because of their applicability to solve large-scale inverse problems [23], [168].

In the Section 1.1, after a brief introduction of a few notions and common approaches in medical image classification, we discuss the contributions of this thesis in the Section 1.2. More in depth literature review will be provided in each chapter depending on the topic of the chapter.

1.1 Literature Review

One of the fundamental limitations in medical image classification is the lack of sufficient training samples relative to the high dimensionality of the data. Therefore, a critical step underlying the success of methods that use high-dimensional pattern classification is effective feature extraction and selection, *i.e.,* dimensionality reduction. The main objective of dimensionality reduction is to find or construct a set of image features for a better representation of group difference, to best differentiate between two or more groups, and to improve generalization of a classification problem.

In this section, we first review dimensionality reduction methods for medical image classification applications. Since a choice of feature reduction method also depends on the type of features, most of our focus is on methods that use features that are similar in nature to what we have used in this thesis; *i.e.,* volumetric features rather than shape [92], [34], [233] or cortical thickness [143], [6] features. Dimensionality reduction methods can be categorized into unsupervised and supervised methods. Unsupervised approaches in which class labels are ignored are vaguely similar to generative methods[2] (see Chapter 2, Section 2.4.1 for discussion). Supervised methods take class labels into account and similar to the "discriminative" approach; they try to approximate a map that best approximates or correlates with the class labels. We avoided using the word "discriminative" because they may not explicitly find a map from input features to the class labels. There are also few methods that combine ideas from both ends of the spectrum either from supervised-unsupervised or generative-discriminative point of view.

Voxel-based analysis (VBA) has been widely used in the medical imaging community for group analysis. It typically consists of mapping image data to a standard template space and then applying voxel-wise linear statistical tests on voxel values. General Linear Model (GLM) is used to identify regions of an anatomy (*e.g.,* brain) that are significantly related to the particular effects under study [84]. Standard parametric statistical procedures ($t-$tests and $F-$tests) can

[2] The reason, we avoid to use the term "generative" is that they may not have any generative assumption.

be used to test the hypotheses within the framework of GLM, whereby a vector of observations is modeled by a linear combination of user specified regressors [84]. GLM can be viewed as a generative method that assumes a linear model between response variable (x_i say determinant of Jacobian at i'th voxel) and set of exploratory variables (y_j say levels of experiment):

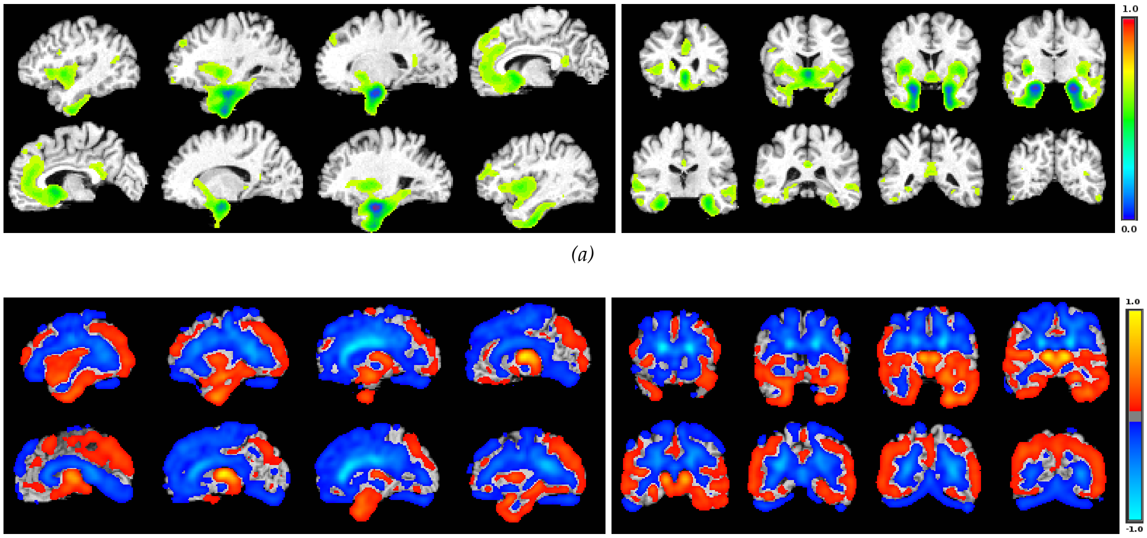
$$x_i = \beta_{i1} y_1 + \beta_{i2} y_2 + \cdots + \beta_{iL} y_L + \epsilon_i$$

where β_j ($1 \leq j \leq L$ and $L < D$ where D number of voxels) are unknown parameters corresponding to exploratory variables. ϵ_i are i.i.d normal random variables $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Two-sample $t-$test is a special case of GLM that assumes $x_{qj} \sim \mathcal{N}(\mu_q, \sigma^2)$, for q'th group; it evaluates the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$. In a typical approach [10, 61, 197, 225], $t-$test is performed at each voxel that results in many statistical tests; hence if no correction is done, the number of false-positives is proportional to the number of independent tests. A False discovery rate (FDR) correction is usually applied to order to compensate for number of tests, but this correction normally does not take spatial smoothness into account. On the whole, VBA identifies regions in which two groups differ (*e.g.,* patients and controls [111]) or regions in which other variables (*e.g.,* disease severity [182]) correlate with imaging measurements. VBA can also be viewed as a correlation-based feature selection [100]. However, VBA has limited ability to identify complex population differences because it does not take into account multivariate relationships in the data [20, 40, 54, 62].

Another popular assumption is that significant brain regions would more likely occur in clusters than in a single voxel [107]. A popular approach is to identify a small number of regions of interest (ROIs) in the brain and aggregate data within these ROIs [89], [86]. Usually data aggregation is done by simply averaging voxel values inside an ROI; with the underlying assumption that the mean of the ROI is a good representation for the whole ROI *i.e.,* $x_i \sim \mathcal{N}(\mu_k, \sigma_k)$ for all i belonging to the k'th ROI region. This is why ROI based methods can be viewed as instances of the generative methods. To define ROIs, an image segmentation is usually done on an *atlas* space and the atlas is registered to the subjects in order to define corresponding areas on the subjects (see Figure 2.5b for examples of ROIs). ROI-based methods usually ignore class labels therefore they can be categorized as unsupervised methods. Plus, ROI's are usually defined based on some cognitive function of a region of brain and do not necessarily follow the boundaries of regions affected by the abnormality; incorrectly defined ROI can cause sever artifacts on the results [87]. Another approach is to use clustering [67], [22] to group voxels into smaller sets. However, a short-coming of the clustering approach is that the clusters cannot overlap: *e.g., region A* cannot belong to *cluster 1* and *cluster 2* at the same time. This can be limiting, for example, in fMRI network discovery because a region of a brain may be involved in multiple networks. In addition, derived clusters might not be optimal for classification. Clustering methods can be viewed as unsupervised generative methods.

Different variations of matrix factorization approaches have been proposed for medical image classification and group analysis purpose. Since the whole image is considered as a high-dimensional sample; such methods are used to reduce the dimensionality. One of the most well-known unsupervised dimensionality reduction method is Principal Component Analysis (PCA) [37,44,105]. PCA can also be combined with ROI analysis, for example [44], first used VBA to identify brain regions with significant difference between two groups of subjects, then applied PCA to the voxels within each ROI. Other variants of matrix factorization such as ICA [25][3] have been also applied particularly for fMRI application [39, 144, 194]. PCA and ICA results are often hard to interpret since they do not specifically attempt to identify localized brain regions, instead, they capture global correlations (see Figure 1.1 for an example). Non-negative Matrix Factorization (NMF) (see Table 2.2) usually improves the representation because of its additive properties that yield part-based representations [146, 235]. Another idea to improve the representation of matrix factorization is to incorporate a sparseness prior. For example, sparse PCA [238], [60] has

[3] Adopting the notation in *Eq.*2.4.4, ICA approximates the data matrix $\mathbf{X}$ as $\mathbf{X} \approx \mathbf{BC}$ s.t. $\|\mathbf{c}_k\|_2 \leq 1$. KL-divergence and negative entropy for $\mathcal{D}(\cdot;\cdot)$ are among common for the divergence terms [114].



(a)

(b)

Figure 1.1: The first and the second rows show examples of applying NMF and SVD on GM RAVEN maps [61] (see Section 2.2.3 for explanation of RAVENS) of brain images. While NMF basis is more localized, SVD eigen basis has non-zero values all over the brain which renders its interpretation very difficult. (a) One of the basis vectors learned by the NMF method on sagittal and coronal cuts and, (b) one of the basis vectors learned by the SVD method on sagittal and coronal cuts.

been applied for modeling anatomical shape variation [191]. However, PCA, ICA, and NMF as unsupervised methods often focus on variations in the data that are irrelevant to the class labels and do not yield the best performance if the main objective is discrimination.

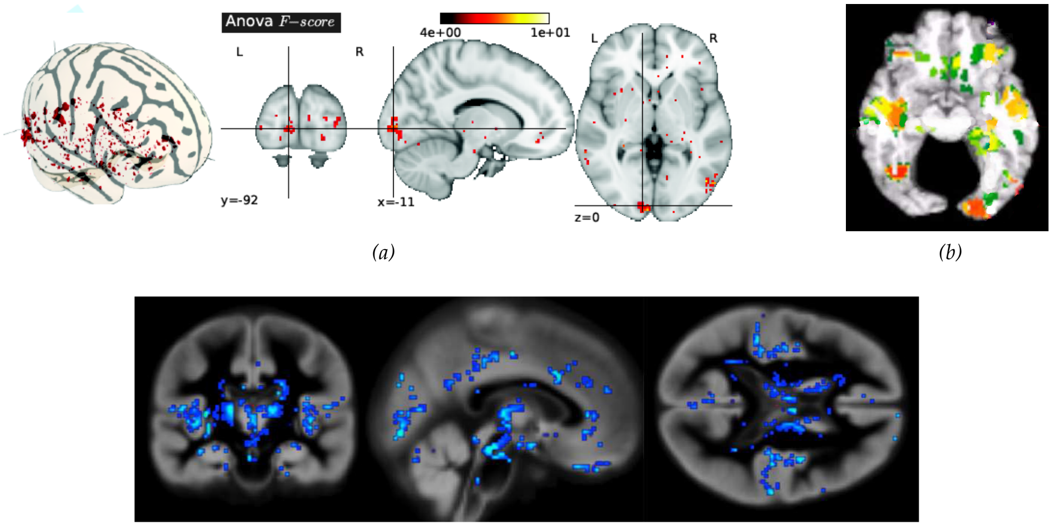
There are also few more formal Bayesian methods particularly applied for fMRI task localization purposes. Lashkari *et al.* [139] proposed a generative Bayesian model using Hierarchical Dirichlet Process [196] as the prior to learn patterns of functional specificity to tasks from fMRI data in a group of subjects. The approach does not need spatial alignment of the subjects to an atlas. It consists of two layers: at the first layer, the functional brain response to each stimulus is modeled as a binary activation variable; and the second layer specifies a prior over sets of activation variables in all subjects. Chen *et al.* [43] proposed a graphical model based method to identify morphological abnormalities automatically, and to find probabilistic associations among voxels in MR images and clinical variables. However, if the objective of a study is classification, such approaches may not perform as well as discriminative methods (see Section 2.4.1).

On the other hand, supervised methods like Linear Discriminant Analysis (LDA) and feature selection methods have been recently applied for medical image analysis [80, 198, 232]. LDA is closely related to ANOVA (analysis of variance) and it approaches the problem by assuming the conditional probability density functions $\mathbb{P}(\mathbf{x}|y = -1)$ and $\mathbb{P}(\mathbf{x}|y = 1)$ are both normally distributed and have the same covariance. Under such assumptions, Bayes optimal solution is to find a threshold c such that for all $\mathbf{x}_i$ in the first class:

$$\mathbf{w}^T\mathbf{x}_i < c, \qquad \mathbf{w} = \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)$$

where μ_1 and μ_2 are the means of the first and the second classes respectively and $\mathbf{\Sigma}$ is the covariance matrix. LDA is a simple method that can be viewed as a supervised generative method. Similar to PCA, LDA may not be able to identify localized abnormal brain regions; in the medical imaging context, the ability of a method to provide an interpretable model is important. In addition, both methods are linear methods and due to the curse of dimensionality, the number of derived basis are limited by the number of subjects which is far smaller than the number of features.

Feature selection methods, on the other hand, output regions that are potentially interpretable. The Recursive Feature Elimination Support Vector Machine (RFE-SVM) ([100], Chapter 5) is an example of feature selection methods. For linear SVM, *i.e.,* $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$, the method boils down to removing the features with the smallest weight in absolute value $|w_i|$. The method is slow for high-dimensional problems such as medical imaging problems. Fan *et al.* [80] proposed a method called COMPARE which is a state-of-the-art algorithm for medical image classification. They suggested first to group pixels into so-called "super-pixels" via applying a watershed segmentation algorithm [209]. The watershed algorithm is used for generating regions according to a discrimination of each local morphological feature. The authors used this measure to improve the robustness of their method against noise and inaccuracy in the registration process.



(a)

(b)

(c)

Figure 1.2: This figure shows examples of feature selection for a few methods: (a) shows maps of Anova (F-score) for the a task prediction in fMRI [156]. (b) shows areas selected by COMPARE [80]; in spite of heuristic used in [80] some areas are too tiny for clinical interpretation. (c) represents voxels picked up by [179] as discriminative ones. In spite of spatial smoothing priors, features tend to be isolated voxels.

The method is a discriminative method that combines a VBM approach with ROI heuristics. In Chapters 3,4 and 5, we compared our method with COMPARE. Vemuri *et al.* [207] proposed an approach called STAND in which the dimensionality is reduced by a sequence of heuristic feature aggregation and selection steps. The heuristic is mostly designed for one kind of features and may fail for other types of features in medical imaging. One of the drawbacks of most feature selection methods is that they are computationally expensive; this is why they mostly rely on some heuristic pre-processing to trim a large portion of the features at the beginning. Another disadvantage of methods using feature selection is that they may produce isolated voxels as relevant features. Voxels are more likely relevant to the class labels as groups rather than isolated voxels and picking isolated voxels as discriminative features may cause over-fitting.

Many machine learning methods with sparsity constraints have been applied to fMRI activation images: Lasso [145], elastic net regression [40], sparse logistic regression [171, 178], or Bayesian regularization [85,227]. Those methods are mostly discriminative methods and usually have the same problem as feature selection methods, *i.e.,* choosing isolated voxels that render interpretablity hard and they are prone to over-fitting. In the context of regression, [156] suggested to use TV-norm (*Eq.*2.1) to incorporate spatial smoothing into objective function; adding such regularization improves spatial contiguity of the detected voxels but since the discriminative methods, in general, ignore the correlation between input features (x_i's), the detected areas may not correspond to any anatomically reasonable region. Sabuncu *et al.* [179] proposed a conditional generating method based on Relevant Vector Machine (RVM[4]) [199]. Unlike RVM, where sparseness is realized by discarding many samples, their approach removes most voxels, retaining only those voxels that are relevant for prediction.

There are few methods that fuse the modeling power of generative approaches with discriminative methods. Argyriou *et al.* [9] used a convex formulation for multi-task classification problems while an orthogonal linear transform of input features is jointly learned with a classifier. In neural networks literature, there are some works on learning compact features with convolutional neural networks [138, 142, 172]. In a different context, a supervised topic model is proposed [32] for movie ratings predicted from reviews, and web page popularity predicted from text descriptions. Very recently, Mairal *et al.* [151] introduced a supervised formulation for learning dictionaries adapted to various tasks instead of dictionaries only adapted to data reconstruction. [151] is very similar in spirit to the work presented in this thesis although in a different context (computer vision) and with a different formulation. In term of formulation, they used different regularization terms than ours and they were used in the objective function rather than as constraints which renders the optimization method significantly different.

1.2 Contributions

Contributions of this thesis can be summarized as follows:

[4] Relevance vector machine (RVM) is a technique that uses Bayesian inference to obtain sparse solutions for regression and classification. The RVM has an identical functional form to the SVM, but provides probabilistic classification.

- **Novel Regularized Matrix Factorization to Extract Informative Features:** As explained in the Section 1.1, most of the existing methods separate feature extraction from classification problem (*e.g.,* [80]). Such separation serves two purposes: 1) feature extraction which is usually done via ROI delineation [89], clustering [67], or segmentation [80] reduces the original dimension (*i.e.,* number of voxels in the training images) significantly, 2) final results are interpretable. In this thesis, we combine those two steps into one framework. We propose a novel formulation that casts the problem as a large-scale constrained matrix factorization which in effect clusters rows (voxels of the images) and classifies columns (subjects) yielding interpretable results. The method finds the clusters that are optimal for a task of interest (*e.g.,* classification or regression) unlike traditional methods in which the feature extraction is done in an unsupervised way and as a pre-processing step. The method allows us to address the curse of dimensionality without compromising classification or producing clinically meaningless results (see Chapters 3,4).
- **Straightforward Extension to Semi-Supervised Learning:** Schematically, the proposed method consists of three building blocks: 1) a generative term in the objective function, 2) a discriminative term in the objective function, and 3) a feasible set. The generative term encourages concise (in our case low-rank and non-negative) reconstruction of the data while the discriminative term encourages good prediction for a task (*e.g.,* class labels in classification). The feasible set encodes prior knowledge by including the set of all acceptable solutions. The modular nature of the method makes it readily extensible for different learning scenarios such supervised, semi-supervised, and unsupervised learning cases. For example, the semi-supervised is useful in medical imaging datasets where there are large sets of subjects not classified as normal but lacking a fully confident disease label[5]. In such cases, a semi-supervised variant of the method can be used to predict future follow-up labels (see Chapter 6). Most of the existing work for semi-supervised learning methods for medical

[5] This is the case for subjects diagnosed as Mild Cognitive Impairment (MCI) who show some impairment in their cognitive scores and have high risk to develop Alzheimer's disease.

imaging does not address which parts of an organ (*i.e.,* brain) undergo changes because they encode only similarities between subjects in which the underlying structure of the image is lost (*e.g.,* [31]).

- **Incorporating Various Clinical Prior as Regularization:** The modular nature of the algorithm also allows various prior knowledge to be encoded in the form of the feasible set. The definition of our feasible set is derived from our generative modeling of the data. Sparsity plays an important role in the definition of the feasible set. We showed how various notions of sparsity can encode which voxels of an image are correlated. Those correlations can be specified by our anatomical understanding about an organ (*e.g.,* connectivity between different regions of a brain). See Chapters 4 and 5 for more details.
- **Discriminative yet Interpretable for Clinical Application:** Unlike feature selection methods that produce good classification accuracy rates in the expense of meaningful anatomical results [80], [179] (see Figure 1.2), our method holds promising good classification rates without compromising anatomical interpretability. This is due to the fact that the generative terms encourages good reconstruction of the data; in fact, our novel formulation chooses a subset of voxels that is optimal for the task (*e.g.,* classification) and also contributes in the reconstruction of the images.
- **Efficient Algorithm for the Large-Scale Optimization:** The algorithm is formulated as a large-scale matrix factorization problem. Finding an optimal solution requires an iterative solution of a few convex optimization problems in order to converge to a local minimum (the formulation is not convex but block-wise convex). The large-scale nature of the problem renders off-the-shelf solver computationally prohibitive; therefore, we proposed a novel solver based on proximal first methods [49]. The technical novelty of the method lies in an almost closed-form solution of a projection sub-problem that is the computational bottle-neck of the algorithm. It turns out even other extensions of the algorithm which are based on different notions of sparsity can exploit the projection algorithm as a module and

inherit its efficiency (see Chapters 4 and 5).

The chapters of this thesis are organized as follows:

Chapter 1: Overview of the thesis and literature review are covered in this chapter.

Chapter 2: Preliminaries and notation are presented in the beginning of the Chapter 2. Due to the multi-disciplinary nature of the problem we address in this thesis, we need to briefly introduce different terms and steps ranging from different modalities of medical imaging used in this thesis to pre-processing steps, various learning paradigms and a few optimization techniques that can be used to solve large-scale optimization problems. Finally, an illustrative example is presented to show the gist of the idea.

Chapter 3: Our novel formulation is detailed in this chapter. We show how the discriminative objective and generative criterion can be cast as a matrix factorization problem. The trade-off between the generative versus discriminative aspects of the formulation is investigated through experiments with synthetic and real data.

Chapter 4: This chapter discusses how different priors can be incorporated as a feasible set of the optimization problem discussed in Chapter 3. It turns out that various notions of sparsity can encode different priors. This chapter also focuses on computational bottle-neck of the large-scale optimization problem and proposes an efficient algorithm for solving it. The algorithm can be extended for other types of applications addressed in this thesis. We also compare classification performance of the algorithm with various choices of the prior with other common or state-of-the-art methods in the literature.

Chapter 5: We extend the basic algorithm proposed in Chapter 3 and 4 in which every subject has one channel image to the case that every subject has a multi-channel image. We view the fMRI time series as an instance of multi-modal image and show how a new notion of sparsity can be defined to incorporate brain connectivity as a prior to guide the inference of functional connectivity.

Chapter 6: The proposed generative-discriminative approach can be readily extended to *semi-supervised learning* where a subset of subjects has labels and a large cohort of subjects are unlabeled but they do contribute in the learning. We showed the applicability of such setting for a medical imaging application.

Chapter 7: Finally, this chapter summarizes what is presented in this thesis and suggests possible avenues for future extensions.

Material presented in this thesis has been published in peer-reviewed conferences and journal papers:

1. Batmanghelich, N.K.; Taskar, B.; Davatzikos, C.; , "Generative-Discriminative Basis Learning for Medical Imaging," Medical Imaging, IEEE Transactions on , vol. 31, no. 1, pp. 51-69, Jan. 2012
2. Batmanghelich, N.K.; Dong, A.; Taskar, B.; Davatzikos, C.; "Regularized Tensor Factorization for Multi-Modality Medical Image Classification," MICCAI, vol. 3, pp. 17-24, Sep. 2011
3. Batmanghelich, N.K.; Dong, H. Y.; Kilian, M. P.; Taskar, B.; Davatzikos, C.; " Disease classification and prediction via semi-supervised dimensionality reduction, " ISBI, pp. 1086-1090, 2011
4. Batmanghelich, N.K.; Taskar, B.; Gooya, A. ; Davatzikos, C.; , " Application of Trace-Norm and Low-Rank Matrix Decomposition for Computational Anatomy, " MMBIA, 2010
5. Batmanghelich, N.K.; Taskar, B.; Davatzikos, C.; , " A General and Unifying Framework for Feature Construction in Image-Based Pattern Classification, " IPMI 2009, pp. 423-432, 2009

Chapter 2

Background

2.1 Notations and Preliminaries

First, we introduce a few notations that will be used throughout this thesis. We use non-capital letters x,y, *etc.*to represent scalar variables. Greek letters (*e.g.*, α,λ) are usually used to represent constants unless stated otherwise; for example relative weights between terms in an optimization problem. Bold lowercase letters denote vectors, *e.g.*, $\mathbf{x} \in \mathbb{R}^N$, and bold uppercase ones represent matrices, *e.g.*, $\mathbf{X} \in \mathbb{R}^{N \times M}$. Subscript and superscript are used to address a column and row of a matrix respectively: *e.g.*, $\mathbf{x}_m \in \mathbb{R}^{N \times 1}$ and $\mathbf{x}^n \in \mathbb{R}^{1 \times M}$. Superscript may also indicate an iteration of a variable in an algorithm. Distinction between, say k'th row of a matrix or k'th iteration of a vector variable, is obvious from the context. Blackboard bold font is used to represent a tensor except the letter $\mathbb{R}$ which is reserved for the set of real values: *e.g.*, $\mathbb{X} \in \mathbb{R}^{N \times M \times K}$, and the letter $\mathbb{P}$ to denote a probability distribution. Calligraphic letters (*e.g.*, $\mathcal{A}$,$\mathcal{B}$) are used sporadically in the text either to denote a set (*e.g.*, $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0\}$) or a special function such $\mathcal{N}$ representing normal distribution. We also use $\{\cdot\}$ to denote a sequence, for example $\{\mathbf{x}^t\}_{t=1}^N$, is a sequence of N variables $\mathbf{x}^1 \cdots \mathbf{x}^N$; if the upper-bound is not given, it means that it is an infinite set (*e.g.*, $\{\mathbf{x}^t\}_{t=1}$). We use $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_p$ to refer to inner product and norm respectively; the subscript of the norm defines types of the norm.

A few different types of norm have been used in this thesis for vectors and matrices. p-norm ($p \geq 1$) of a vector is defined as follows:

$$\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x_i|^p)^{1/p} \tag{2.1.1}$$

where $|x|$ is the absolute value of x. Three common examples of such norm are: 1) ℓ_2 norm: $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ where $\langle \cdot, \cdot \rangle$ denotes inner product; 2) ℓ_1 norm: $\|\mathbf{x}\|_1 = \sum_{n=1}^{N} |x_n|$; 3) ℓ_∞ norm: $\|\mathbf{x}\|_\infty = \max_i\{|x_i|\}$. If $p < 1$, $\|\mathbf{x}\|_p$ does not satisfy properties of a norm, however with abuse of notation, we still call it norm with the same formulation as *Eq.*2.1.1. Perhaps the most interesting example of such a norm is the so-called ℓ_0 norm that counts the number of non-zeros entries of a vector. It is also conceivable to rotate and rescale the vector $\mathbf{x}$ before feeding it to the norm. Namely, for a given semi-definite matrix $\mathbf{Q}$, Mahalanobis (semi)norm $\|\mathbf{x}\|_\mathbf{Q}$ is defined by

$$\|\mathbf{x}\|_\mathbf{Q} = \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}} \tag{2.1.2}$$

Since $\mathbf{Q}$ is positive semidefinite, it can be written as $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where $\mathbf{U}$ is orthogonal and $\mathbf{D}$ is diagonal with non-negative entries. Thus the positive semidefinite root $\mathbf{P} = \mathbf{U}\sqrt{\mathbf{D}}\mathbf{U}^T$ is unique. Therefore, $\|\mathbf{x}\|_\mathbf{Q}^2 = \mathbf{x}^T\mathbf{Q}\mathbf{x} = (\mathbf{P}\mathbf{x})^T(\mathbf{P}\mathbf{x}) = \|(\mathbf{P}\mathbf{x})\|_2^2$. Hence, computing $\|\mathbf{x}\|_\mathbf{Q}$ is equivalent to replace $\mathbf{x}$ with $\mathbf{P}\mathbf{x}$ under ℓ_2-norm; *i.e.*, a rotation and shrinking/stretching of the original $\mathbf{x}$. An interesting example of such a norm for image processing purposes is Total Variation (TV) semi-norm. Assuming that an image is concatenated into a vector $\mathbf{x}$, $TV_2^{1/2}$-norm can be defined as follows:

$$TV_2^{1/2}(\mathbf{x}) = \sum_{j=1}^{N} (\sum_{i \in \mathcal{N}(j)} (x_j - x_i)^2)^{1/2} \tag{2.1.3}$$

where $\mathcal{N}(j)$ is the set of neighbors of the j'th pixel in an image domain. The idea is illustrated

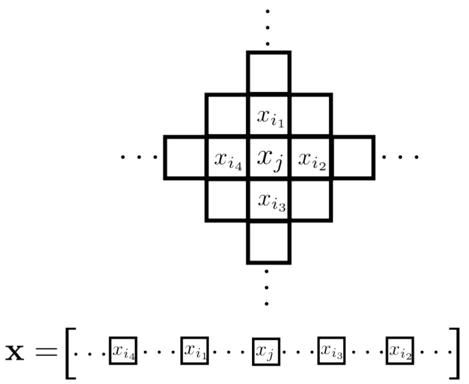


Figure 2.1: TV-norm calculation: the image is concatenated to a vector $\mathbf{x}$. The top figure shows neighbors of the j'th pixel in the image domain: $\mathcal{N}(j) = \{i_1, i_2, i_3, i_4\}$ and corresponding coordinates in the vector $\mathbf{x}$ (bottom) .

in Figure 2.1. In fact, TV-norm measures smoothness of $\mathbf{x}$; the smoother the image, the smaller its gradient; hence the smaller TV-norm. In this case $\mathbf{Q}$ is the Laplacian of a graph representing the image grid. With abuse of the notation, we can define TV_p^q-norm, which may not be even a semi-norm

$$TV_p^q(\mathbf{x}) = \sum_{j=1}^{N} (\sum_{i \in \mathcal{N}(j)} (|x_j - x_i|)^p)^q \tag{2.1.4}$$

For some applications, the indices of the coordinates of a vector may be grouped into a few predefined subsets. We call each subset a group (g_i) and $\mathcal{G}$ denotes the set of all groups (*i.e.,* $\mathcal{G} = \{g_1, g_2, \cdots, g_N\}$). The coordinates of $\mathbf{x}$ within each group are represented as $\mathbf{x}_{|g}$. $p, q-$ group-norm can be defined as follows: 1) q-norm is used to combine entries of each group to a single value ($\|\mathbf{x}_{|g}\|_q$). This results in a $|\mathcal{G}|$-dimensional tuple (or $|\mathcal{G}|$-dimensional vector). 2) *Group-Norm* is defined as the p-norm of the tuple:

$$\|\mathbf{x}\|_{p,q} = (\sum_{g \in \mathcal{G}} \|\mathbf{x}_{|g}\|_q^p)^{1/p} \tag{2.1.5}$$

Groups in $\mathcal{G}$ may or may not overlap (see for examples Figure 2.2). Four common examples

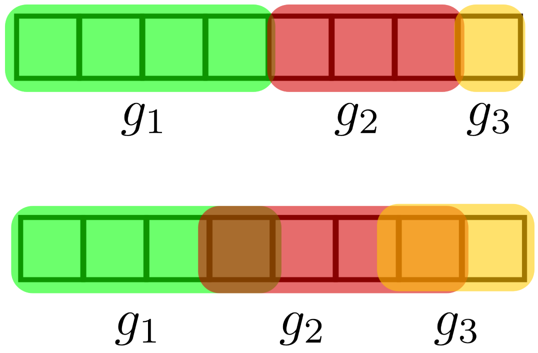


Figure 2.2: Example of overlapping and non-overlapping groups for $\mathbf{x} \in \mathbb{R}^8$ and $\mathcal{G} = \{g_1, g_2, g_3\}$. Top: the non-overlapping group; bottom: overlapping groups.

of group-norms are given:

$$\|\mathbf{x}\|_{1,2} = \sum_{g \in \mathcal{G}} \eta_g \|\mathbf{x}_{|g}\|_2, \quad \|\mathbf{x}\|_{\infty,2} = \max_{g \in \mathcal{G}} \eta_g \|\mathbf{x}_{|g}\|_2$$
$$\|\mathbf{x}\|_{1,\infty} = \sum_{g \in \mathcal{G}} \eta_g \|\mathbf{x}_{|g}\|_\infty, \quad \|\mathbf{x}\|_{\infty,1} = \max_{g \in \mathcal{G}} \eta_g \|\mathbf{x}_{|g}\|_1 \tag{2.1.6}$$

where η_g are constants that can compensate for discrepancy between sizes of the groups.

Similar to *Eq.*2.1.1, we can define p-norm for a matrix $\mathbf{X}$:

$$\|\mathbf{X}\|_p = (\sum_{n=1}^{r} (\sigma_n(\mathbf{X}))^p)^{1/p} \tag{2.1.7}$$

where r is the rank of $\mathbf{X}$ and σ_n is its n'th singular value. This norm is also called Schatten-norm. An example of Schatten norm is the Frobenius norm: $\|\mathbf{X}\|_F := \|\mathbf{X}\|_2 = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$, where $\langle \mathbf{X}, \mathbf{Y} \rangle = trace(\mathbf{X}^T\mathbf{Y})$. Another example is nuclear norm: $\|\mathbf{X}\|_1 = \sum_{n=1}^{r} \sigma_n(\mathbf{X})$ which is simply the sum of the singular values of $\mathbf{X}$.

There is a type of norm for matrices which is similar to the group norm for vectors. This norm is defined on rows or columns of matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ as follows ($p, q \geq 1$):

$$\|\mathbf{X}\|_{p,q} = (\sum_{m=1}^{M} \|\mathbf{x}_m\|_q^p)^{1/p} \tag{2.1.8}$$

where $\mathbf{x}_m$ is the m'th column of the matrix. Unlike the schatten-norm, the norm in *Eq*.2.1.8 is not rotation invariant except the special case $\|\mathbf{X}\|_{2,2} = \|\mathbf{X}\|_F$.

2.2 Pre-Processing of Medical Images

2.2.1 Medical Image Modalities

Many medical image modalities have been developed over the recent decades to quantify various aspects of anatomy and function of tissues. Introducing all medical image modalities and related applications is beyond the scope of this thesis, nevertheless, we briefly present a few modalities used in different applications in this thesis. One may categorize medical image modalities into two general classes: Structural imaging and Functional imaging. Structural imaging reveals anatomical characteristics of underlying tissues, and functional imaging centers on visualizing physiological activities within a certain tissue or organ by measuring changes in metabolism, blood flow, or absorption of different substances (so-called "tracers"). Each modality is generally sensitive toward a particular material hence it provides good contrast for a particular tissue. For example, Magnetic Resonance Imaging (MRI) provides good contrast between the different soft tissues of the body, which makes it especially useful in imaging the brain, muscle, and heart. Even within a modality (*e.g.*, MRI), there are several sub-types specialized for particular tissue types; for example,

- **T1 Weighted Imaging (T1WI):** Water molecules which largely compose the body have two protons. When a person goes inside of a powerful magnetic field, average magnetic moment of those protons becomes aligned with the direction of the field. T1-weighted scans are a standard basic scans designed to differentiate fat from water in a tissue. It is one of the basic pulse sequences in MRI and demonstrates the differences in the T1 relaxation time of tissues [218]. The T1 relaxation time (also known as the spin-lattice relaxation time) is a measure of how quickly the net magnetization vector (NMV) recovers in the direction of the main magnetic field [217] after a perturbation by the pulse.

- **Fluid Attenuated Inversion Recovery (FLAIR):** Fluid attenuated inversion recovery (FLAIR) is a pulse sequence used in magnetic resonance imaging. The pulse sequence is an inversion recovery technique that nulls fluids. For example, it can be used in brain imaging to suppress signals from cerebrospinal fluid (CSF) in the image, so as to bring out hyperintense lesions, such as multiple sclerosis (MS) plaques [220], [13].

- **Diffusion Tensor Imaging (DTI):** Diffusion MRI is a magnetic resonance imaging (MRI) method that produces in vivo images of biological tissues weighted with the local microstructural characteristics of water diffusion, which is capable of showing connections between brain regions [101], [219]. Diffusion tensor imaging (DTI) is important when a tissue such as the neural axons of white matter in the brain has an internal fibrous structure analogous to the anisotropy of some crystals [219].

A few examples of structural medical image modalities for human brain are shown in Figure 2.3.

Functional imaging usually employs tracers or probes to reflect spatial distribution of metabolism within the body. Amount of these tracers are often proportional to some chemical compounds, like glucose, within the body. As examples of functional imaging we can name:

- **PET:** Positron emission tomography (PET) is a nuclear medicine imaging technique that produces a three-dimensional image of functional processes in the body. First, a tracer is introduced into the body on a biologically active molecule, then, the system detects pairs of gamma rays emitted indirectly by a positron-emitting radionuclide. Three-dimensional images of tracer concentration within the body are then constructed by computer analysis. For example, if the biologically active molecule chosen for PET is ^{18}F-Fluorodeoxyglucose (known as FDG), which is an analogue of glucose, the concentrations of tracer give tissue metabolic activity, in terms of regional glucose uptake [221].

- **SPECT:** Single-photon emission computed tomography (SPECT) is a nuclear medicine to-

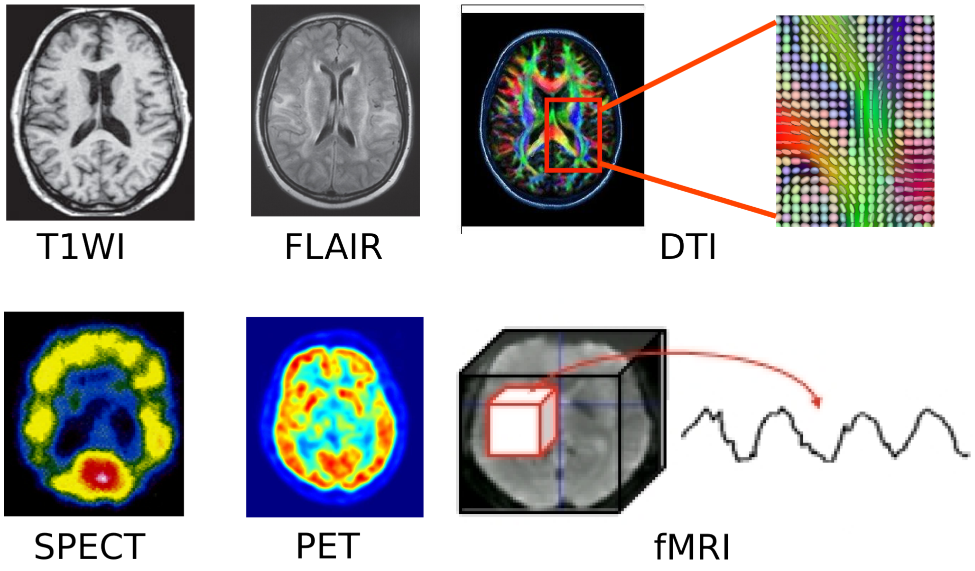


Figure 2.3: Figures on the bottom row are examples of structural image modalities of brain image. Each voxel of DTI represents diffusion properties in that voxel which is represented as a 3×3 matrix. A 3×3 matrix can be represented as a 3-dimensional ellipsoid axes of which are aligned with the eigen vectors of the diffusion matrix; radii of the ellipsoid are proportional to eigen values of the matrix. Figures on the top row show examples of functional image modalities of brain: SPECT, PET and fMRI. fMRI image is basically 4D image (*i.e.,* three dimensions to index location and one to index time); therefore, each voxel contains a time series. PET and SPECT figures are courtesy of [221] and [48] respectively.

mographic imaging technique using gamma rays. The basic technique requires injection of a gamma-emitting radioisotope (called radionuclide) into the bloodstream of the patient [222].

- **fMRI:** Functional magnetic resonance imaging or functional MRI (fMRI) is an MRI procedure which measures brain activity by detecting associated changes in blood flow. The primary form of fMRI uses the blood-oxygen-level-dependent (BOLD) contrast. fMRI is used to map neural activity in the brain or spinal cord of humans or animals by imaging the change in blood flow (hemodynamic response) related to energy use by brain cells [223].

A few examples of medical image modalities for human brain are shown in Figure 2.3.

2.2.2 Pre-Processing

Pre-processing steps is required before applying any algorithm but the actual steps may vary significantly depending on image modalities, anatomy that is being studied, and obviously the application of interest. Here, we limit ourself to brain imaging modalities, which were introduced on the previous section and focus on applications directly related to the purpose of this these, *i.e.,* image classification and group analysis. We only mention common pre-processing steps that might be necessary for understanding other section of the thesis. Discussing details of each step or any extra steps are beyond the scope of this chapter and will be mentioned in each chapter if it is necessary.

The diagram showed in Figure 2.4 represents a typical pre-processing pipeline used for group analysis and classification purposes in brain imaging. There are many other blocks that can be added to the diagram but we only mentioned the most general ones:

- **Image Enhancement:** A common step in a pre-processing pipeline is image enhancement. We use this step in a broad sense; *i.e.,* any step that improves image quality can be a part of this block. For example, denoising or bias field correction[1], or histogram equalization[2] or motion correction[3] can be viewed as an image enhancement step. There might be several of such blocks in a typical pre-processing pipeline; it can be done before or after image registration.
- **Tissue Segmentation:** This step can be viewed as a part of feature extraction step. Since it is very common step particularly for brain image analysis, we introduce it as a separate step. The fundamental task in tissue segmentation is to classify the voxels in the volumetric MR data into subclasses of tissue types, *e.g.,* gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) tissue types [12], [234] (see Figure 2.5a for an example).

[1] Bias field signal is a low-frequency and very smooth signal that corrupts MRI images [210], [36].

[2] Histogram equalization is a method in image processing to adjust contrast using the image's histogram [103], [5].

[3] One of the major sources of error in the analysis of functional Magnetic Resonance images is the presence of spurious activation arising due to patient head movement at the time of image acquisition. Motion correction algorithms are designed to remove this artifact [161].

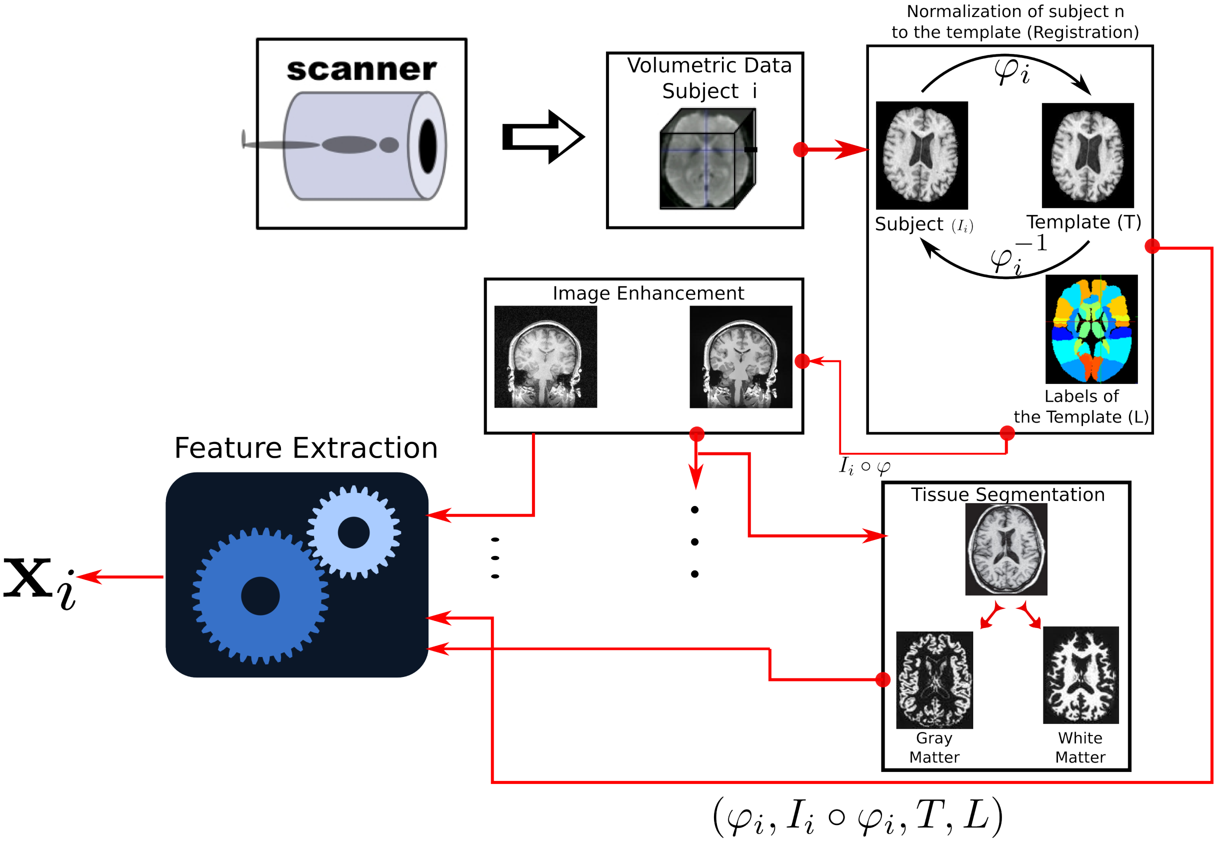
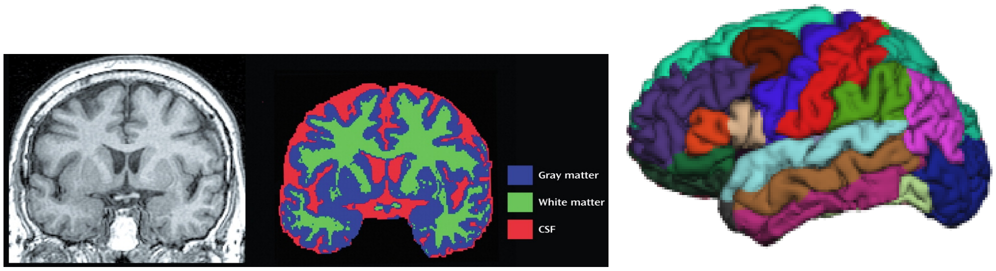


Figure 2.4: The figure shows a general pre-Processing pipeline in medical imaging application for group-analysis and classification purpose: to establish voxel-wise correspondence, volumetric image of the i'th subject (I_i) should be aligned (warped) to a template (T). This normalization process is called *registration* and produces a mapping (φ_i) for the i'th subject. The map (φ_i), the warped subject ($I_i \circ \varphi_i$), the template image (T), and the corresponding label image (L) are feed to a black-box for feature extraction. The warped image ($I_i \circ \varphi_i$) can also be used by many other blocks two of the most common of which are shown here: Image enhancement and Tissue segmentation. The image enhancement block may include histogram equalization, bias field correction, or any other procedure to enhance the quality of an input image. Enhanced image may be used for tissue segmentation (or any other block). This tissue segmentation block classifies voxels of the image into various tissue sub-types: White matter, Gray matter, *etc.*. All results are *optionally* provided to the feature extraction block that in turn produces the feature vector $\mathbf{x}_i$.

- **Registration:** In order to compare images of different subjects, one may need to maintain voxel-wise correspondence. For example, to compare subject i with subject j, we need to know which coordinate of, say the i'th subject, corresponds to, say coordinate (z_1, z_2, z_3) of the j'th subject. Therefore, a one-to-one mapping (φ) representing the correspondence is computed during the registration process. Instead of having pair-wise maps between all pairs of subjects, it is common to find a map to a common image called *Template* or



(a) *(b)*

Figure 2.5: (a) shows an example of tissue segmentation (courtesy of [14]). (b) shows an example of structural segmentation; each color denotes a structure. Each segment can be used as a region of interest (ROI) for feature extraction step.

Atlas. Image registration is a topic of research on its own right and here we only give a brief introduction (see [192], [237], [149], [57], [129] and references therein for a survey on medical image registration methods).

A subject image (I) and the template (T) are viewed as a function that maps compact domains (*i.e.,* Ω_1 and Ω_2 respectively) to a set of real values, namely: $I : \Omega_1 \to \mathbb{R}$ and $T : \Omega_2 \to \mathbb{R}$ where $\Omega_1, \Omega_2 \subset \mathbb{R}^3$ (assuming that the image is a volumetric image). A registration algorithm solves the following optimization problem:

$$\min_{\theta \in \Theta} \mathcal{D}(I_i \circ \varphi(\theta); T) \tag{2.2.1}$$

where $\varphi(\theta) : \Omega_2 \to \Omega_1$ is the one-to-one mapping[4] parametrized by θ and Θ is the set of all possible parameters and $\mathcal{D}(\cdot;\cdot)$ is a measure of distance (a divergence function); *e.g.,* $\mathcal{D}(\cdot;\cdot) = \int_{\Omega_2} \|T(\mathbf{z}) - (I \circ \varphi(\theta))(\mathbf{z})\|_2 d\mathbf{z}$. $I \circ \varphi(\theta)$ means composition (warping) the subject image according to the mapping function. The idea is pictorially represented in Figure 2.6a.

[4] One-to-one mapping is usually not enough and φ needs to be smooth too. Mathematically speaking, φ should be a *Diffeomorphic* map: it is a bijection map that is differentiable and its inverse is also differentiable.

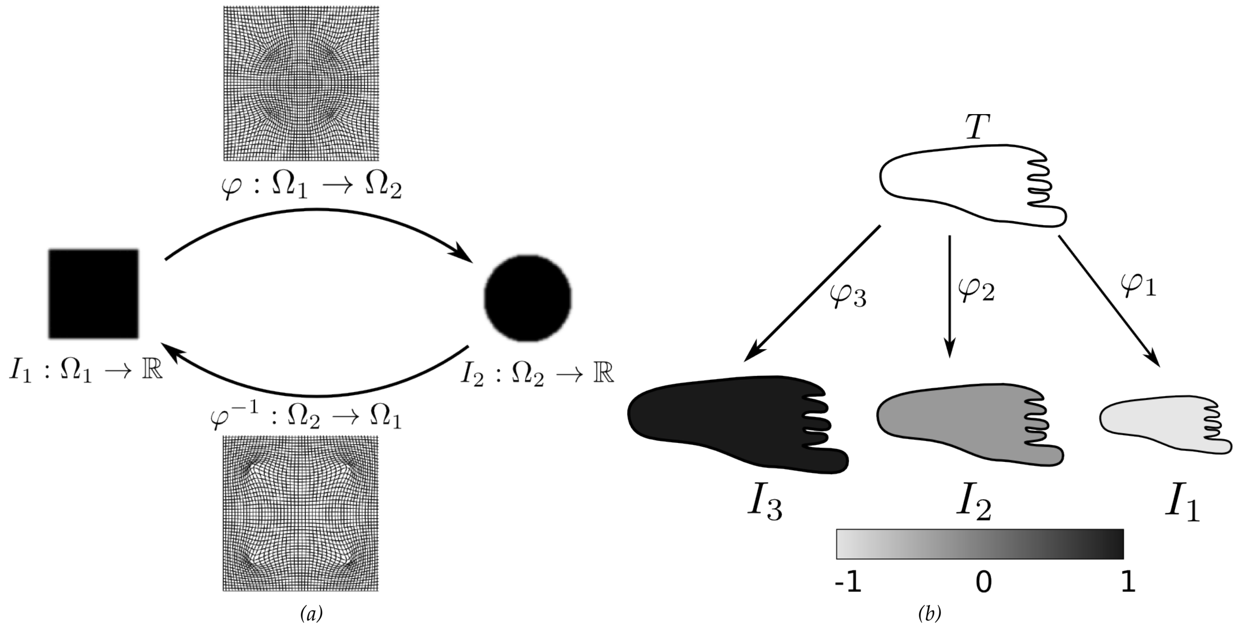

$\varphi : \Omega_1 \to \Omega_2$
$I_1 : \Omega_1 \to \mathbb{R}$
$I_2 : \Omega_2 \to \mathbb{R}$
$\varphi^{-1} : \Omega_2 \to \Omega_1$
(a)
T
φ_3
φ_2
φ_1
I_3
I_2
I_1
-1
0
1
(b)

Figure 2.6: (a) show registration concept: φ and φ^{-1} map the box to the circle and vice-versa respectively. Warped grids show local deformation. (b) shows the idea of determinant of Jacobian; the template object T is mapped to the three objects (I_1,I_2 and I_3). The color encodes logarithm of the determinant of the Jacobian of the transformations. If T is expanded, the determinant of the Jacobian is greater than one (its logarithm is positive), and it is less than one if the template object is shrunk.

2.2.3 Feature Extraction

There are abundant feature extraction methods for medical imaging application. Choice of the features and the algorithm depends on the modality and the application. Giving an exhaustive list of algorithms and features is beyond the scope of this chapter; therefore, we limit ourselves to features mentioned in this thesis.

- **Intensity:** In some modalities intensity value of an image is informative. For example, in Positron emission tomography (PET) voxel intensity encodes concentration of a tracer (*e.g.,* Fluorodeoxyglucose, an analogous of glucose) in a particular location of a tissue (*e.g.,* brain). Sometimes, intensity value should be mapped to a meaningful value. For example, in Diffusion Tensor Imaging (DTI) (see Figure 2.3), each voxel is not a scalar but a positive semi-definite matrix (diffusion tensor). From a DTI image, one can compute fractional anisotropy[5] (FA) [115], [212] or trace[6] maps that measure relative degree of anisotropy and total diffusivity in a voxel respectively. Many other features can be extracted from DTI images [216].

- **Deformation Features:** For some aims, voxel intensity itself may not be directly informative. For example, measuring deformation (*e.g.,* shrinkage or expansion) of a brain structure with respect to population average might be meaningful for detecting neurodegenerative diseases. There are various approaches to quantify deformation of a structure. One approach is to segment structures of interest and study them as independent objects using shape analysis methods [93], [124], [41], [167]. Figure 2.5b shows examples of brain segmentation into cortical and sub-cortical regions. One can study deformation of each region separably or together.

  Volumetric approaches suggest another way to address the problem [46], [94], [10]. In a volumetric approach, the images in a dataset are registered to a template[7] and determinant of Jacobian of the deformation fields are extracted as informative features. Determinant of Jacobian is a non-negative value for a diffeomorphic map (*e.g.,* deformation field of a registration map) that quantifies local shrinkage or expansion of tissues [61], [80]. The idea is shown in Figure 2.6b. If a part of a tissue undergoes shrinkage, the determinant of the Jacobian of deformation for voxels inside of that part are less than one (their logarithm are negative) and vice-versa for areas undergoing expansion. Alternatively in this thesis, we use a feature, so-called *RAVEN* map, that uses both deformation field and tissue segmentation to quantify local expansion and shrinkage of the tissue types. RAVEN has the advantage of accounting for imperfect registration by taking residual (error) of the imperfect registration into account [61].

[5] Each voxel in DTI is a positive semi-definite matrix. Assuming that $\lambda_1, \lambda_2, \lambda_3$ are eigen values, $FA = \sqrt{\frac{(\lambda_1-\lambda_2)^2+(\lambda_1-\lambda_3)^2+(\lambda_2-\lambda_3)^2}{4(\lambda_1^2+\lambda_2^2+\lambda_3^2)}}$. The idea is that spherical diffusion voxel takes FA value close to 0. Elongated diffusion voxel FA takes value close to 1.

[6] Assuming $\lambda_1, \lambda_2, \lambda_3$ are eigen values of the diffusion tensor, the $Trace = \lambda_1 + \lambda_2 + \lambda_3$.

[7] The template image is either chosen or estimated from the database in a unbiased way [147].

- **Features extracted from time series:** In fMRI, each voxel contains a time series. There are various approaches to extract features from such datasets; obviously a design of features depends on the application. A popular approach is to fit a general linear model (GLM) to find correlated voxels with a task and use the parameters of the regressions (β-map) as features [84], [72] [157]. Alternatively, parameters of time series models (*e.g.,* autoregressive model [176]) or even time series itself [104] can be used to extract features. Another alternative is to apply a spatio-/temporal transformation on the data first, and use its coefficient as features [195].

2.3 Optimization with Sparsity

In this section, we provide some background material related to optimization and sparsity methods used in this thesis. First in Section2.3.1, the relationship between norm for regularization and sparsity is introduced; we illustrate why they yield sparse solutions. Specific focus of this thesis is on medical imaging applications that call for usually large-scale optimization problems; Section 2.3.2 briefly presents an efficient first-order optimization framework for such applications.

2.3.1 Sparsity-Inducing Norms

Finding a subset of covariates that correlates with a quantitative response has been a staple of statistical analysis for a long time [81]. In machine learning, this problem is usually referred as *Feature Selection* [100]. Feature selection is usually performed to select relevant features to 1) gain predictive accuracy, 2) gain knowledge about the process that generated the data or simply visualize the data, 3) limit storage requirements and increase algorithm speed [81]. Reviewing feature selection methods is beyond the scope of this section (see [81] and references therein). Here, we only focus on an optimization point of view of feature selection. The optimization problem usually consists of fitting some model parameters $\mathbf{w} \in \mathbb{R}^p$ to training data while using few parameters of $\mathbf{w}$:

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \ell(f_{\mathbf{w}}(\mathbf{x}_i); y_i) + \lambda \Omega(\mathbf{w}) \tag{2.3.1}$$

where $(\{(\mathbf{x}_i, y_i\}_{i=1}^{N}))$ is pair of training features $(\mathbf{x}_i)$ and observations (*e.g.,* class labels) respectively. $f_{\mathbf{w}(\mathbf{x})}$ is the model parametrized by $\mathbf{w}$ and $\ell(\cdot;\cdot)$ is a loss function measuring mismatch between the model and the observations. $\Omega(\mathbf{w})$ is a regularizer designed to control the complexity of $f_{\mathbf{w}}(\cdot)$. λ controls the trade-off between the loss function and the regularizer. In order to promote sparsity, a natural choice for $\Omega(\mathbf{w})$ is ℓ_0 norm but it renders solving *Eq.*2.3.1 computationally intractable. Therefore, one needs to approximate the solution using greedy methods or ℓ_1 convex relaxation methods (Section 2.3.2). To understand why ℓ_1 relaxation yields a sparse solution, consider the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{y} - \mathbf{A}\mathbf{w}\|_2^2 \\ \text{subject to:} \quad & \|\mathbf{w}\|_1 \leq \alpha \end{aligned} \tag{2.3.2}$$

where α and $\mathbf{A}$ are a parameter and a constant matrix respectively. To compare *Eq.*2.3.2 with *Eq.*2.3.1, notice that the loss function is ℓ_2 norm and $\Omega(\mathbf{w}) = \{0, \text{ if } \|\mathbf{w}\|_1 \leq \alpha; \infty, \text{ otherwise}\}$. In fact, even if ℓ_1 is used as a regularizer (*i.e.,* $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$) rather than a constraint, *Eq.*2.3.1 and *Eq.*2.3.2 follow the same regularization path[8]. From convex optimization [35], we know that at the optimal solution, $\mathbf{w}^*$, the level-set[9] of the objective function corresponding to $\mathbf{w}^*$ is tangent to ℓ_1 ball of radius α. The idea is pictorially represented in Figure 2.7. The figure shows this tangency on the balls of ℓ_0-, ℓ_q-$(0 < q < 1)$ and ℓ_1- and ℓ_2-norms in the two dimensional case. Due to the anisotropic behavior of ℓ_q-norms $(0 \leq q \leq 1)$, they encourage solutions to be on one of

[8] Two optimization problems: 1) $\mathbf{w}^* = \arg\min_{\mathbf{w}} f(\mathbf{w}) + \lambda\Omega(\mathbf{w})$, and 2)$\mathbf{w}^{**} = \arg\min_{\mathbf{w}} f(\mathbf{w}) \quad \text{s.t. } \Omega(\mathbf{w}) \leq \alpha$, follow the same regularization path if $\forall\lambda > 0, \exists\alpha > 0$ such that $\mathbf{w}^* = \mathbf{w}^{**}$.

[9] Level set of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ corresponding to $\mathbf{w}$ is: $\mathcal{C}(\mathbf{w}) = \{\mathbf{x} \in dom(f) : f(\mathbf{x}) \leq f(\mathbf{w})\}$ [35]; where $\mathcal{C}(\mathbf{w})$ is a set parametrized by $\mathbf{w}$ and $dom(f)$ denoted domain of f.

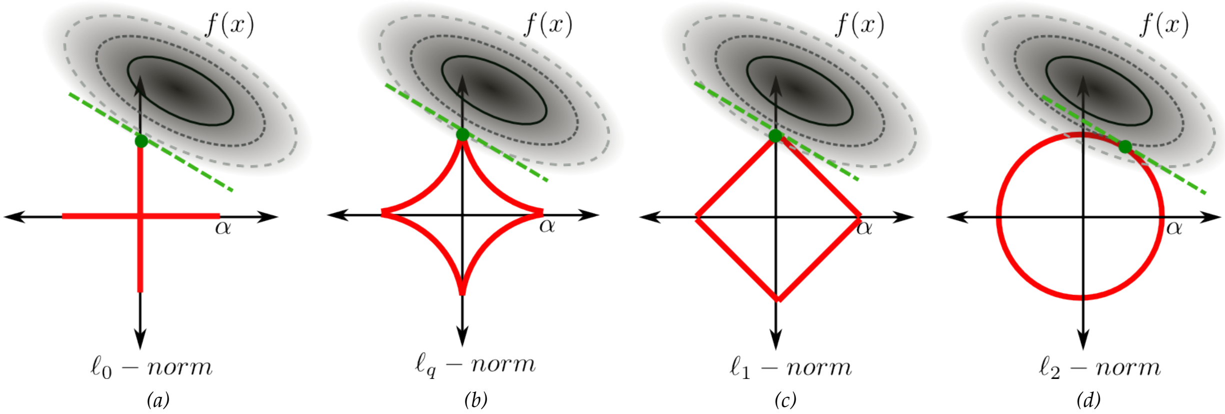


Figure 2.7: The figure shows balls of radius α for different norms. The green dots represent the optimal points and the green dashed lines are tangent lines to level sets of the objective function ($f(x)$). ℓ_0, ℓ_q ($0 < q < 1$), and ℓ_1 encourage sparse solutions because it is more "likely" for a tangent line to touch on the α-balls on the corners.

the axis which corresponds to a sparse solution (because the value of the variable corresponding to the other axes are zero). However, ℓ_2-norm is isotropic and does not enjoy the same property.

If $\Omega(\mathbf{w})$ is replaced with ℓ_0-norm, the resultant optimization problem is NP-hard in general; however some greedy procedures have been proposed for a sub-class of *Eq.*2.3.1, namely $\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{A}\mathbf{w}\|_2^2$ s.t. $\|\mathbf{w}\|_0 \leq \alpha$. Forward selection technique in statistics [214], Matching Pursuit (MP) [155], and Orthogonal Matching Pursuit [154] are examples of such algorithms; some optimality guarantees have been shown for a few cases [201]. While greedy algorithms are the right choice for small dimensional problems, they may not be applicable for medical imaging applications. In fact, the dimensionality of medical imaging problems inflicts a high computational cost. In addition, it is not easy to incorporate constraints such as non-negativity or complicated group-norm regularizations[10] into the optimization problem.

Next, we focus on methods that solve convex relaxation for large-scale optimization problems with a sparsity term which are specifically useful in the problems presented in this thesis.

[10] Recently Lozano *et al.* [148] proposed an algorithm for group-sparsity norm with ℓ_0-norm.

2.3.2 Convex Relaxation for Sparse Algorithm

We mentioned in Section 2.3.1 that variants of ℓ_1 regularization can be used as a surrogate for ℓ_0 norm. Due to its anisotropic behavior, it encourages sparsity while being computationally tractable for small- and medium-scale problems. The computational complexity of the convex relaxation depends on the exact form of the optimization problem. There are several generic solvers such as Linear Programming (LP), Quadratic Programming (QP), Second-Order Cone Programming (SOCP), *etc.* [35] that can address various forms of the objective function and constraints. Most of such methods are based on the barrier method to handle constraints [35] and need to solve a system of linear equations (Newton system) to incorporate second order information[11]. However, except in cases where the Newton system is low-rank, memory requirements and computational complexity are cost prohibitive for large-scale problems. Therefore, a first order method needs to be employed for the optimization.

Naive first order methods (*i.e.,* gradient descent) yield very slow convergence rate. To improve convergence rate, we use the Proximal method [49] that generalizes the first order descent algorithm and can handle non-smooth components in the objective. To introduce proximal method, let us assume that we want to solve the following optimization problem:

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \tag{2.3.3}$$

where both $f(\cdot)$ and $\Omega(\cdot)$ are convex functions and $f(\mathbf{w})$ is smooth and $\nabla f(\mathbf{w})$ is L-Lipschitz [12]. Notice that $\Omega(\mathbf{w})$ does not need to be smooth; it may be a non-smooth function or a representation

[11] For a non-constrained problem: $\min_{\mathbf{x}} f(\mathbf{x})$, Newton method computes the descent direction ($\Delta\mathbf{x}_{nt}$) as follow: $\nabla^2 f(\mathbf{x})\Delta\mathbf{x}_{nt} = -\nabla f(\mathbf{x})$. In a constrained case, this equation is replaced with the KKT equations [35].

[12] A function $f(\mathbf{x})$ is called L-Lipschitz if: $\forall \mathbf{x}_1, \mathbf{x}_2 \in dom(f), |f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|_2$.

of a feasible set (*e.g.,* $\mathcal{W}$),

$$\Omega(\mathbf{w}) = \begin{cases} \infty & \text{if } \mathbf{w} \notin \mathcal{W}; \\ 0 & \text{if } \mathbf{w} \in \mathcal{W}. \end{cases} \tag{2.3.4}$$

which maintains the following equivalence: $\min_{\mathbf{w}} f(\mathbf{w}) + \Omega(\mathbf{w}) \equiv \min_{\mathbf{w} \in \mathcal{W}} f(\mathbf{w})$.

The most basic Proximal method for solving *Eq.*2.3.3 generates a sequence of iterates $\{\mathbf{w}^t\}_{t=1}$ element of which solves the following subproblem

$$\mathbf{w}^{t+1} \in \arg\min_{\mathbf{z}} f(\mathbf{w}^t) + (\mathbf{z} - \mathbf{w}^t)^T \nabla f(\mathbf{w}) + \frac{\eta_t}{2}\|\mathbf{z} - \mathbf{w}^t\|_2^2 + \lambda\Omega(\mathbf{w}^t) \tag{2.3.5}$$

where $f(\mathbf{w}^t) + (\mathbf{z} - \mathbf{w}^t)^T \nabla f(\mathbf{w})$ is a linear approximation of f around the current estimate $\mathbf{w}^k$. To keep the linear approximation correct, the quadratic term $\|\mathbf{z} - \mathbf{w}^t\|_2^2$ keeps the update of $\mathbf{w}$ in a vicinity of current solution $\mathbf{w}^k$. In fact, $f(\mathbf{w}^t) + \frac{\eta_t}{2}\|\mathbf{z} - \mathbf{w}^t\|_2^2$ can be viewed as a quadratic approximation of f around $\mathbf{w}^t$ assuming a simple diagonal Hessian approximation $\eta_t \mathbf{I}$; we come back to this in the sequel.

Removing all terms irrelevant to the optimization in *Eq.*2.3.5 (*e.g.,* $f(\mathbf{w}^t)$) and absorbing the linear term into the quadratic term, *Eq.*2.3.5 is equivalent to

$$\begin{aligned} \mathbf{w}^{t+1} &\in \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{u}^t\|_2^2 + \frac{\lambda}{\eta_t}\Omega(\mathbf{z}) \\ \text{where } \mathbf{u}^t &= \mathbf{w}^t - \frac{1}{\eta_t}\nabla f(\mathbf{w}^t) \end{aligned} \tag{2.3.6}$$

Therefore in every iteration of the algorithm, a sub-problem called *Proximal operator* needs to be solved. More formally, *proximal operator* or *proximity operator* for Ω is

$$\mathcal{P}_{\lambda\Omega}(\mathbf{u}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{z} - \mathbf{u}\|_2^2 + \lambda\Omega(\mathbf{z}) \tag{2.3.7}$$

where $\lambda > 0$.

In fact, the subproblem in *Eq.*2.3.6 is proximal operator for $\frac{\lambda}{\eta_t}\Omega$. Proximal operators are generalizations of the orthogonal projection operator[13]; Table 2.1 shows examples of proximal operators for a few popular regularization functions; see [159], [49], [226], [23], and [150] for more discussion and applications.

Table 2.1: The table shows a few examples of popular regularization functions and corresponding proximal operators. The entries of the first column are the functions and the entries of the second column are $\mathcal{P}_{\lambda\Omega}$. λ is a positive constant and $\prod_{\mathcal{C}}(\mathbf{x})$ denotes the orthogonal projection of $\mathbf{x}$ on the set $\mathcal{C}$.

$\lambda\Omega(\mathbf{x})$	$\mathcal{P}_{\lambda\Omega}(\mathbf{x})$	Description
$\lambda\lVert\mathbf{x}\rVert_1$	$\frac{\max\{\lvert\mathbf{x}\rvert-\lambda\}}{\max\{\lvert\mathbf{x}\rvert-\lambda\}+\lambda}\mathbf{x}$	Known as "soft-thresholding operator" ($\lvert\mathbf{x}\rvert$ indicates element-wise absolute value)
$\lambda\lVert\mathbf{x}\rVert_2$	$\begin{cases} 0 & \text{if } \lVert\mathbf{x}\rVert_2 \leq \lambda; \\ \frac{\lVert\mathbf{x}\rVert_2-\lambda}{\lVert\mathbf{x}\rVert_2}\mathbf{x} & \text{if } \lVert\mathbf{x}\rVert_2 > \lambda. \end{cases}$	Zero if it is inside λ-ball of ℓ_2, otherwise rescaling
$\lambda\lVert\mathbf{x}\rVert_\infty$	$\mathbf{x} - \prod_{\lVert\mathbf{x}\rVert_1 \leq \lambda}$	Notice the relationship with the dual norm
$\begin{cases} \infty & \text{if } \mathbf{x} \notin \mathcal{C}; \\ 0 & \text{if } \mathbf{x} \in \mathcal{C}. \end{cases}$	$\prod_{\mathcal{C}}(\mathbf{x})$	Orthogonal projection on the set $\mathcal{C}$
$\lambda\phi(\mathbf{x})$	$\mathbf{x} - \prod_{\phi^*(\cdot)\leq\lambda}(\mathbf{x})$	$\prod_{\phi^*(\cdot)\leq\lambda}(\mathbf{x})$ is the orthogonal projector onto the ball of radius λ of the dual function ϕ^* (proof in [50])

Choosing η_t can also be done automatically. One approach is to use the method proposed by Nestrov *et al.* [159], or Beck *et al.* [23] (FISTA) that uses a practical line search. Another approach is to use the Barzilai-Borwein (BB) spectral method [16]. Remember in *Eq.*2.3.5, $f(\mathbf{w}^t) + \frac{\eta_t}{2}\|\mathbf{z} - \mathbf{w}^t\|_2^2$ can be viewed as a quadratic approximation of f around $\mathbf{w}^t$ assuming simple diagonal Hessian approximation $\eta_t\mathbf{I}$. BB method suggests to optimally estimate η_t. Assuming $\mathbf{s}^t = \mathbf{w}^t - \mathbf{w}^{t-1}$ and

$$\mathbf{r}^t = \nabla f(\mathbf{w}^t) - \nabla f(\mathbf{w}^{t-1})$$

If the diagonal approximation is a good approximation, $\eta_t\mathbf{s}^t \approx \mathbf{r}^t$. In least-square sense:

$$\eta_t = \arg\min_{\eta} \|\eta\mathbf{s}^t - \mathbf{r}^t\|_2^2 = \frac{(\mathbf{s}^t)^T\mathbf{r}^t}{(\mathbf{s}^t)^T\mathbf{s}^t} \tag{2.3.8}$$

[13] Orthogonal projection of given parameter $\mathbf{u}$ on a set $\mathcal{C}$ solves the following optimization problem: $\prod_{\mathcal{C}}(\mathbf{u}) = \arg\min_x \|\mathbf{u} - \mathbf{x}\|_2$ s.t. $\mathbf{x} \in \mathcal{C}$

This is not the only way to approximate the Hessian as a diagonal matrix and there are other variants of the BB method (see [16], [59], [226]). The BB method is usually a non-monotone method (*i.e.*, it is not a descent method), occasional increase in the objective appears to be essential in good the performance of BB method. Similar to SpaRSA [226], we use a safeguard method to ensure that $\eta_t \in [\eta_{min}, \eta_{max}]$. Our acceptance criteria for a step $\mathbf{w}^t$ is also similar to [226], namely

$$\phi(\mathbf{w}^{t+1}) \leq \max_{i=\max(t-M,0),\cdots,t} \phi(\mathbf{w}^i) - \frac{\sigma}{2}\eta_t \|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2, \tag{2.3.9}$$

where ϕ is the objective function we want to minimize and M is a constant. The meaning of *Eq*.2.3.9 is that ϕ^{t+1} must be decreasing with respect to maximum of the last M objective values but increase with respect to the ϕ^t is allowed. Notice that if $M = 1$, it is a descent method. The convergence behavior of *Eq*.2.3.9 is studied in [97] and [226]. Its good performance is also shown empirically in this thesis and also in SpaRSA [226] and many other studies [213], [215].

2.4 Learning

This section reviews different learning paradigms in machine learning community. Section 2.4.1 compares generative and discriminative learning approaches. Section 2.4.2 briefly introduces graphical models as a modeling approach for distributions. Section 2.4.3 presents matrix factorization as an instance of generative framework and Section 2.4.4 lays out Support Vector Machine as a discriminative framework.

2.4.1 Generative vs Discriminative Approaches

For this section, we assume that N samples $(\mathbf{x}_i)$ and corresponding class labels (y_i) are given; each sample is potentially in multi-dimensional space $(\mathbf{x}_i \in \mathbb{R}^D)$. We represent this setting as a set of N pairs: $\mathcal{Z} = \{\mathbf{z}_i \equiv (\mathbf{x}_i, y_i)\}_{i=1}^N$. An example of such a setting could be $\mathbf{x}_i$ representing an image acquired from the MR scan of a subject and y_i, a label denoting normal or abnormal status of the same subject. In order to learn a relationship between $\mathbf{x}$ and y, one can choose a method from two schools of thought in machine learning [119]: generative and discriminative approaches[14]. Providing a comprehensive survey over generative or discriminative approaches is beyond the scope of this chapter; therefore, we supply sufficient background for methods used in this thesis (see [119] for more in depth discussion). In this section, we give a brief introduction to generative and discriminative methods. Section 2.4.2 is devoted to a short introduction to graphical models as a modeling approach for generative models; section 2.4.3 presents the matrix factorization framework as a model to represent data in a generative manner. Section 2.4.4 introduces Support Vector Machine (SVM) as one of the most popular discriminative methods.

Assuming generative and discriminative methods lay on two ends of a spectrum, generative models at one end attempt to estimate a distribution over all variables; for example in our setting for both $\mathbf{x}$ and y. In a generative approach, covariates ($\mathbf{x}$) and the observation (y) and potentially hidden variables are modeled by a joint probability distribution: $\mathbb{P}(\mathbf{z}) = \mathbb{P}(\mathbf{x}, y)$. Examples of such approaches are mixture of Gaussians [29], hidden Markov Models (HMM) [169], naive Bayes and, Markov random fields [230]. These models are usually parametrized by a set of parameters Θ. Given the training dataset ($\mathcal{Z}$), the parameters should be estimated. Generative learning has different varieties [119], ranging from local estimation that only considers performance on the given training data (maximum likelihood): $\Theta^* = \arg\max_\Theta \mathbb{P}(\Theta|\mathcal{Z})$, to combination of a fitness term and a prior term (maximum a posteriori): $\Theta^* = \arg\max_\Theta \mathbb{P}(\Theta|\mathcal{Z})\mathbb{P}(\Theta)$, to fully weighted averaging over all possible hypothesis in a hypothesis space (Bayesian inference): $\mathbb{P}(\mathbf{z}|\mathcal{Z}) = \int d\mathbb{P}(\mathbf{z}, \Theta|\mathcal{Z})$. There are several ways to constrain the joint distribution and reduce degrees of freedom: ranging from identifying conditional independence and encoding it into a graph structure (see Section 2.4.2), to parametrically constraining the distribution by assuming a prior distribution over parameters and hyper-parameters (see Section 2.4.3). Generative

[14] The two competing formalisms have also been called discriminative versus informative approaches [177].

models handle classification or regression by manipulating variables using standard basic axioms of probability using marginalization, conditioning and Bayes rules. Generative models can be viewed as one extreme of learning spectrum that attempt to estimate a distribution over all variables ($\mathbf{x}$ and y). Although the completeness of generative models might be appealing, they might be inefficient particularly if the conditional distribution of output given input is needed (*i.e.,* $\mathbb{P}(y|\mathbf{x})$). Thus, in this case, more reductionist approaches such as conditional learning[15] or even more minimalist methods as discriminative learning may be more appropriate.

Unlike generative methods, discriminative approaches do not attempt explicitly to model the underlying distribution between input ($\mathbf{x}$) and output (y). Instead, they focus on finding a mapping from inputs to output (*e.g.,* given features, finding optimal classifier) [177], [119]. Thus, such techniques only consider distance from a decision boundary or goodness of approximation of the regression function as evaluation measures to find optimal parameters. Since discriminative methods do not spend computational resources on the intermediate steps of computing conditional distributions and such, they are potentially more efficient. Examples of discriminative methods are Support Vector Machine (SVM) [204], Gaussian Processes [88], logistic regression [122]. The discriminative models usually lack the elegant probabilistic concept of priors, structure and such concepts are usually replaced with the notions of regularization and loss function [119]. Therefore, it is difficult to incorporate our prior knowledge into such methods and they are not as explicit or visualizable as generative models [119].

It motivates fusing flexibility and diversity of the generative models with efficiency and power of discriminative framework. Several approaches have been proposed to combine these two frameworks. One technique is to combine generative modeling with a subsequent SVM classifier using Fisher kernels [117]. However in such an approach, there is no iteration between generative and discriminative phases and there is a chance that discriminative information is

[15] Unlike Generative methods that learn $\mathbb{P}(\mathbf{x}, y)$, conditional methods focus on $\mathbb{P}(y|\mathbf{x})$ assuming that one is only interested in input ($\mathbf{x}$) and output (y) relationship. They are potentially more efficient than generative settings. This is not quite discriminative learning because we have a generative model of $\mathbb{P}(y|\mathbf{x})$ as compared to a discriminative setting in which we are only interested in a mapping from $\mathbf{x}$ to y [119].

lost during generative modeling. Another approach is proposed by Jebara *et al.* [116], [119] so-called "maximum entropy discrimination (MED)". The model that we proposed in this thesis ([20], [17], [21] and [19]) is rather different. It learns two maximum posteriori models for generative and discriminative models simultaneously.

2.4.2 Graphical Model: An Approach to Model a Distribution

Graphical models have been a very popular tool over the last decade [165], [53], [140], [131]. They allow to handle complicated dependencies between variables of a multi-variate distribution using a graph representation. Nodes of the graph represent random variables and edges between nodes symbolize dependencies between the variables. Exploiting independence between variables allows compact representation. The graph provides a modeling language to incorporate those independencies. Graphical model is a wide research topic on its own and here we only provide very brief introduction to some notions we have used in this thesis; for more in depth introduction see [131], [140].

In a graphical model, some nodes correspond to the observed variables and others denote latent (hidden) missing variables. There might be also nodes representing parameters (Θ) and hyper-parameters[16]. As a convention in this thesis, we represent observed variables with gray circles and latent ones with white circles (see Figure 2.8a,2.8c,2.8d for examples). The two most common classes of graphical models are Bayesian Networks (BN) which are based on directed acyclic graphs (DAG) and Markov networks which are based on undirected graph[17].

Let us assume that we have defined a Bayesian network with a DAG ($\mathcal{G}$) on D variables, $[x_1, \cdots, x_D] = \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^D$, the distribution over all variables in BN can be factorized as a

[16] Hyper-parameters describe distributions over parameters.

[17] It is less common to use a mixed directed and undirected graph.

product

$$\mathbb{P}(\mathbf{x}) = \mathbb{P}(x_1, \cdots, x_D) = \prod_{i}^{D} \mathbb{P}(x_i|\pi(x_i)) \tag{2.4.1}$$

where $\mathbb{P}(x_i|\pi(x_i))$ is the conditional probability of x_i conditioned on its parents nodes $\pi(x_i)$[18]. Examples of BN are shown in Figure 2.8a,2.8c.

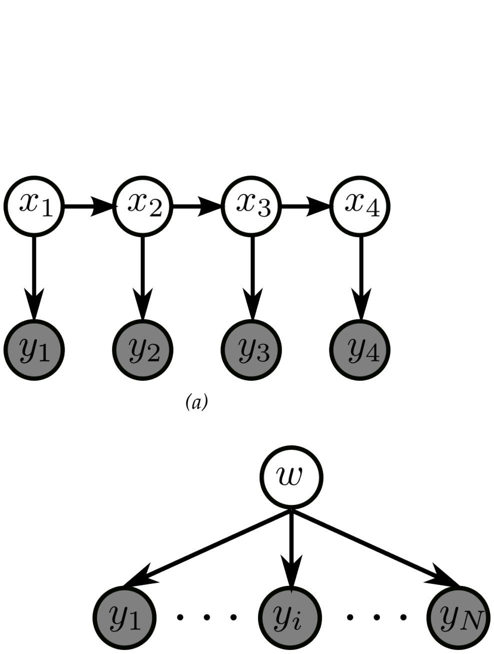
For the Markovian network (*i.e.,* undirected edges and cycle is allowed), the distribution can be factorized according to the product of non-negative potential functions:

$$\mathbb{P}(\mathbf{x}) = \mathbb{P}(x_1, \cdots, x_D) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi(x_C), \quad Z = \int_{\mathcal{X}} \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi(x_C) \tag{2.4.2}$$

where $\mathcal{C}$ is the largest set of fully connected sub-graph (maximal cliques), Z is just a normalizer to produce a proper distribution, and x_C is the set of all random variables in a clique C (see Figure 2.8b for an example). In order to apply a graphical model, one needs to know how to perform *Learning* and *Inference* algorithms over the graph. Giving a comprehensive survey over learning and inference algorithms is beyond scope of this chapter. Here, we provide very brief explanation for each.

Inference is about computing *queries* from the model. Both directed and undirected graphs are full joint probability of all variables. However, one might want to have a specific query from the model. The most common queries are *conditional probability query* and *most probable query*. In *conditional probability query*, we have some observations over a subset of random variables and we would like to compute the conditional probability over another set of variables, namely $\mathbb{P}(x_{\mathcal{C}_1}|x_{\mathcal{C}_2} = z)$ where $\mathcal{C}_2$ is the set of observed group, z is the observed value, and $\mathcal{C}_1$ is the set of variables we are interested in. In "most probable query", we are interested in finding the most probable value given an observation. An obvious example of such query is *maximum a posterior* (MAP) which is mentioned earlier in Section 2.4.1, namely $\arg\max_{x_{\mathcal{C}_1}} \mathbb{P}(x_{\mathcal{C}_1}|x_{\mathcal{C}_2} = z)$. Computing

[18] Remember that $\mathcal{G}$ is a directed acyclic graph



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

Figure 2.8: This figure shows a few examples of graphical models. Figures (a), (c), and (d) are examples of Bayesian Network constructed with a Directed Acyclic Graph (DAG); more specifically (a) represents a Hidden Markov Model (HMM) [169]. (c) and (d) are equivalent, (d) is more compact representation; the box in (c) denotes repetition of N variables. (b) represents an example of Markov network constructed with an undirected graph. All gray nodes (y_i's) are observed variables and the white notes are the latent variables.

an exact inference for a general graph is intractable for large number of models; for this reason, we resort to approximations. In general, there are two frameworks for approximate inference: optimization-based and sampling-based. In optimization-based approach, a class of "easy" distributions is defined, and then the objective of the optimization is to best approximate the query within that class. KL-divergence[19] is usually used to measure distance between distributions. In sampling-based algorithms, the joint distribution is approximated as a set of instantiations to all or some of the variables in the graph. The instantiations (*i.e.,* samples) represent part of the probability mass. The query function can usually be presented as an expectation. The approximation is done via generating M samples[20] and computing empirical expectation (see [131] for more in

[19] Recall that relative entropy between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as $\mathcal{D}(\mathbb{P}_2|\mathbb{P}_2) = \mathbb{E}_{\mathbb{P}_1}[\ln \frac{\mathbb{P}_1(\mathbf{x})}{\mathbb{P}_2(\mathbf{x})}]$

[20] For example Markov chain Monte Carlo (MCMC) is an approach for generating samples from the posterior distribution.

depth discussion).

Learning in graphical models includes two aspects: parameter estimation and structure learning. In parameter estimation, it is assumed that general structure of the graph is given (*i.e.*, dependencies between variables) and the task is to find the parameters given a training data $\mathcal{Z}$, In structure estimation, the objective is to extract both structure as well as parameters of Bayesian network or Markov network given the training data. In this thesis, whenever we use a graphical model, the structure is given and rationalized through a few arguments; see [131] for discussion about structure learning in graphical models. Parameter estimation can be done with maximum likelihood estimation (MLE) or Bayesian approaches. The difference between the two approaches is that in Bayesian approach, a prior distribution is assumed over parameters to improve robustness against over-fitting. Nevertheless, the key ingredient for both is the likelihood function: the probability of the data given the model. Assuming that there are m independent training samples, MLE maximizes $J(\Theta; \mathcal{Z}) = \prod_{i=1}^{m} \mathbb{P}(\mathbf{z}_i|\Theta)$ and Bayesian objective is to maximize $\mathbb{P}(\Theta) \prod_{i=1}^{m} \mathbb{P}(\mathbf{z}_i|\Theta)$. The factorization formulations in *Eq.*2.4.1 and *Eq.*2.4.2 can now be exploited to decompose $\mathbb{P}(\mathbf{z}_i|\Theta)$ further. While estimation of the parameters in BN can be solved efficiently thanks to decomposability of parents and children random variables in *Eq.*2.4.1, estimation of parameters in Markov network usually involves iterative inference and local parameter estimation; therefore it is more expensive than parameter estimation in BN (see [131] for more details).

2.4.3 Generative Model: Matrix Factorization

"A" generative approach to model for high dimensional samples ($\mathbf{x}_i \in \mathbb{R}^D$) is to arrange them as columns (or rows) of a matrix (say $\mathbf{X} \in \mathbb{R}^{D \times N}$) and derive a decomposition as an abstract summarization of the data. In the mathematical discipline of linear algebra, matrix factorization is decomposition of a matrix to a canonical form, *e.g.*,

$$\mathbf{X} = \mathbf{BC} + \mathbf{E}$$

$$\mathbf{B} \in \mathcal{B}, \quad \mathbf{C} \in \mathcal{C}, \quad \mathbf{E} \in \mathcal{E} \tag{2.4.3}$$

where $\mathcal{B}, \mathcal{C}$, and $\mathcal{E}$ denote the sets of feasible choices for $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{E}$ respectively. Different choices for $\mathcal{B}, \mathcal{C}$, and $\mathcal{E}$ yield various decomposition methods. For example assuming $\mathbf{E} = \mathbf{0}$ (*i.e., Eq.*2.4.3 is an exact decomposition): 1) if $\mathcal{B}$ is the set of all "orthogonal" matrices and $\mathcal{C}$ is the set of all "orthonormal" ones, *Eq.*2.4.3 is the Singular Value Decomposition (SVD) method[21]; 2) If $\mathcal{B}$ is set of lower triangular matrices and $\mathcal{C}$ is set of upper triangular ones, then *Eq.*2.4.3 is *LU* decomposition; *etc.*. There are many flavors of matrix factorization and here we only focus on low-rank matrix approximation. By low-rank matrix approximation, we mean: $rank(\mathbf{X}) > rank(\mathbf{BC})$ and $\mathbf{E}$ denotes error or noise matrix entries of which should be close to zero; hence $\mathbf{X} \approx \mathbf{BC}$. We need a measure of distance ($\mathcal{D}(\cdot;\cdot)$) (a divergence) to measure the quality of the approximation; therefore *Eq.*2.4.3 can be written as an optimization problem

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{C}} \qquad \mathcal{D}(\mathbf{X};\mathbf{BC}) \\ &\text{subject to:} \mathbf{B} \in \mathcal{B}, \quad \mathbf{C} \in \mathcal{C} \end{aligned} \tag{2.4.4}$$

Here we show a few examples of popular algorithms that can be cast out as low-rank matrix approximation. Most of the dictionary learning methods can be viewed as variations of *Eq.*2.4.4, k-SVD [7], Non-negative Matrix Factorization [141], Independent Component Analysis (ICA) [25], *etc.* [75], [190]. Table 2.2 represents some other examples of popular methods that can be described by $\mathbf{X} \approx \mathbf{BC}$ (for more examples see [190]). Just for illustration purposes, we derive a matrix factorization for k-means clustering which is widely known as a straightforward and fairly efficient method for solving unsupervised learning problems:

Example 1: $k-$means clustering is a method of cluster analysis which aims to partition N observations into K clusters, in which each observation belongs to the cluster with the nearest

[21] Typically SVD is represented as $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{D\times r}$ and $\mathbf{V} \in \mathbb{R}^{N\times r}$ are orthonormal matrices and Σ is a $r \times r$ diagonal matrix with positive diagonal entries. Σ can be absorbed into $\mathbf{U}$ which make it just orthogonal.

Table 2.2: This table shows examples of well-known methods that can be viewed as matrix factorization: Singular Value Decomposition (SVD), k-means/medians, Probabilistic Latent Semantic Indexing (pLSI), Non-negative Matrix Factorization (NMF). In the table, $\|\cdot\|_F^2$ denotes Frobenius norm and Λ is a diagonal matrix and KL denotes *Kullback-Leibler* divergence [65].

Method	$\mathcal{D}(\mathbf{X};\mathbf{BC})$	$\mathcal{B}$	$\mathcal{C}$
SVD	$\|\mathbf{X}-\mathbf{BC}\|_F^2$	$\mathbf{B}^T\mathbf{B}=\mathbf{I}$	$\mathbf{CC}^I=\Lambda$
k-means	$\|\mathbf{X}-\mathbf{BC}\|_F^2$	-	$\mathbf{CC}^T=\mathbf{I}$, $c_{ij}=\{0,1\}$
k-medians	$\|\mathbf{X}-\mathbf{BC}\|_1$	-	$\mathbf{CC}^T=\mathbf{I}$, $c_{ij}=\{0,1\}$
pLSI [109]	$KL(\mathbf{X};\mathbf{BC})$	$\mathbf{1}^T\mathbf{B}\mathbf{1}=1$ $b_{ij}\geq 0$	$\mathbf{1}^T\mathbf{C}=\mathbf{1}$ $c_{ij}\geq 0$
NMF [141]	$KL(\mathbf{X};\mathbf{BC})$	$b_{ij}\geq 0$	$c_{ij}\geq 0$

mean. Difference between the $k-$means algorithm and its soft version is that the variable describing how data points belong to clusters takes "degree" values instead of binary (0 and 1) values. Assuming that each of the N observations ($\mathbf{x}_i$) belongs to a D-dimensional feature space ($\mathbf{x}_i \in \mathbb{R}^D$):

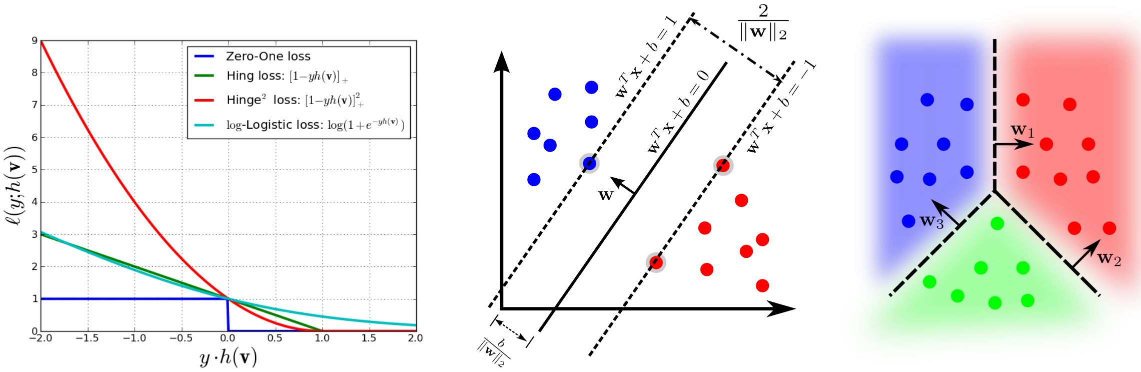
$$\begin{array}{ll}
\text{hard } k\text{-means:} & \text{soft } k\text{-means:} \\
\min_{c_{ki},\mathbf{b}_k} \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_{k=1}^{K} \mathbf{b}_k c_{ki}\|_2^2 & \min_{c_{ki},\mathbf{b}_k} \sum_{i=1}^{N} \|\mathbf{x}_i - \sum_{k=1}^{K} \mathbf{b}_k c_{ki}\|_2^2 \\
\text{s.t.: } \sum_{k=1}^{K} c_{ki} = 1, \quad c_{ki} \in \{0,1\} & \text{s.t.: } \sum_{k=1}^{K} c_{ki} = 1, \quad c_{ki} \geq 0
\end{array} \tag{2.4.5}$$

where $\mathbf{b}_k$ are the centroids of the clusters, c_{ki} are cluster membership values. Because of the constraint, $\{c_{ki}\}_{k=1}^{K}$ can be viewed as the probability or membership values.

Alternatively, one can view *Eq.*2.4.5 as a constrained matrix factorization problem:

$$\begin{array}{ll}
\min_{\mathbf{C},\mathbf{B}} & \|\mathbf{X}-\mathbf{BC}\|_F^2 \\
\text{subject to} & \mathbf{C} \in \mathcal{C}
\end{array} \tag{2.4.6}$$

where $\mathcal{C} := \{\mathbf{c}_k : \mathbf{c}_k \geq 0, \quad \mathbf{1}^T\mathbf{c}_k = 1, \quad 1 \leq k \leq K\}$ for soft $k-$means and $\mathcal{C} \in \{0,1\}^{K\times N}$ for hard $k-$means; $\mathbf{X} \in \mathbb{R}^{D\times N}$ is matrix holding the observations; each column of the $\mathbf{X}$ is a sample.



(a) *(b)* *(c)*

Figure 2.9: (a) shows some of common choices for the discriminative loss function. Notice that *zero-one* loss function is a sign function. (b) shows maximum margin hyperplane and margins for an SVM trained with samples of the two classes. (c) shows an example of loss function for multi-class classification.

Similarly, the columns of $\mathbf{B} \in \mathbb{R}^{D\times K}$ are cluster centroids and the columns of $\mathbf{C} \in \mathbb{R}^{K\times N}$ hold the membership values. For brevity of notation, $\mathcal{C}$ encodes the feasible set for the columns of $\mathbf{C}$ that was shown earlier in *Eq.*2.4.5; $\mathbf{c}_k$ are columns of the matrix $\mathbf{C}$. In matrix nomenclature, $\mathbf{B}$ and $\mathbf{C}$ can be called *basis matrix* and *coefficient matrix* respectively. Notice that from the matrix factorization point of view, *Eq.*2.4.6 clusters the columns of $\mathbf{X}$ and the constraints are defined on the columns on $\mathbf{C}$. If the constraint is defined on rows of $\mathbf{B}$ instead, the matrix factorization clusters the rows of the $\mathbf{X}$ instead of the columns, and the rows of $\mathbf{C}$ play the role of centroids while the rows of $\mathbf{B}$ hold membership values.

2.4.4 Discriminative Model: Support Vector Machine

One of the most popular discrimination methods is Support Vector Machine (SVM) [204]. SVM is a minimalist (non-probabilistic, see Section 2.4.1) classifier that maps given input features to a class label. It can be used for classification, regression, or other tasks. Intuitively, a good separation in a classification problem is achieved by the hyperplane that has the largest distance to the nearest training data points of any class; therefore, the larger the margin the lower the generalization error of the classifier (see Figure 2.9b).

For illustration, we study linear binary classification. Let us assume we are given some training data $\mathcal{Z} = \{\mathbf{z}_i = (\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^D, y_i \in \{-1, 1\}\}_{i=1}^N$ where $\mathbf{x}_i$ are features and y_i are the class labels. The objective is to find a linear classifier parametrized by $\mathbf{w}$ and b ($h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$) with the least number of mis-classification. We can define a so-called a *loss* function to measure goodness of fit between $h(\mathbf{x}_i)$ and corresponding y_i, *i.e.,* $\ell(y_i; h(\mathbf{x}_i))$:

$$\min_{\mathbf{w},b} \frac{1}{N} \sum_{i=1}^{N} \ell(y_i; h(\mathbf{x}_i)) \tag{2.4.7}$$

For a binary classification, the loss can be written as a function number of disagreement between y_i and $h((\mathbf{x}_i))$, *i.e.,* $\ell(y_i; h(\mathbf{x}_i)) = \ell(y_i h(\mathbf{x}_i))$. Several possible choices for the loss function are given in Figure 2.9a. The loss function that actually counts the number of disagreements is the *sign* function but it renders *Eq.*2.4.7 computationally inefficient even for small number of samples[22]. Other choices for the loss function in Figure 2.9a approximate the sign function. They inflict penalties that are bigger or equal to the sign function (they upper bound the sign function; see Figure 2.9a). Let us study one of the surrogates which is called *hinge* loss function, namely $\ell(y_i; h(\mathbf{x}_i)) = [1 - y_i h(\mathbf{x}_i)]_+$. Choosing the hinge loss function, we can write the SVM optimization problem as:

$$\min_{\mathbf{w},b} \frac{1}{N} \sum_{i=1}^{N} [1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)]_+ + \lambda \|\mathbf{w}\|_2^2 \tag{2.4.8}$$

where $[a]_+ = \max\{0, a\}$ and $\|\mathbf{w}\|_2^2$ is a regularizer added to the objective function to improve the generalization and λ balances between the loss function and the regularization. To see why $\|\mathbf{w}\|_2^2$ improves generalization, notice that $\frac{b}{\|\mathbf{w}\|_2}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$ and we want to choose $\mathbf{w}$ and b to maximize the margin line (distance between the parallel hyperplanes) that are as far apart as possible while still separating

[22] If the sign function is chosen as the loss function, one may resort to algorithms to solve the mixed integer programming in *Eq.*2.4.7(*e.g.,* Cutting-Plane [120]), but the integer programming algorithms do not scale well to medium- or large-scale optimization problems.

the data. If two classes are separable, the first term in *Eq.*2.4.8 is zero and the optimization in *Eq.*2.4.8 minimizes $\|\mathbf{w}\|_2$ which is equivalent to maximizing the margin (see Figure 2.9b).

SVM can be extended beyond the binary classification. For multi-class case (*i.e.*, $y_i \in \{1, 2, \cdots, L\}$), the most common approach is to reduce the problem into several binary classifications (*e.g.*, *one-vs-one* or *one-vs-all* [70], [110]). Nevertheless, we prefer the method proposed by Crammer *et al.* [55] that casts multi-class classification into a single optimization rather than multiple binary ones. In [55], for each class (say l'th class) there is a set of parameters (say $\mathbf{w}_l$); therefore instead of a vector parameterizing a classifier, we have a matrix $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_L] \in \mathbb{R}^{D \times L}$ holding parameters of a multi-class classifier. The class label is decided by finding the maximum value classifier, namely $h(\mathbf{x}) = \arg\max_{l=1,\cdots,L} \{\mathbf{w}_l^T \mathbf{x} + b_l\}$ (see Figure 2.9c). Before casting it as a single optimization, let us re-write *Eq.*2.4.8 slightly differently:

$$\begin{aligned} &\min_{\xi_i, b, \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \xi_i + \lambda \|\mathbf{w}\|_2^2 \\ &\text{s.t.:} \;\; 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{2.4.9}$$

*Eq.*2.4.9 and *Eq.*2.4.8 are equivalent. In effect, ξ_i's account for samples that are on the wrong side of the separating hyperplane (miss-classification) or within the margin (sample classified correctly but fell within the margin area). Crammer *et al.* [55] suggest to modify *Eq.*2.4.9 with:

$$\begin{aligned} &\min_{\xi_i, b, \mathbf{W}} \frac{1}{N} \sum_{i=1}^{N} \xi_i + \lambda \|\mathbf{W}\|_F^2 \\ &\text{s.t.:} \quad 1 + \mathbf{w}_l^T \mathbf{x}_i + b_i - \mathbf{w}_{y_i}^T \mathbf{x}_i - b_{y_i} \leq \delta_{il} + \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{2.4.10}$$

δ_{il} is 1 if $i = l$ and 0 otherwise. $\mathbf{w}_l$ and $\mathbf{w}_{y_i}$ are the l and y_i'th columns of $\mathbf{W}$. If there are N samples and L class label, the number of constraints on *Eq.*2.4.10 in $N \times L$. In *Eq.*2.4.10, if i'th sample is classified as $\hat{y}_i \neq y_i$, it means $\mathbf{w}_{\hat{y}_i}^T \mathbf{x}_i + b_{\hat{y}_i} > \mathbf{w}_{y_i}^T \mathbf{x}_i + b_{y_i}$, hence $\xi_i = 1 + \mathbf{w}_{\hat{y}_i}^T \mathbf{x}_i + b_{\hat{y}_i} - \mathbf{w}_{y_i}^T \mathbf{x}_i - b_{y_i}$ is inflicted to the objective of *Eq.*2.4.10.

2.5 Toward the Proposed Method: Generative-Discriminative Learning

One of the aims of the proposed method is classification; this problem falls into the discriminative learning paradigm. In addition to the generalization performance in term of classification, we desire a method that is expressive for clinical purposes. Expressiveness serves two goals: 1) it allows us to validate and compare what is found to be important by our model to that of clinical knowledge[23] thus this qualitative measure can be used in tandem with the quantitative measure (*i.e.,* classification accuracy). 2) it allows clinical knowledge to be incorporated into the model as a prior[24]; this prior knowledge can be instrumental to alleviate the curse of dimensionality of the original problem. A generative framework (*e.g.,* Bayesian) is more appropriate to satisfy the "expressiveness" criterion. In this thesis, we combine those two learning paradigms, generative and discriminative, and address related challenges for medical image classification applications.

One of the fundamental limitations in medical image classification is the lack of sufficient training samples relative to the high dimensionality of the data. Therefore, a dimensionality reduction is required to achieve a good generalization performance for the classification task. We adopted a matrix factorization framework to reduce the dimensionality in both the discriminative and the generative tasks. Section 2.5.1 provides a simple illustrative example showing a generative application of matrix factorization for a medical imaging task. In light of that example, Section 2.5.2, briefly discusses the general idea behind the proposed method.

2.5.1 An Illustrative Example

In this example, let us assume that we are given an ensemble of N brain images with a common pathology. All brain images are *registered* to a common template. Let us assume that the pathol-

[23] For example, for some abnormalities such as AD, areas related to memory are usually affected.

[24] For example, a pathology may only affect the gray-matter part of brain

ogy we are interested in, affects similar areas in patients' brains[25] and we are asked to segment regions where the pathology is mostly observed. Since we may not have a priori information about the location of the pathology, an appropriate solution would be to apply a data-driven approach, for example $k-$means clustering, to segment voxels of the same type. *Signatures* of voxels across subjects [26] can be used as features. Recall **Example 1** in Section 2.4.3, we can cast the $k-$means algorithm as matrix factorization. Let us assume that each column of matrix $\mathbf{X}$ holds all voxels of a subject. The only difference is that we want to cluster the voxels but not the subjects. We can simply transpose data matrix ($\mathbf{X}$) and try to factorize it into centroids and membership values. However, for the sake of this example, we do not transpose the data matrix. Instead, we change our interpretation of the basis and the coefficient matrices by moving the constraint which was defined earlier on $\mathbf{C}$ in *Eq.*2.4.6 to $\mathbf{B}$. The idea is shown in Figure 2.10. New formulation would be as follows:

$$\begin{aligned} \min_{\mathbf{C},\mathbf{B}} \quad & \|\mathbf{X} - \mathbf{BC}\|_F^2 \\ \text{subject to} \quad & \mathbf{B} \in \mathcal{B} := \{\mathbf{b}^d : \mathbf{b}^d \geq 0, \quad \mathbf{1}^T \mathbf{b}^d = 1, \quad 1 \leq d \leq D\} \end{aligned} \tag{2.5.1}$$

in which $\mathbf{b}^d$ denotes rows of the basis matrix. Notice that in *Eq.*2.5.1, each row of the $\mathbf{X}$ is one sample while in *Eq.*2.4.6, each column is one sample; in other words, columns of $\mathbf{X}$ in *Eq.*2.5.1 index features while rows of $\mathbf{X}$ denote features in *Eq.*2.4.6. Therefore, the constraint in *Eq.*2.4.6 moves from columns of the $\mathbf{C}$ to rows of $\mathbf{B}$. In *Eq.*2.5.1, rows of $\mathbf{B}$ are membership values. In fact the formulation in *Eq.*2.5.1 segments or clusters voxels into groups in a generative (unsupervised) way. Observe that changing the feasible set changes the meaning of the algorithm and consequently alters the roles of its blocks ($\mathbf{B}$ and $\mathbf{C}$).

[25] An example of such a pathology is vascular lesions in elderly patients which occur mostly around ventricles.

[26] Here by *signature*, we mean features extracted for each voxel; for example, simply intensity of the voxel. By signatures of the voxel *across subjects*, we mean, for each voxel location, features extracted from every subject are concatenated into a vector to build a feature vector for that voxel location. We call this feature vector, signature of the voxel across subjects for that location.

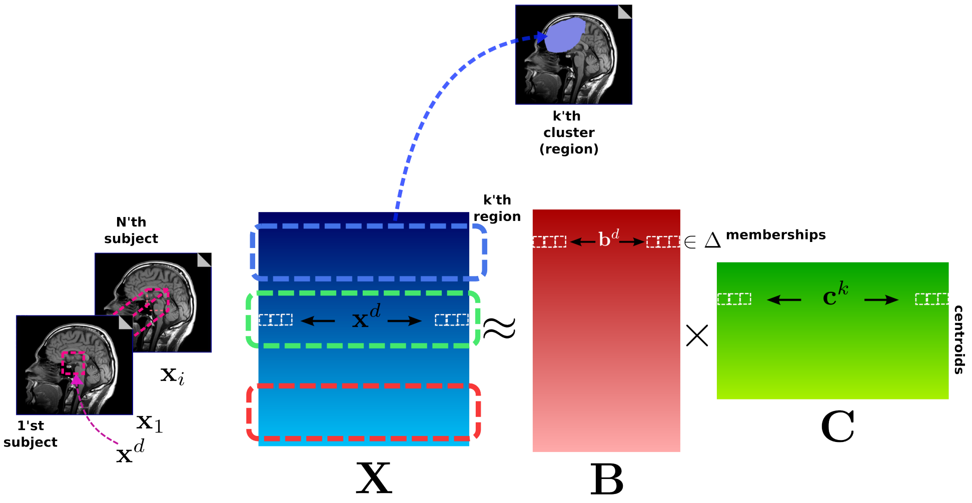


Figure 2.10: In this figure, each row in the data matrix ($\mathbf{X} \in \mathbb{R}^{D \times N}$) is one observation which is an N dimensional row vector representing the signature of a voxel across N subjects. This figure shows how $k-$means clustering of the voxels can be viewed as matrix factorization. Therefore, the rows of $\mathbf{C}$ are centroids of the clusters and rows of $\mathbf{B}$ are membership values. Hence, the rows of $\mathbf{B}$ (*i.e.,* $\mathbf{b}^k$) belong to K-dimensional simplex ($\mathbf{b}^k \in \Delta$), *i.e.,* it is positive and sums to one. This algorithm clusters rows of the $\mathbf{X}$ into clusters (regions).

2.5.2 Merge Some Rows, Classify All Columns

As we saw in the previous example, matrix factorization can be used to formulate a data-driven approach for pathology localization. However, the problem we address in this thesis is more challenging than clustering or segmentation:

- Not only we are interested in localization but also we would like to find regions that can be used to extract discriminative features.
- Not all voxels (*i.e.,* rows of the data matrix) are relevant to the abnormality (*e.g.,* background voxels). Therefore, it calls for a different formulation than clustering or segmentation (*i.e.,* set partitioning); the problem rather falls into the category of subset selection. It renders the simplex constraint on rows of the $\mathbf{B}$ irrelevant (see Chapter 3 for details).

Nevertheless, the example in the previous section can give us an insight. In effect, we would like to *merge some of the rows (voxels)* but not all them because only some are relevant.

Notice the formulation in *Eq.*2.5.1, leads to cluster the rows. If the simplex constraint is defined on the columns of the $\mathbf{C}$ instead of row of $\mathbf{B}$, it leads to cluster subjects (columns) into homogeneous groups. However, we would like our cluster assignment to be consistent with the class labels. In other words, we would like subjects of the same class to be clustered together. Therefore from this point of view, it is a classification problem not unsupervised clustering. More specifically, we adopt a discriminative approach rather than the generative approach and we would like to make as few misclassification as possible. Consequently, since our objective is different than that of the first example, we do not use the definition of the feasible set in *Eq.*2.4.6. Feasible sets will be discussed in Section 3.4 and more elaborately in Chapter 4.

In a similar theme to this thesis, various formulations of matrix factorization have already been proposed for collaborative filtering [132], [181], [231], and [173], multi-way clustering [185], [74]. However, they serve different purposes than what we are interested in this thesis. The collaborative filtering method has mostly been used for recommendation system, for example, to recommend movies to customers[27] or people with similar interests in dating websites[28] [132]. In recommendation systems, the objective is to fill unknown entries of a matrix and the low-rank assumption ($rank(\mathbf{X}) < \min(D, K)$) is imposed for regularization purposes or to improve robustness of the algorithm. In some of them [193], there is no direct access to the basis and coefficient matrix but they are implicitly regularized.

In the context of dictionary learning, there has been similar matrix factorization problems. Duarte *et al.* [71] have learned dictionaries for compressive sensing purpose. Most of the dictionary learning methods [98], [76], [152] are used for denoising or restoration of signals or images. Some authors [170], [98], [228] used learned dictionaries for classification tasks but atoms of the dictionaries were not learned for the specific task (*i.e.*, classification). In addition, it is not proper for our purpose because: first, they mostly focus on the large-sample size problem while in our case, number of samples is much smaller than the dimensionality; second, the resultant dictio-

[27] Netflix: www.netflix.com

[28] eHarmony: www.eharmony.com

nary atoms cannot be used to illustrate differences between the classes.

To the best of our knowledge, the closest publication that exploits matrix factorization with a relatively similar approach is a paper by Mairel *et al.* [151]. Nevertheless, there are significant differences between our formulation and that of Mariel both in the term of the objective function, the constraints and the optimization approach. In addition, the main goal in [151] is to apply the method in the cases that the number of samples is large; while in medical imaging applications the number of features are much larger than the number of samples.

Part I

General Framework

Chapter 3

Generative Discriminative Matrix Factorization

3.1 General Framework

We adopt a regularized matrix factorization framework for our purposes. In regularized matrix factorization, the objective is to decompose a matrix into two or more matrices subjected to some constraints or priors such that the decomposition describes the matrix as accurately as possible. Assuming that each column of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_i \cdots \mathbf{x}_N]$ represents an observation (*i.e.,* a sample image that is vectorized), the columns of matrix $\mathbf{B}$ can be viewed as basis vectors and the i'th column of $\mathbf{C}$ contains corresponding loading coefficients of the basis vectors for the i'th observation:

$$\mathbf{X} \approx \mathbf{BC} \qquad \mathbf{B} \in \mathcal{B}, \quad \mathbf{C} \in \mathcal{C}, \tag{3.1.1}$$

in which $\mathbf{X}$ is decomposed into two matrices $\mathbf{B}$ and $\mathbf{C}$, each of which has its own feasible set, $\mathcal{B}$ and $\mathcal{C}$ respectively.

In order to define the feasible sets in *Eq.*(3.1.1), we need to elaborate the requirements that our model should satisfy:

1) The basis vectors must be anatomically meaningful. This means that a constructed basis vector should correspond to contiguous anatomical regions preferably in areas which are biologically related to a pathology of interest. Remember our example in Section 2.5.1 where we wanted to cluster voxels into regions. There are similarities between our formulation and that of the second example in a sense that we want to have local spatial support; nevertheless we use different constraints than those of *Eq*.2.5.1. Constraints in *Eq*.2.5.1 enforce voxels to be exclusively member of one cluster. This enforcement is applied hardly or softly depending on whether hard or soft k-mean is used. However in our application, there might be two or more overlapping clusters of voxels that are relevant to an abnormality. In addition, parts of an image may not be relevant to the abnormality at all (*e.g.,* background). Therefore, we would like to allow each voxel to belong to none or more than one cluster of voxels.

We do not use the same definition of the feasible set as *Eq*.2.5.1 but we still want to have local spatial support.

2) The basis must be discriminative: we are interested in finding features, *i.e.,* projections onto the basis vectors, that construct spatial patterns which best differentiate between groups, *e.g.,* patients and controls or activation and baseline.

3) The basis vectors must be representative of the data as much as possible, while maintaining their discriminatory ability.

In subsequent sections, we will introduce appropriate priors that encourage the aforementioned properties, but we first lay out our framework. This framework is represented pictorially in Figure3.2 and as a graphical model in Figure3.1. Let us assume that we collect images into columns of matrix $\mathbf{X}$, therefore a column $\mathbf{x}_i$ represents one sample image whose label (class) is represented by y_i. Entries of each column of $\mathbf{X}$ ($\mathbf{x}_i$) are image features based on which we can define regions. For example, it can be the determinant of Jacobian of a deformation field that warps a subject to a common template (see Section 3.6), a tissue density map representing region

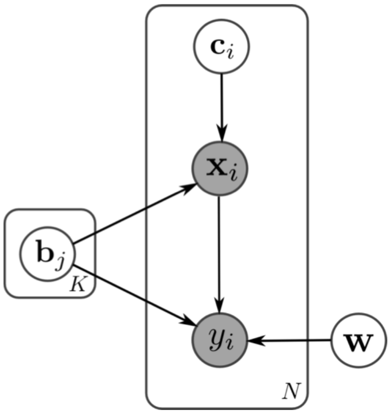


Figure 3.1: Graphical model representing our model: $\mathbf{x}_i$ is the i'th sample (out of N samples) and y_i is the corresponding class label. $\mathbf{b}_j$ is the j'th basis vector (out of K basis vectors) and $\mathbf{c}_i$ is the loading coefficient for the i'th sample; $\mathbf{w}$ parametrizes the class-likelihood, *i.e.*, $p_{\mathbf{w}}(\mathbf{y}|\cdot)$; in other words, it parametrizes the classifier. Since samples and corresponding labels are observed variables, they are shaded with gray while unobserved variables (*i.e.*, $\mathbf{b}_j$, $\mathbf{c}_i$, and $\mathbf{w}$) are white.

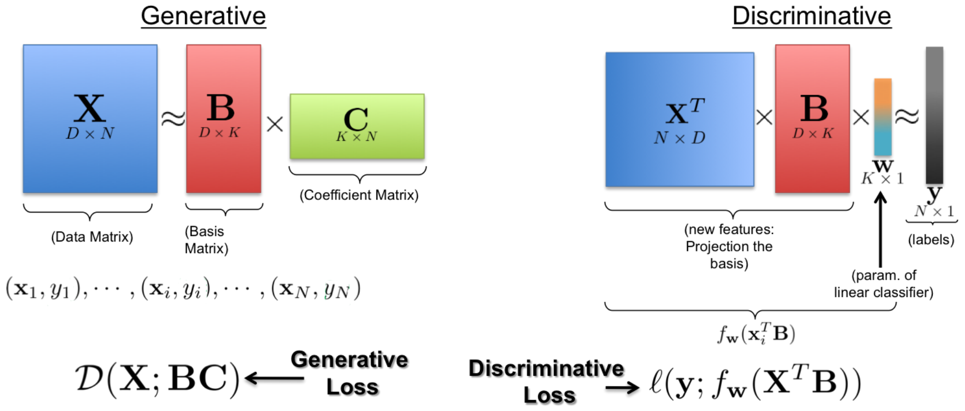

Generative
$\mathbf{X}$ $D \times N$
$\approx$
$\mathbf{B}$ $D \times K$
$\times$
$\mathbf{C}$ $K \times N$
(Coefficient Matrix)
(Data Matrix)
(Basis Matrix)
$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_i, y_i), \cdots, (\mathbf{x}_N, y_N)$
Discriminative
$\mathbf{X}^T$ $N \times D$
$\mathbf{B}$ $D \times K$
$\mathbf{w}$ $K \times 1$
$\mathbf{y}$ $N \times 1$
(labels)
(new features: Projection the basis)
(param. of linear classifier)
$f_{\mathbf{w}}(\mathbf{x}_i^T \mathbf{B})$
$\mathcal{D}(\mathbf{X}; \mathbf{BC})$
Generative Loss
Discriminative Loss
$\ell(\mathbf{y}; f_{\mathbf{w}}(\mathbf{X}^T \mathbf{B}))$

Figure 3.2: It shows the idea of general framework a matrix factorization. The objective function consists of *Generative* term that approximate $\mathbf{X} \approx \mathbf{BC}$; and *Discriminative* term that approximate labels ($\mathbf{y}$). In order to measure goodness of fit for each term, we define the generative loss function $\mathcal{D}(\cdot;\cdot)$ and discriminative loss $\ell(\cdot;\cdot)$.

volume (see [90] and [61]), or fMRI of an activation task.

Assuming that each image consists of D voxels that are concatenated in lexicographical order, each column of $\mathbf{X}$ is a D-dimensional vector. If the dataset includes N samples, matrix $\mathbf{X}$ is a $D \times N$ matrix. In this part of the thesis, we assume that $\mathbf{x}_i$'s reside in the positive quadrant (in most cases, images, or determinants of Jacobian of diffeomorphic transformations derived from them, are non-negative). The goal is to decompose the data, $\mathbf{X}$, into a matrix $\mathbf{B}$, whose columns are optimized basis vectors, and a loadings matrix $\mathbf{C}$, which holds corresponding loadings of the basis vectors, namely $\mathbf{X} \approx \mathbf{BC}$. At the same time the projections on the basis, $\mathbf{B}$, are used to predict the labels $\mathbf{y}$ using $\mathbf{w}$ as we describe below, thus trading off generative and discriminative criteria. Without additional constraints, the decomposition is ill-posed and has infinitely many solutions; hence regularization is necessary. Given conditional independence depicted in Figure3.1, we formulate the problem as a MAP (Maximum a Posteriori) estimation problem as follows:

$$\begin{aligned}
&\mathbf{b}_k \in \mathbb{R}^D, \qquad \mathbf{c}_i \in \mathbb{R}^K, && \mathbf{w} \in \mathbb{R}^K \qquad (3.1.2)\\
&\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_K], \mathbf{C} = [\mathbf{c}_1 \cdots \mathbf{c}_N], && \mathbf{w} = [w_1 \cdots w_K]\\
&(\mathbf{B}^*, \mathbf{C}^*, \mathbf{w}^*) = \arg\max_{\mathbf{B},\mathbf{C},\mathbf{w}} \log p(\mathbf{B}, \mathbf{C}, \mathbf{w}|\mathbf{X}, \mathbf{y}) = && \arg\max_{\mathbf{B},\mathbf{C},\mathbf{w}} [\log p(\mathbf{X}|\mathbf{B}, \mathbf{C}) + \log p(\mathbf{y}|\mathbf{X}, \mathbf{B}, \mathbf{w})\\
& && + \log p(\mathbf{B}) + \log p(\mathbf{C}) + \log p(\mathbf{w})]
\end{aligned}$$

in which $\mathbf{w}$ is a vector that parametrizes class-likelihood ($p(\mathbf{y}|\mathbf{X}, \mathbf{B}, \mathbf{w})$), or, in other words, it parametrizes a classifier that will be explained later (Section3.3). Instead of maximizing the logarithm of the posterior, we can minimize the negative of the logarithm of the posterior that yields:

$$\begin{aligned}
(\mathbf{B}^*, \mathbf{C}^*, \mathbf{w}^*) = \arg\min_{\mathbf{B},\mathbf{C},\mathbf{w}} & \mathcal{D}(\mathbf{X}; \mathbf{B}, \mathbf{C}) + \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}, \mathbf{w}) + \mathcal{R}(\mathbf{B}, \mathbf{C}, \mathbf{w})\\
\text{subject to:} \qquad & \mathbf{B} \in \mathcal{B} \quad \mathbf{C} \in \mathcal{C} \quad \mathbf{w} \in \mathcal{W}, \qquad (3.1.3)
\end{aligned}$$

in which the first term is a divergence term that encourages good data approximation, which will be referred to as the *generative* term. This idea is represented in Figure3.2. The second term is a loss function that encourages good classification, which will be referred to as the *discriminative* term. The last term in the objective of *Eq.*(3.1.3) is a combination of prior terms on $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{w}$. Due to conditional independence assumed in our model (Figure 3.1), this term can be decomposed into addition of priors over each of the terms. Observe that in *Eq.*(3.1.3), the regularization terms are redundantly added only for future references. This perspective is consistent with *Eq.*(3.1.2) because every constraint can be transformed to a prior by imposing an infinite cost for points outside the feasible set and zero for points inside the feasible set. For example:

$$\mathbf{B} \in \mathcal{B} \quad \equiv \quad \mathcal{R}(\mathbf{B}) = \begin{cases} \infty & \text{if } \mathbf{B} \notin \mathcal{B}; \\ \\ 0 & \text{if } \mathbf{B} \in \mathcal{B}. \end{cases} \tag{3.1.4}$$

Some examples of well-known methods in Table 2.2 that can be viewed as regularized matrix decomposition and can be formulated as *Eq.*(3.1.3). Note that the examples in Table 2.2 are all generative methods, hence $\mathbf{w}$, and consequently its feasible set, $\mathcal{W}$, is omitted.

3.2 Generative Term

In this section, we will explain $\mathcal{D}(.;.)$ (the generative term) that measures the divergence between the data ($\mathbf{X}$) and its decomposition ($\mathbf{BC}$), that is

$$\mathbf{X} = \mathbf{BC} + \mathbf{E}$$

where $\mathbf{E}$ represents approximation error (noise).

Various divergence choices can model different noise assumptions. Basically, any Bregman divergence can be used for $\mathcal{D}(\cdot;\cdot)$:

Definition 3.2.1. For any strictly convex function $\phi : S \subseteq \mathbb{R} \to \mathbb{R}$ that has a continuous first derivative, the corresponding *Bregman* divergence $\mathcal{D}_\phi : S \times int(S) \to \mathbb{R}_+$ is defined as $\mathcal{D}_\phi(x, y) \triangleq \phi(x) - \phi(y) - \nabla\phi(y)(x - y)$, where $int(S)$ is the interior of set S [65].

Bregman divergences are non-negative, convex in the first argument and zero if and only if $x = y$. For matrices, we can define separable Bregman divergences as $\mathcal{D}_\phi(\mathbf{X}, \mathbf{Y}) = \sum_{ij} \mathcal{D}_\phi(x_{ij}, y_{ij})$.

Notice that $x_{ij}, y_{ij} \in dom\phi \cap \mathbb{R}_+$. Different choices of ϕ lead to various divergence terms. For example $\phi(x) = \frac{1}{2}x^2$ yields Frobenius norm and $\phi(x) = x\log(x)$ gives element-wise Kullback-Leibler (KL) divergence.

We assume Gaussian noise between observations ($\mathbf{X}$) and their reconstructions ($\mathbf{BC}$), *i.e.,* $p(\mathbf{X}|\mathbf{B},\mathbf{C}) = \mathcal{N}(\mathbf{BC}, \frac{1}{\lambda_1}\mathbf{I})$. It is shown in [47] that the Frobenius norm is optimal for additive Gaussian noise. Hence, the generative term is:

$$-\log p(\mathbf{X}|\mathbf{B},\mathbf{C}) = \lambda_1 \mathcal{D}(\mathbf{X};\mathbf{B},\mathbf{C}) = \lambda_1 \|\mathbf{X} - \mathbf{BC}\|_F^2 \tag{3.2.1}$$

Observe that the divergence term is a convex function with respect to $\mathbf{B}$ if $\mathbf{C}$ is fixed, and vice-versa, but it is not jointly convex with respect to both $\mathbf{B}$ and $\mathbf{C}$. Other assumptions of noise between observation and reconstruction, *e.g.,* Poisson, can be modeled by various choices for the divergence term, *e.g.,* Kullback-Leibler (KL) divergence [65].

3.3 Discriminative Term

The idea behind the discriminative term is to encourage discriminative basis vectors. In other words, if an image, $\mathbf{x}_i$, is projected on basis vectors yielding new features (*e.g.,* $\mathbf{v}_i$), such new features should be discriminative. In other words, for new features ($\mathbf{v}$), there exists a classifier parametrized by, say $\mathbf{w}$, that minimizes a loss function, $\ell(y_i; h_{\mathbf{w}^*}(\mathbf{v}_i))$, for an optimal set of parameters $\mathbf{w}^*$. Here, we use a linear classifier, namely

$$h_{\mathbf{w}}(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v} \rangle \tag{3.3.1}$$

where $\langle \cdot, \cdot \rangle$ represents inner product and entries of $\mathbf{v}$ are new features after projection.

Ideally, $\mathbf{v}$ can be written as a projection operator acting on $\mathbf{x}_i$ to project it on the subspace spanned by $\mathbf{b}_j$'s. However, in our formulation, we set $v_j = \langle \mathbf{x}, \mathbf{b}_j \rangle$ or, in matrix notation, $\mathbf{v} =$

$\mathbf{B}^T\mathbf{x}$. It is not a proper projection unless the basis vectors are orthonormal; nevertheless, as it will become clear in the next section, due to the positivity constraint and the fact that basis vectors act like indicator functions, $\langle \mathbf{x}, \mathbf{b}_j \rangle$ is proportional to the weighted sum of features in a non-zero area of a basis vector, which is the quantity we are interested in using as new features. There is also two more reasons for defining the classifier as it is:

- Remember our example in Section 2.5.1 in which the objective was to define regions (clusters) on images and rows of the $\mathbf{B}$ were membership values of the clusters. In that context, $v_j = \langle \mathbf{x}, \mathbf{b}_j \rangle = \sum_{d=1}^{D} b_{dj} x_d$ computes weighted average of j'th cluster for a given image ($\mathbf{x}$). In other words, it is one way to extract a feature value for each cluster.
- This formulation allows us to have two different types of features on the generative and discriminative terms. In other words, it is possible to have $v_j = \langle \psi(\mathbf{x}), \mathbf{b}_j \rangle$. instead of *Eq*.3.3.1. This situation arises when original features used to define regions ($\mathbf{x}_i$) are not necessary the discriminative ones and perhaps a mapping of that $\psi(\cdot)$ must be fed as features to the classifier.

Therefore, the classifier function is:

$$h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{B}^T\mathbf{x} \rangle = \mathbf{w}^T\mathbf{B}^T\mathbf{x}, \tag{3.3.2}$$

in which $\mathbf{x}$ is an image concatenated into a D-dimensional vector and $\mathbf{w} \in \mathbb{R}^K$ is a vector with the same dimensionality as the number of basis vectors. In fact, $\mathbf{B}^T\mathbf{x}$ reduces the dimensionality from D to K. $\mathbf{w}$ is linearly related to the classifier, $h_{\mathbf{w}}(\cdot)$, because of computational reasons; more specifically, $\ell(\cdot)$ becomes convex with respect to $\mathbf{B}$ when $\mathbf{w}$ is fixed.

The loss term $\ell(.;.)$ penalizes misclassification of data by comparing estimated classification with class labels, y. Many choices are possible for the loss function in SVM. Here, we choose the squared hinge loss function, namely $\ell(y; h_{\mathbf{w}}(\mathbf{v})) = [1 - yh_{\mathbf{w}}(\mathbf{v})]_+^2 = \max(0, 1 - yh_{\mathbf{w}}(\mathbf{v}))^2$. Differentiability of this loss function is one of the reason for our choice and any other differentiable loss, *e.g.,* log-logistic, can also be used; this reason will be discussed more in Section 3.5. For the binary case (*i.e.,* $y_i \in \{-1, 1\}$), the discriminative loss function would be:

$$\begin{aligned} \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}, \mathbf{w}) &= \frac{1}{N} \sum_{i=1}^{N} \ell(y_i; h_{\mathbf{w}}(\mathbf{B}^T \mathbf{x}_i)) \\ &= \frac{1}{N} \sum_{i=1}^{N} [1 - y_i \mathbf{w}^T \mathbf{B}^T \mathbf{x}_i]_+^2 \end{aligned} \tag{3.3.3}$$

Notice that this loss function can also be written as:

$$\begin{aligned} \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}, \mathbf{w}) = \quad & \min_{\mathbf{w}, \xi_i \geq 0} \quad \frac{1}{N} \sum_{i=1}^{N} \xi_i^2 \\ & \text{subject to} \quad \xi_i \geq 1 - y_i \mathbf{w}^T \mathbf{B}^T \mathbf{x}_i \end{aligned} \tag{3.3.4}$$

in which the loss function is written as the optimal value of *Eq*.3.3.5 which is obviously convex with respect to ξ_i. Notice that if $\|\mathbf{w}\|_2^2$ is added to the objective function of *Eq*.3.3.5 and the optimization is performed with respect to $\mathbf{w}$ and ξ_i's jointly, *Eq*.3.3.5 is the optimization problem of Support Vector Machine (SVM) when $\mathbf{B}^T \mathbf{x}_i$ are features.

This loss function can be easily extended to the multi-class case (*i.e.,* $y_i \in \{1, 2, \cdots, L\}$) [56]:

$$\begin{aligned} \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}, \mathbf{W}) = \quad & \min_{\mathbf{w}_l, \xi_i \geq 0} \quad \frac{1}{N} \sum_{i=1}^{N} \xi_i^2 \\ & \text{subject to} \quad \mathbf{w}_{y_i}^T \mathbf{B}^T \mathbf{x}_i - \mathbf{w}_l^T \mathbf{B}^T \mathbf{x}_i \geq e_i^l - \xi_i \quad i = 1, \cdots, N \end{aligned} \tag{3.3.5}$$

where

$$e_i^l = \begin{cases} 0 & \text{if } y_i = l; \\ 1 & \text{if } y_i \neq l. \end{cases}$$

therefore, there is a $\mathbf{w}_l$ corresponding to the l'th class label. *Eq*.3.3.5 reduces to *Eq*.3.3.5 for the binary case. The decision function is

$$\arg\max_{l=1,\cdots,L} \mathbf{w}_l^T \mathbf{v}$$

Other possibilities for the loss function (*e.g.,* logistic, hinge, *etc.*) are not investigated in this thesis. For more diverse choices of the loss function, please see [78] and references therein. Some of the examples of the loss functions are shown in Figure2.9a.

3.4 Priors

Various feasible sets for $\mathbf{B}$ (*i.e.,* $\mathcal{B}$) can be defined for different applications some of which will be addressed in more details in Chapter 4. For experiments in this chapter, we use the following definition for the feasible set and postpone further explanations to Chapter 4:

$$\mathbf{b}_k \in \mathcal{B}_\lambda := \{\mathbf{b} \in \mathbb{R}^D : 0 \leq \mathbf{b} \leq 1, \quad \|\mathbf{b}\|_1 \leq \lambda\}, \quad (1 \leq k \leq K) \tag{3.4.1}$$

where $\mathbf{b}_k$ denotes the k'th column of matrix $\mathbf{B}$ and λ_3 specifies the sparsity of $\mathbf{b}_k$(for more detail see Chapter 4). $\mathcal{B}_\lambda$ indicates that it depends on λ.

We mainly focus on the regularization terms for $\mathbf{w}$ and $\mathbf{C}$ in this section. We choose ℓ_2^2 for $\mathbf{w}$, namely $\|\mathbf{w}\|_2^2$ similar to ℓ_2-SVM [38]. The rationale behind using this type of regularization for $\mathbf{w}$ is similar to that of ℓ_2-SVM. It can be shown [38] that adding this regularization for SVM encourages a linear classifier in the feature space that maximizes the margin between two classes and the decision boundary while minimizing the loss function. Another common option for regularization of $\mathbf{w}$ is ℓ_1-norm [78] that favors a sparser $\mathbf{w}$ (or fewer features). However, given that the basis vectors, $\mathbf{B}$, have already reduced the dimensionality significantly from D to K, a sparse $\mathbf{w}$ is not preferable here.

For $\mathbf{C}$, we simply impose a non-negativity constraint. Lee *et al.* [141] demonstrated that Non-

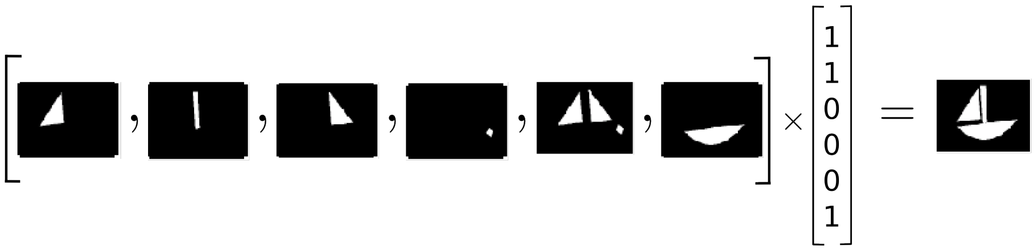


Figure 3.3: Due to non-negativity constraints, only the addition operation is allowed. If a *part* is added to an image, it cannot be subtracted; thus the algorithm must choose proper basis vectors to represent an image.

negative Matrix Factorization (NMF) is able to learn parts of faces and semantic features of text. NMF is distinguished from the other factorization methods, *e.g.,* PCA and Vector Quantization (VQ) which learn holistic but not parts-based representations, by its use of non-negativity constraints that leads to a parts-based representation because it allows only additive, not subtractive, combinations (this idea is intuitively represented in Figure3.3[1]). Donoho *et al.* [68] showed that under certain conditions, basically requiring that some of the samples are spread across the faces of the positive orthant, result in a unique decomposition. Nevertheless, as explained in the examples in Section 2.5.1, constraints on $\mathbf{C}$ and $\mathbf{B}$ depends on our modeling and interpretation of the blocks of the algorithm. We will come back to this notion in more detail in Chapter 4.

3.5 Optimization

Given the generative term (*Eq.*(3.2.1)), the discriminative term (*Eq.*(3.3.3)), and the regularization on $\mathbf{w}$ ($\|\mathbf{w}\|_2^2$), on $\mathbf{C}$ ($\mathbf{C} \geq \mathbf{0}$), and $\mathbf{B}$ (that we abstractly represent as $\mathcal{B}$), we form the optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{B},\mathbf{C}} \quad & \lambda_1 \mathcal{D}(\mathbf{X};\mathbf{B},\mathbf{C}) + \lambda_2 \ell(\mathbf{y};\mathbf{X},\mathbf{B},\mathbf{w}) + \|\mathbf{w}\|_2^2 \\ \text{subject to:} \quad & \mathbf{B} \in \mathcal{B}, \quad \mathbf{C} \geq \mathbf{0} \end{aligned} \tag{3.5.1}$$

[1] Pictures of parts of the boat shown in the figure are borrowed from presentation of a paper by Biggs *et al.* [27].

where $\mathcal{D}(\cdot,\cdot)$ and $\ell(\cdot;\cdot)$ are given in *Eq.*(3.2.1) and *Eq.*(3.3.3) respectively and $\mathcal{B}$ is the the abstract definition of the feasible set for $\mathbf{B}$ that shall be explained in Chapter 4. λ_1 and λ_2 are relative weights to control importance of the three terms in the objective function. The ratio $\frac{\lambda_2}{\lambda_1}$ controls the discriminative power vs. the generative power of the model: the higher the ratio, the more discriminative the model. Throughout the experiments, λ_1 and λ_2 are normalized by the number of samples (*i.e.,* $\lambda_1, \lambda_2 \propto \frac{1}{N}$). Note that the objective in *Eq.*(3.5.1), is comprised of three terms; thus, two regularization weights suffice to control the relative ratio of the terms.

Although this optimization is not jointly convex with respect to all variables, it is a block-wise convex program; *i.e.,* if any pair of blocks of variables is fixed, it is a convex optimization problem with respect to the other block. For example, if $\mathbf{w}$ and $\mathbf{C}$ are fixed, it is a convex optimization problem with respect to $\mathbf{B}$. Therefore, we propose a block coordinate descent (*BCD*) scheme shown in Alg.1. However, a block-wise optimization does not converges to local minimum in general. For example when the objective function is non-differentiable on joint terms between blocks, BCD may not converges (see Figure3.4 for an example). However, the following theorem guarantees that BCD converges to a local minimum but we need a lemma first [202].

Theorem 3.5.1. *(ref. [202]) The objective function of an optimization can be written as:*

$$f(\mathbf{x}_1, \cdots, \mathbf{x}_N) = f_0(\mathbf{x}_1, \cdots, \mathbf{x}_N) + \sum_{k=1}^{K} f_k(\mathbf{x}_k)$$

for some $f_k : \mathbb{R}^{n_1+\cdots+n_N} \to \mathbb{R} \cup \{\infty\}, k = 1, \cdots, N$ *and some* $f_k : \mathbb{R}^{n_k} \to \mathbb{R} \cup \{\infty\}, k = 1, \cdots, N$ *and we assume that* f *is proper, i.e.,* , $f \not\equiv \infty$. *Suppose that* $f, f_0, f_1, \cdots, f_N$ *satisfy:*

(A1) f_0 *is continuous on* $dom(f_0)$.

(A2) For each $k \in \{1, \cdots, N\}$ *and* $(\mathbf{x}_j)_{j \neq k}$, *the function* $\mathbf{x}_k \to f(\mathbf{x}_1, \cdots, \mathbf{x}_N)$ *is quasiconvex and hemivariate* [2]

(A3) $f_0, f_1, \cdots, f_N$ *are lower semicontinuous (lsc)* [3]

[2] A function is called hemivariate if it is not constant on any line segment over its domain.

[3] A function is called lower semicontinuous (lsc) if for $\forall \mathbf{x}_0 \in dom(f)$, we have $\liminf_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0)$; where

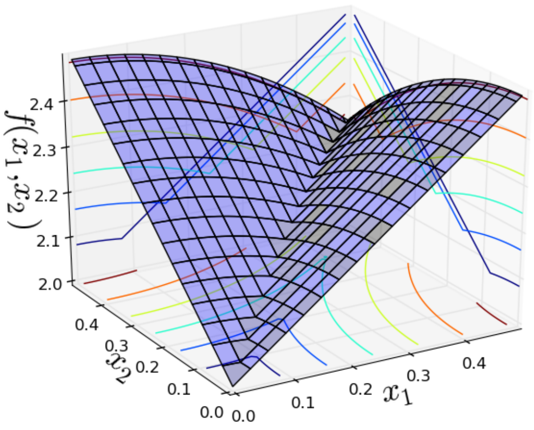


Figure 3.4: This figure shows an example of a function (f) in two dimensional space: $f(x_1, x_2) = f_0(x_1, x_2) + f_1(x_1) + f_2(x_2)$ that is not differentiable on the join term (*i.e.,* $f_0(x_1, x_2)$); therefore, a BCD applied on x_1 and x_2 alternatively may stuck in the ridge area.

and f_0 satisfies either assumption:

(B1) dom(f_0) is open and f_0 tends to ∞ at every boundary point of dom(f_0).

(B2) dom$(f_0) = \mathcal{Y}_1 \times \cdots \mathcal{Y}_N$, for some $\mathcal{Y}_k \subseteq \mathbb{R}^{n_k}, k = 1, \cdots, N$.

Also, assume that the sequence $\{\mathbf{x}^r = (\mathbf{x}_1^r, \cdots, \mathbf{x}_N^r)\}_{r=0,1,\cdots}$ generated by the BCD method. Then, either $\{f(\mathbf{x}^r)\} \to \infty$, or else every cluster point $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_N)$ is a coordinate-wise minimum point of f.

In our case:

$$f_0(\mathbf{B}, \mathbf{C}, \mathbf{w}) = \lambda_1 \mathcal{D}(\mathbf{X}; \mathbf{B}, \mathbf{C}) + \lambda_2 \ell(\mathbf{y}; \mathbf{B}, \mathbf{w}) + \|\mathbf{w}\|_2^2$$

$$f_1(\mathbf{B}) = \begin{cases} 0, & \text{for } \mathbf{B} \in \mathcal{B} \\ \infty, & \text{for } \mathbf{B} \notin \mathcal{B} \end{cases}$$

$$f_2(\mathbf{C}) = \begin{cases} 0, & \text{for } \mathbf{C} \in \mathcal{C} \\ \infty, & \text{for } \mathbf{C} \notin \mathcal{C} \end{cases}$$

f_0 is defined everywhere and it is continuous (hence lsc). f_1 and f_2 are both lsc. Because of

lim inf is the limit inferior (of the function f at point $\mathbf{x}_0$).

Algorithm 1 Block-wise Optimization

```
Require: Data ($\mathbf{X}$), Labels ($\mathbf{y}$), Regularization ($\lambda$'s)
  initialize $\mathbf{B}$, $\mathbf{C}$, $\mathbf{w}$
  $k \leftarrow 0$
  repeat
    $\mathbf{B}^{k+1} \leftarrow \arg\min_{\mathbf{B}} J_3(\mathbf{B}; \mathbf{C}^k, \mathbf{w}^k)$ (Eq.(3.5.4))
    $\mathbf{C}^{k+1} \leftarrow \arg\min_{\mathbf{C}} J_2(\mathbf{C}; \mathbf{B}^k, \mathbf{w}^k)$ (Eq.(3.5.3))
    $\mathbf{w}^{k+1} \leftarrow \arg\min_{\mathbf{w}} J_1(\mathbf{w}, \mathbf{B}^k, \mathbf{C}^k)$ (Eq.(3.5.2))
    $k \leftarrow k + 1$
  until some convergence criteria satisfied
```

our choice of the loss function in *Eq.*3.3.3, if we fix all blocks of f except one (*e.g.*, $\mathbf{w}$), the function is quadratic and hence convex and hemivariate which satisfies A2. This rationalizes our choice for the loss function in *Eq.*3.3.3 and why we preferred squared hinge versus hung loss function.

The optimization is straightforward with respect to two of the blocks ($\mathbf{C}$ and $\mathbf{w}$) but challenging with respect to the others ($\mathbf{B}$) that will be discussed in detail subsequently. Optimization of $\mathbf{B}$ depends on the definition of its feasible set that will be discussed in detail in Chapter 4

Optimization w.r.t. $\mathbf{w}$

We start with the most straightforward block. In the k'th iteration, fixing $\mathbf{B}$ and $\mathbf{C}$, the optimization should find the global minimum of the following convex function:

$$J_1(\mathbf{w}; \mathbf{B}^k, \mathbf{C}^k) = \lambda_2 \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}^k, \mathbf{w}) + \|\mathbf{w}\|_2^2 \tag{3.5.2}$$

in which $\ell(\cdot;\cdot)$ is the loss function defined in *Eq.*(3.3.5). Solving this optimization problem with respect to $\mathbf{w}$ is not challenging because it is basically a linear SVM classifier with ℓ_2^2 regularization applied on new features, namely $\mathbf{B}^T\mathbf{x}_i$. It yields a constrained quadratic optimization and any off-the-shelf quadratic solver can solve *Eq.*3.5.2 in a reasonable time. One option can be a multi-class linear SVM solver because computational complexity of such a solver is a function of the number of new features (K) and number of samples (N), which are not large in our application. We use `LIBLINEAR` [78] as the solver.

Optimization w.r.t. $\mathbf{C}$

Fixing $\mathbf{B}$ and $\mathbf{w}$ in the k'th iteration, we need to find the global optimum of the following objective with respect to $\mathbf{C}$:

$$\begin{aligned} & J_2(\mathbf{C}; \mathbf{B}^k, \mathbf{w}^k) = \|\mathbf{X} - \mathbf{B}^k \mathbf{C}\|_F^2 \\ & \text{subject to:} \qquad \mathbf{C} \geq \mathbf{0} \end{aligned} \tag{3.5.3}$$

This problem can be easily formulated as a non-negative quadratic optimization problem with $K \times N$ variables. Hessian of the objective function is $I_N \otimes (\mathbf{B}^T \mathbf{B})$ where $\otimes$ is Kronecker product and I_N is $N \times N$ identity matrix. Given that N is not typically large in medical imaging applications and K is also not large, any off-the-shelf solver (*e.g.*, `MOSEK` [1]) can solve this problem. Since the optimization problem is not very large scale for $\mathbf{C}$, an interior-point method which are known to be fast can be used to solve *Eq*.3.5.3. There is also abundant supply of options for non-negative least squared solvers.

Optimization w.r.t. $\mathbf{B}$

Fixing $\mathbf{C}$ and $\mathbf{w}$ in the k'th iteration, a constrained convex programming problem needs to be solved to find optimal $\mathbf{B}$:

$$\begin{aligned} & J_3(\mathbf{B}; \mathbf{C}^k, \mathbf{w}^k) = \lambda_1 \|\mathbf{X} - \mathbf{B}\mathbf{C}^k\|_F^2 + \lambda_2 \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}; \mathbf{w}^k) \\ & \text{subject to:} \qquad \mathbf{b}_j \in \mathcal{B}, \quad 1 \leq j \leq K \end{aligned} \tag{3.5.4}$$

where $\mathcal{B}$ is the feasible set for the columns of $\mathbf{B}$. Several choices for $\mathbf{B}$ are discussed in Chapter 4. For the experiments in this chapter, we use *Eq*.3.4.1.

Complexity of the algorithm needed to solve *Eq*.3.5.4 depends on choice of $\mathcal{B}$ that will be discussed in detail in Chapter 4. Nevertheless for large number of choices of non-trivial $\mathcal{B}$, *Eq*.3.5.4

can be a difficult optimization problem due two reasons: 1) high-dimensionality: the number of variables is at least $D \times K$ (number of voxels by number of basis vectors) plus variables introduced by the non-differentiability of the constraints or objective, and 2) constrained programming subject to a non-smooth feasible set. In general, constrained optimization problems are more expensive to solve than unconstrained optimization problem. In this section, we will introduce the general method to solve the problem above and explain the details in Chapter 4.

First we need to introduce *proximal* operator for an extended convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$:

$$\mathcal{P}_h(\mathbf{x}) = \arg\min_{\mathbf{y}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + h(\mathbf{y}) \tag{3.5.5}$$

The Projected Gradient (PG) algorithm combines a proximal step with a gradient step. PG algorithm can be used to solve constrained optimization problems. It can also be used to solve non-smooth problem at linear rates. Assuming that an objective function can be decomposed as:

$$J(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x}) \tag{3.5.6}$$

where f is smooth and h is a convex extended real valued function. Let us assume that ∇f is Lipschitz so that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.

In PG scheme, the algorithm starts from a feasible point, $\mathbf{x}_0$. Let $\alpha_1, \alpha_2, \cdots$ be a sequence of positive step sizes. The $k + 1$'th iteration of the PG scheme:

$$\mathbf{x}_{k+1} = \mathcal{P}_{\alpha_k h}(\mathbf{x}_k - \alpha_k h(\mathbf{x}_k)) \tag{3.5.7}$$

The algorithm alternates between taking gradient steps (*i.e.,* computing $\nabla f(\mathbf{x})$) and then taking proximal point steps. Notice that the decomposition in *Eq.*3.5.6 matches with *Eq.*3.5.4; in this case f is simply the objective function and h is the indicator function of $\mathcal{B}$ (similar to 3.1.4); therefore its proximal operator become projection on $\mathcal{B}$. We represent this special proximal operator as $\prod_{\mathcal{B}}(\mathbf{x})$:

$$\prod_{\mathbf{B}}(\mathbf{x}) \triangleq \mathcal{P}_h(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathcal{B}} \|\mathbf{x}-\mathbf{y}\|_2^2$$

The objective function in *Eq*.3.5.4 consists of two terms: 1) the generative term ($\mathcal{D}(\mathbf{X};\mathbf{BC})$), 2) and the discriminative term ($\ell(\mathbf{y};\mathbf{X},\mathbf{w},\mathbf{B})$). Derivative of the generative term with respect to $\mathbf{B}$ is:

$$\nabla_{\mathbf{B}}\mathcal{D}(\cdot;\cdot) = 2(\mathbf{BC}) \odot (\mathbf{X}-\mathbf{BC})$$

In general case, when a divergence term is used, the derivative of the divergence term is:

$$\nabla_{\mathbf{B}}\mathcal{D}(\cdot;\cdot) = \phi''(\mathbf{BC}) \odot (\mathbf{X}-\mathbf{BC})$$

where ϕ'' is the second derivative of $\phi(\cdot)$ which is define earlier in Def. 3.2.1 and $\odot$ is element-wise matrix multiplication.

Derivative of the discriminative term with respect to k'th column of $\mathbf{B}$ in a binary case (*i.e.,* $y_i \in \{-1,+1\}$) is:

$$\begin{aligned}\nabla_{\mathbf{b}_k}\ell(\cdot;\cdot) &= \sum_{i\in\mathcal{I}}^{N}(1-y_i\sum_{j=1}^{K} w_j\mathbf{b}_j^T\mathbf{x}_i)(-y_i w_k\mathbf{x}_i)\\ &= \sum_{i\in\mathcal{I}}^{N}(\sum_{j=1}^{K} w_j\mathbf{b}_j^T\mathbf{x}_i - y_i)(w_k\mathbf{x}_i)\end{aligned}$$

in which $\mathcal{I} \equiv \{i|1-y_i\mathbf{w}^T(\mathbf{B}^T\mathbf{x}_i) > 0\}$. It also follows similarly for the multi-class case (*i.e.,*

$y_i \in \{1, 2, \cdots\}$):

$$\nabla_{\mathbf{b}_k} \ell(\cdot;\cdot) = \sum_{i\in\mathcal{I}}^{N} (\sum_{j=1}^{K} 1 + (w_{j\hat{y}_i} - w_{jy_i})(\mathbf{b}_j^T \mathbf{x}_i))(w_{k\hat{y}_i}\mathbf{x}_i - w_{ky_i}\mathbf{x}_i) \tag{3.5.8}$$

in which $\mathcal{I} \equiv \{i | 1 + \mathbf{w}_{\hat{y}_i}^T(\mathbf{B}^T\mathbf{x}_i) - \mathbf{w}_{y_i}^T(\mathbf{B}^T\mathbf{x}_i) > 0\}$ and w_{jy_i} is the j'th element of classifier corresponding to the class label y_i and $w_{j\hat{y}_i}$ is the j'th element of the class label (incorrectly) associated with the i'th sample; the incorrect label is $\hat{y}_i$.

Projected Gradient (PG) [26] is a first order method that can be used for a constrained problem. However, PG can be slow particularly for non-smooth feasible sets. The newton method is used to accelerate first-order solvers [26]. The Interior Point (IP) method is a variant of the Newton method for a constrained problem [35]. However, the IP method implemented naively fails to solve *Eq.*(3.5.4) because IP involves computation and inversion of a Hessian matrix which is prohibitive in term of computation and memory costs. In our experiments, more sophisticated implementations like `MOSEK` [1] fail to find a point in the feasible set in a reasonable time. Our chosen alternative is use to use Spectral Projected Gradient (SPG) [28] that is a modification of the classical PG method which differs in two essential ways: 1) It uses a non-monotone line search that measures descent with respect to a fixed number of previous iterations instead of just the last iteration. This may lead to a temporary increase in the objective while ensuring overall convergence. 2) It uses spectral step length introduced by Barzilai-Borwein (BB) [16] that gives an initial step length. In the BB approach, the step length (α_t) in t'th iteration is chosen such that $\alpha_t^{-1}\mathbf{I}$ mimics the Hessian of the objective over the most recent step. Similar approaches have been taken recently by Schmidt *et al.* [183] and Wright *et al.* [226] for large-scale non-smooth problems. There are several choices for BB step length [59], in this thesis, we choose the following method to compute it [205]:

$$\mathbf{s}^k = vec(\mathbf{B}^k), \;\; \mathbf{g}_t = vec(\nabla J_3(\mathbf{B}^k))$$

Algorithm 2 Spectral Projected Gradient Solver

```
Require: Initial point, step-length bounds $0 < \alpha_{\min} < \alpha_{\max}$, $\nu$, $M$
  repeat
    $\mathbf{d} \leftarrow \prod_{\mathcal{B}}(\mathbf{s}^k - \alpha\mathbf{g}^k) - \mathbf{s}^k$
    $\gamma \leftarrow 1$
    $M \leftarrow \max_{k-M \le i \le k}\{J_3(\mathbf{s}^i)\}$
    while $J_3(\mathbf{s}^k + \gamma\mathbf{d}) > M + \nu\gamma\langle\mathbf{g}^k, \mathbf{d}\rangle$ do
      Choose $\gamma \in (0,1)$ with quadratic interpolation [96]
    end while
    $\mathbf{s}^k \leftarrow \mathbf{s}^k + \gamma\mathbf{d}$
    compute step-length: $\alpha_k \leftarrow \min\{\alpha_{\max}, \max\{\alpha_{\min}, \alpha_{bb}\}\}$ ($\alpha_{bb}$ in Eq.(3.5.9))
    $k \leftarrow k + 1$
  until some convergence criteria satisfied
```

$$\mathbf{q}^k = \mathbf{s}^k - \mathbf{s}^{k-1}, \mathbf{p}^k = \mathbf{g}^k - \mathbf{g}^{k-1}$$
$$\alpha_{bb} = \frac{\|\mathbf{q}^k\|_2}{\|\mathbf{p}^k\|_2} \tag{3.5.9}$$

where $vec(.)$ is an operator that reorders elements of a matrix into a vector.

Our proposed algorithm is shown in Alg.(2). It is conceivable that the bottleneck of the algorithm is the projection ($\mathcal{P}_{\mathcal{B}}(\cdot)$) because it should be performed in each iteration. In Chapter 4, we will discuss various choices for $\mathcal{B}$ and their practical implications. We will also propose efficient approach to compute $\mathcal{P}_{\mathcal{B}}(\cdot)$ for each choice.

3.6 Experiments

In this section, we first perform some experiments to study different aspects of the proposed framework. In the first set of experiments, we generate some synthetic images that are combinations of some *normal* and *abnormal* variations (*i.e., effect*). Effectiveness of the proposed method in recovery of the correct effect is investigated in different ratios of variations. In the second set of experiments, we apply the method on a benchmark set of images of digits[4] to represent a few examples of basis vectors. The effect of the balance between the generative term to the discrimi-

[4] The US Postal (USPS) handwritten digit dataset is derived from a project on recognizing handwritten digits on envelopes [106], [113].

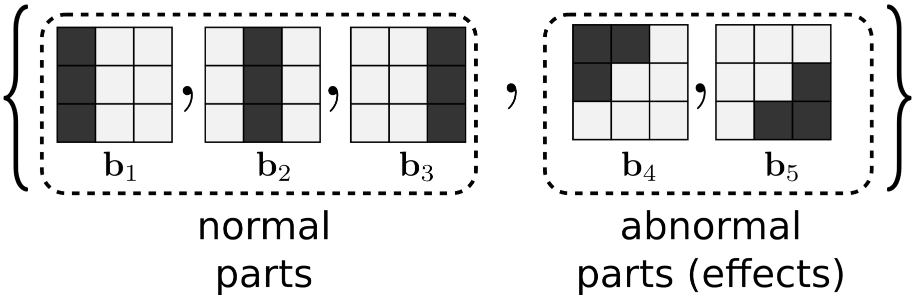


Figure 3.5: This figure represents the *normal* ($\mathbf{b}_1$,$\mathbf{b}_2$,$\mathbf{b}_3$) and the *abnormal* (*effect*) ($\mathbf{b}_4$,$\mathbf{b}_5$) parts (basis vectors) used for simulation. Normal images are allowed to use only two out of three normal parts. Abnormal images are addition of normal parts and abnormal parts. Contribution of parts in image are specified with random coefficients.

native term is studied. We also study the influence of the generative discriminative ratio on the classification rate. At the end of this chapter, we apply the method on high dimensional real brain images to show practical application of the method for classification purposes. The dataset consists of brain images from two cohorts of subjects: subjects diagnosed with Alzheimer disease and normal controls. We evaluate the results quantitatively in terms of classification accuracy and qualitatively by comparing with findings to prior clinical reports and facts. Sensitivity analysis is also reported on varying the ratio between the generative and the discriminative terms.

In all of the experiments in this section, the feasible set of $\mathbf{B}$ ($\mathcal{B}$) is defined as the intersection of ℓ_1 and ℓ_∞ norms in the positive orthant; it can be represented mathematically as:

$$\mathbf{B} \in \mathcal{B}_\lambda = \{\mathbf{b}_k : 0 \leq \mathbf{b}_k \leq 1, \|\mathbf{b}_k\|_1 \leq \lambda\} \subset \mathbb{R}^{D \times K}, \quad \forall 1 \leq k \leq K \tag{3.6.1}$$

We call this feasible set *Boxed-Sparsity*. The reason for such a choice for the feasible set will be elaborated in Chapter 4. Since we are interested to investigate the impacts of the generative and the discriminative terms in this section, we keep the definition of the feasible set for $\mathbf{B}$ the same in all experiments and set λ to a reasonable value that is specified by each experiment.

3.6.1 Synthetic Data: Effect Recovery

In this set of experiments, parametric consistency of the algorithm is studied empirically. In other words, under certain generative assumptions to generate samples (*i.e.,* images), we empirically study if the algorithm can successfully recover the correct parameters (*i.e.,* basis vectors). Here, we assume a simple setting in which there is *normal* variation in the population that is represented as a non-negativity linear combination of basis vectors (*i.e.,* parts). For simplicity, we assume that there are three parts available; each part is represented as a vertical box occupying one third of horizontal axis of the image domain and the whole vertical axis (as shown in Figure3.5). Each image consists of two parts randomly selected from available three parts and added to the image with a non-negative random coefficient ($c_{ik}, k = 1, 2, 3$). Images generated under such assumptions constitute the normal cohort; *i.e.,* $y_i = 1$. Generative scheme is the same for so-called *abnormal* images (*i.e.,* $y_i = -1$) except that $\mathbf{b}_4$ and $\mathbf{b}_5$ are also added to the image with random contribution ($\hat{c}_{i4}$ and $\hat{c}_{i5}$ for $\mathbf{b}_4$ and $\mathbf{b}_5$ respectively):

$$\begin{aligned}
&s_k \in \{0,1\}, && s_1 + s_2 + s_3 = 2, \mathbb{P}(s_1) = \mathbb{P}(s_2) = \mathbb{P}(s_3) \\
&c_{ik} \sim \mathbb{U}[0, \gamma_1], && \hat{c}_{ik} \sim \mathbb{U}[0, \gamma_2], \qquad \varepsilon_i \sim \mathbb{U}[0, \epsilon] \\
&\mathbf{x}_i = \sum_{k=1}^{3} (s_k c_{ik}) \mathbf{b}_k + \varepsilon_i, && y_i = 1 \qquad 1 \le i \le N_1 \\
&\mathbf{x}_i = \sum_{k=1}^{3} (s_k c_{ik}) \mathbf{b}_k + \sum_{k=4}^{5} (s_k \hat{c}_{ik}), \mathbf{b}_k + \varepsilon_i, && y_i = -1 \qquad N_1 + 1 \le i \le N
\end{aligned} \tag{3.6.2}$$

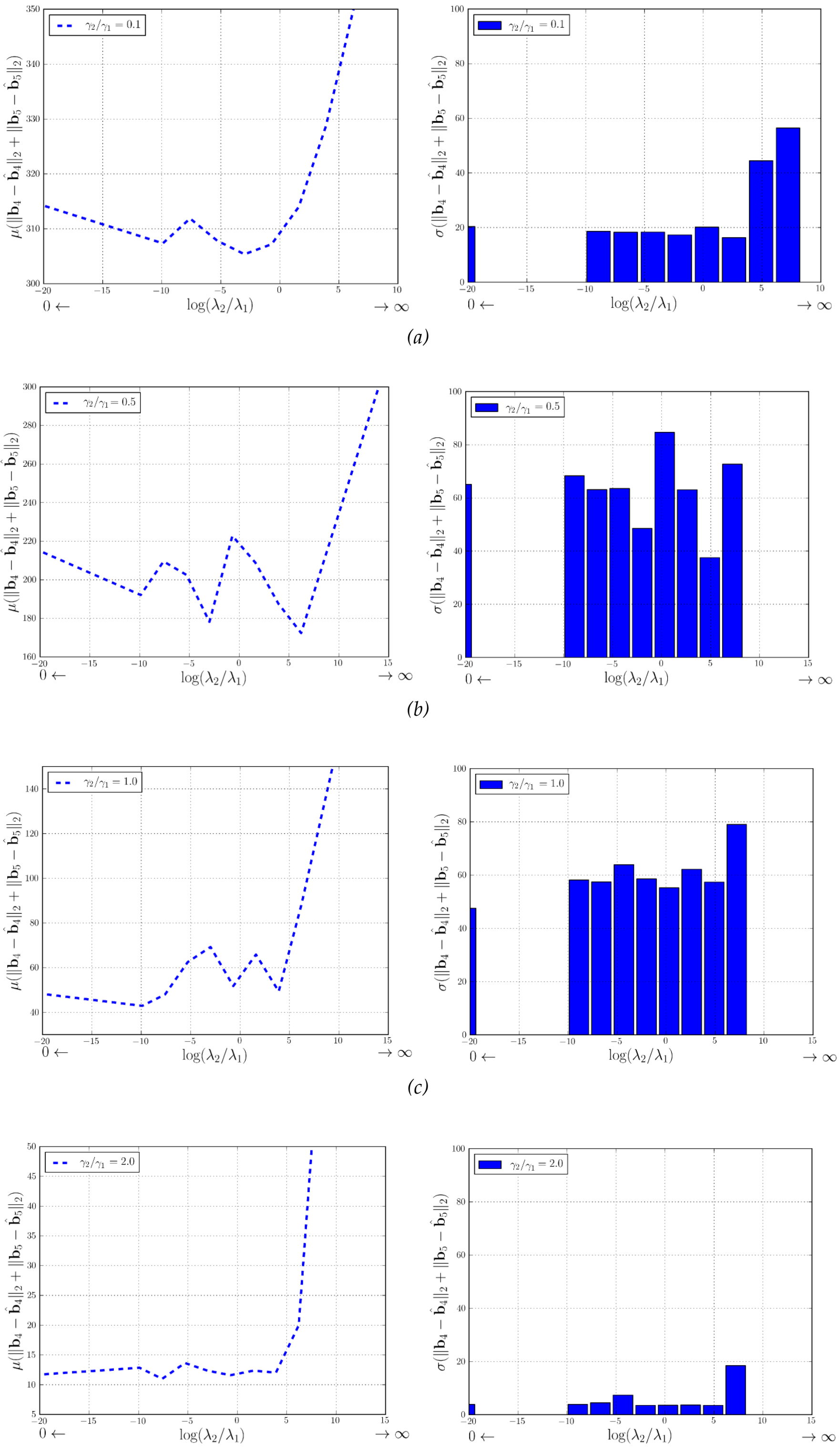
where s_k are selector variables ($s_k \in \{0,1\}, k = 1, 2, 3$) and only two out of three can be one ($\sum_k s_k = 2$) with the same probability. $\mathbb{U}[a, b]$ is a uniform random distribution between a and b, γ_1 and γ_2, and ϵ are constants denoting maximum magnitudes coefficients of normal parts, abnormal parts and noise respectively. The ratio $\frac{\gamma_2}{\gamma_1}$ controls the effect regime: the higher the ratio, the stronger the effect. Each box in basis vectors represented in Figure3.5 denotes a 30×30 pixel image; hence each basis vector is a 90×90 pixel image.

The ratio γ_2/γ_1 denotes different strength of the effect with respect to the original signal; the higher the ratio, the more salient the effect of the abnormal parts (see Figure3.5). The other ratio, λ_2/λ_1, indicates the balance between discriminative and generative terms; the higher the ratio, the stronger the discriminative term. We designed the following experiment: for different ratios of γ_2/γ_1, we change λ_2/λ_1, and studied how successful we the method can detect the abnormal part. The aim of the experiment is to study the effect of balance between generative and discriminative term on detecting the abnormal part for varying strength of the abnormal coefficients. We repeated the experiment 10 times by regenerating the sample images according to the generative scheme given in *Eq*.3.6.2. For evaluation, we found the closest basis vectors to $\mathbf{b}_4$ and $\mathbf{b}_5$ in ℓ_2-norm sense; let us call them $\hat{\mathbf{b}}_4$ and $\hat{\mathbf{b}}_5$ respectively. $\|\mathbf{b}_4 - \hat{\mathbf{b}}_4\|_2 + \|\mathbf{b}_5 - \hat{\mathbf{b}}_5\|_2$ is used as a measure to quantify how well the algorithm was able to capture the actual effects. Figure3.6 reports the means and standard deviations of the measure for different values of γ_2/γ_1.

As expected, decreasing (γ_2/γ_1) which means weaker effect, deteriorates the average detectability of the effect signal in general. However, unless the effect signal is very strong (*e.g.*, Figure3.6-d) optimal ratios λ_2/λ_1 lie somewhere (or in multiple places) between the ends of the spectrum ($[0, \infty)$) and deteriorates at both ends. If effect is very strong, *e.g.*, Figure3.6-d where the effect is twice as strong as the maximum variations in the normal image, the generative term is enough to capture effect basis vectors. In such cases, effect basis vectors have such strong variations that the generative term dedicates, which only tries to explain the data, dedicates a few basis vectors to explain them. This experiment shows that in general, the discriminative term is useful to recover the actual effects unless in extreme cases where the effects dominate the variation in a dataset.

3.6.2 Experiment on a Benchmark Data: Handwritten Digits

In order to show a simple yet illustrative example for application of the algorithm, we apply it on the US Postal (USPS) handwritten digit dataset that is derived from a project on recognizing



(a)

(b)

(c)

(d)

Figure 3.6: Figures plot average and standard deviations of distance between the closest basis vectors to the effect basis vectors on y-axis with respect to different ratios of the discriminative versus the generative terms (λ_2/λ_1) on x-axis. (a)-(d) plots represent different rates of γ_2/γ_1; *i.e.*, different strength of effect.

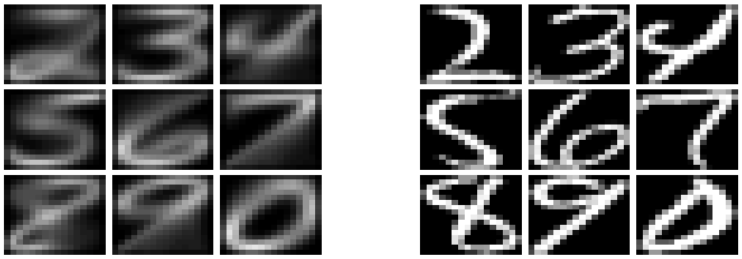


Figure 3.7: The figure on the left shows pixel-wise average of the 9 digits available in USPS dataset and the figure on the left shows an example of each from the dataset.

handwritten digits on envelopes [106], [113]. Similar to the experiments in the previous section and all of the experiments in this chapter, we do not change the definition of the feasible set for $\mathbf{B}$ (*i.e.,* we use *Eq.*3.6.1) and its parameters are set to a reasonable value. Therefore, we only focus on the roles of the generative and discriminative terms in this chapter.

The USPS data set consists of 1,100, 16×16 gray-scale images of handwritten digits (1,100 images of each digit 0 through 9). All images are aligned with affine transformation. Figure3.7 shows average images of the digits in the database.

In order to illustrate the effect of the generative and the discriminative terms, 100 images of “6” and “8” were selected randomly to form the training sample. To investigate the effect of the discriminative term, λ_1 was set to constant value and λ_2 was varied over a large range $(0, \infty)$. For this experiment, we set the number of basis vectors to 16 ($K = 16$) and λ_3 was set to 20% of number of pixels (*i.e.,* $\lambda_3 = 0.2D$, where $D = 16$ is number of pixels). Some results are shown in Figure3.8. In addition to basis vectors, we have also shown $\mathbf{B}|\mathbf{w}| = \sum_{k=1}^{K} \mathbf{b}_k |w_k|$, where $|w_k|$ denotes the absolute value of the k'th element of $\mathbf{w}$. This measure can be viewed as a qualitative measure for how well the algorithm can delineate area of difference between two characters; if the algorithm dedicates some of basis vectors to minimize $\ell(\cdot, \cdot)$, they should high contribution in the loss function and hence have values for $|w_k|$.

For very small ratio (*i.e.,* $\lambda_2/\lambda_1 \to 0$) which is equivalent to the very generative formulation, the algorithm generates only part-based representation of the training set (see examples of the

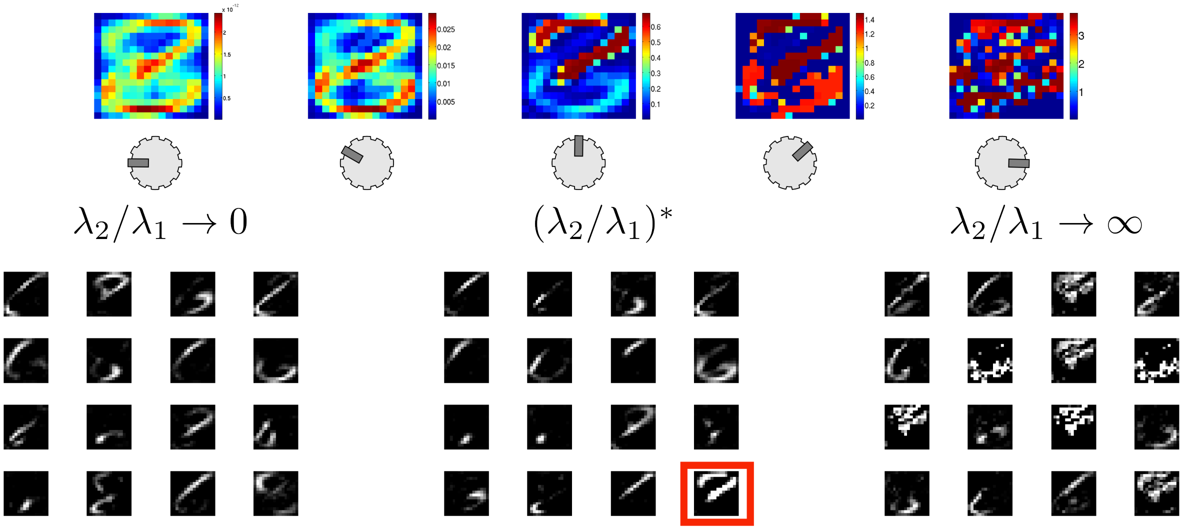


Figure 3.8: The figure shows images of $\mathbf{B}|\mathbf{w}| = \sum_{k=1}^{K} \mathbf{b}_k|w_k|$ as a qualitative measure of performance of the algorithm in detecting *discriminative parts*. The knobs symbolize the ratio of the discriminative to the generative term from almost zero (generative) on the left (discriminative). The figures on the bottom show basis vectors ($K = 16$). Notice that for a so-called *optimal* ratio ($(\lambda_2/\lambda_1)^*$), $\mathbf{B}|\mathbf{w}|$ denotes areas of difference between "6" and "8" with hot colors and some of basis vectors (highlighted by a red box) represent the discriminative parts. For small values of the ratio, the model is mostly generative and $\mathbf{B}|\mathbf{w}|$ has a lot of non-zero values all over the image with very small magnitude. For large values of λ_2/λ_1, the algorithm tries to over-fit for the labels of the training samples by adding as many pixels as it can to decrease the $\ell(\cdot;\cdot)$ on the training data.

basis vectors in bottom left of Figure3.8). However, such basis are not *necessarily* optimal for classification. In addition, they weakly delineate areas in which "6" and "8" differ as it can be seen in $\mathbf{B}|\mathbf{w}|$. For very large values of the discriminative term (*i.e.*, $\lambda_2/\lambda_1 \to \infty$), the algorithm tries to be purely discriminative and it is not loyal in term of representation of the dataset (see examples of the basis vectors in bottom right of Figure3.8). It can be seen in $\mathbf{B}|\mathbf{w}|$ that aggressively adds pixels that are even slightly discriminative features for the training sample. Large ratio of λ_2/λ_1 drives the algorithm to only reconstruct the labels ($\mathbf{y}$) on the training set that are not necessarily good in term of generalization on a test data. For a range value of the ratio λ_2/λ_1, the algorithm can both optimally outline area of the image in which "6" and "8" differ and at the same time represent the dataset. In fact, it generates a *discriminative part* as one of the basis vectors in $\mathbf{B}$. Areas of difference also stand out in $\mathbf{B}|\mathbf{w}|$; it shows that "6" and "8" differ mostly on the top and oblique band in the middle of the image and features extracted from these areas are the most discriminative ones. The figures in the middle bottom of Figure3.8 show examples of such basis vectors, notice the basis vector highlighted by the red box in Figure3.8 that represents the discriminative part of “6” and “8”.

3.6.3 Generative versus Discriminative Trade-Off

The images used in this experiment are structural MR brain images (`T1` image) obtained from Alzheimer’s Disease Neuroimaging Initiative (ADNI[5]). 63 normal control (NC) individuals and 54 AD patients were pre-processed via the same pre-processing pipeline. The pre-processing pipeline is designed according to previously validated and published techniques by Goldszal *et al.* [90]. It includes the following steps: 1) alignment of images to the AC-PC plane; 2) removal of extra-cranial material (skull-stripping); 3) tissue segmentation into gray matter (GM), white matter (WM), and cerebral fluid (CSF), using a brain tissue segmentation method proposed in Pham *et al.* [166]; 4) non-rigid image warping using the method proposed by Shen *et al.* [186] to a standardized coordinate system, a brain atlas (template) that was aligned with MNI coordinate space [125]; 5) formation of regional volumetric maps, named RAVENS maps (see [90] and [61]), using tissue-preserving image warping [90]. RAVENS maps quantify the regional distribution of a GM, WM, and CSF, since one RAVENS map is formed for each tissue type. A RAVENS map quantifies an expansion (or contraction) of the tissue modeled by a transformation that warps the image from the original space to the template space. Consequently, voxel values of a RAVENS map in a template space are directly proportional to the volume of the respective structures in the original brain scan. Although this map can be formed for CSF, WM, and GM, we only used maps corresponding to the GM tissue type. An example of GM, WM, and ventricle RAVENS map is shown in Figure3.9.

In order to investigate the effect of the hybrid generative-discriminative model, we modified

[5] www.loni.ucla.edu/ADNI

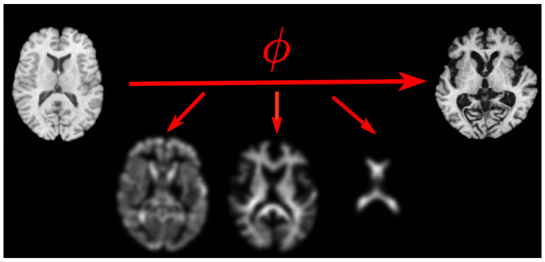
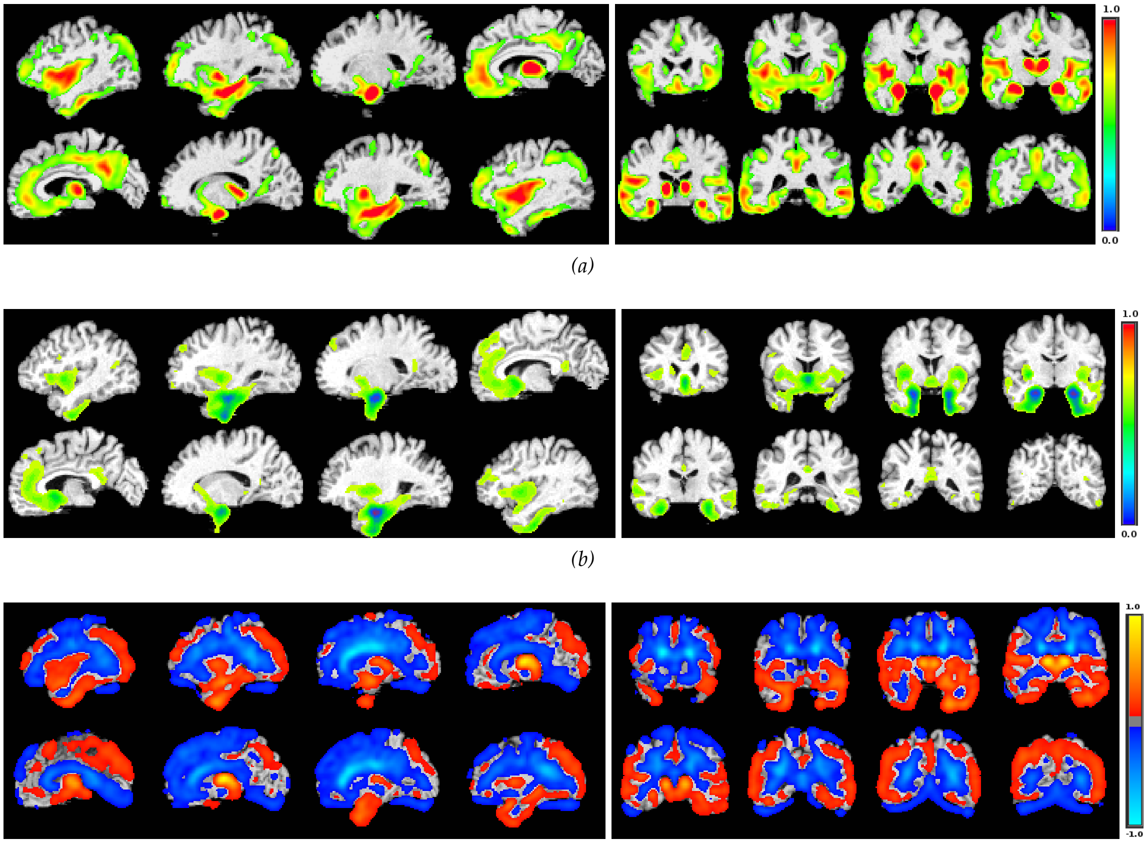


Figure 3.9: Examples of RAVENS maps for the tissue types created from the transformation (ϕ) that warp the template (top, left) to the subject (top, right). The image shows the RAVEN maps for the tree tissue type: Gray Matter (GM, bottom left), White Matter (WM, bottom middle), and Cerebral Spinal Fluid (CSF, bottom right).

the λ_2/λ_1 ratio for various numbers of basis vectors (K). In this experiment, Boxed-Sparsity was used as the sparsity regularization and λ_3 was set to 20% (*i.e.,* $\lambda_3/D = 1/5$). The number of basis vectors (K) was chosen from set of $\{5, 10, 15, 20, 30, 40, 50\}$ to examine robustness of the algorithm to different numbers of basis vectors. As mentioned earlier in the methods section, the proposed algorithm can be viewed as a dimensionality reduction from an original large dimension (D) to smaller but more discriminative and representative dimensions (K); hence so-called *projection* $\mathbf{B}^T\mathbf{x}$ can be viewed as feature extraction. While the original dimension may be too large to apply a non-linear classifier on, we can simply apply a classifier (in this experiment Logistic Model Trees [137] [6]) on the extracted features (K-dimensional instead of D-dimensional) to boost the performance. For each setting, *i.e.,* a particular ratio of λ_2/λ_1 and number of basis vectors (K), data was split into 10-folds; training including learning ($\mathbf{B}, \mathbf{C}, \mathbf{w}$) and training a classifier on the extracted features ($\mathbf{B}^T\mathbf{x}_i$), was conducted on 9-fold and the test was carried on the remaining fold. This process was repeated 10 times to compute an average classification accuracy; hence, each point in Figure3.11 is the 10-fold cross-validation accuracy. Results are shown in Figure3.11. In order to avoid occlusion of the Figure3.11a, error-bars (*i.e.,* standard deviations of the accuracy rates) are added as a separate figure (Figure3.11b).

In Figure3.11, as number of basis vector (K) increases, the accuracy rates also increase but they reach a plateau around $K \in (20, 40)$. An excessively discriminative model (yellow and

[6] This classifier is called Simple Logistic in Weka [102].



(a)

(b)

(c)

Figure 3.10: Three examples of basis vectors with three different methods ($\lambda_3/D = 20\%$): (a) one of the basis vectors learned by the proposed method on sagittal and coronal cuts and; (b) one of the basis vectors learned by the NMF method on sagittal and coronal cuts and, (c) one of the basis vectors learned by the SVD method on sagittal and coronal cuts.

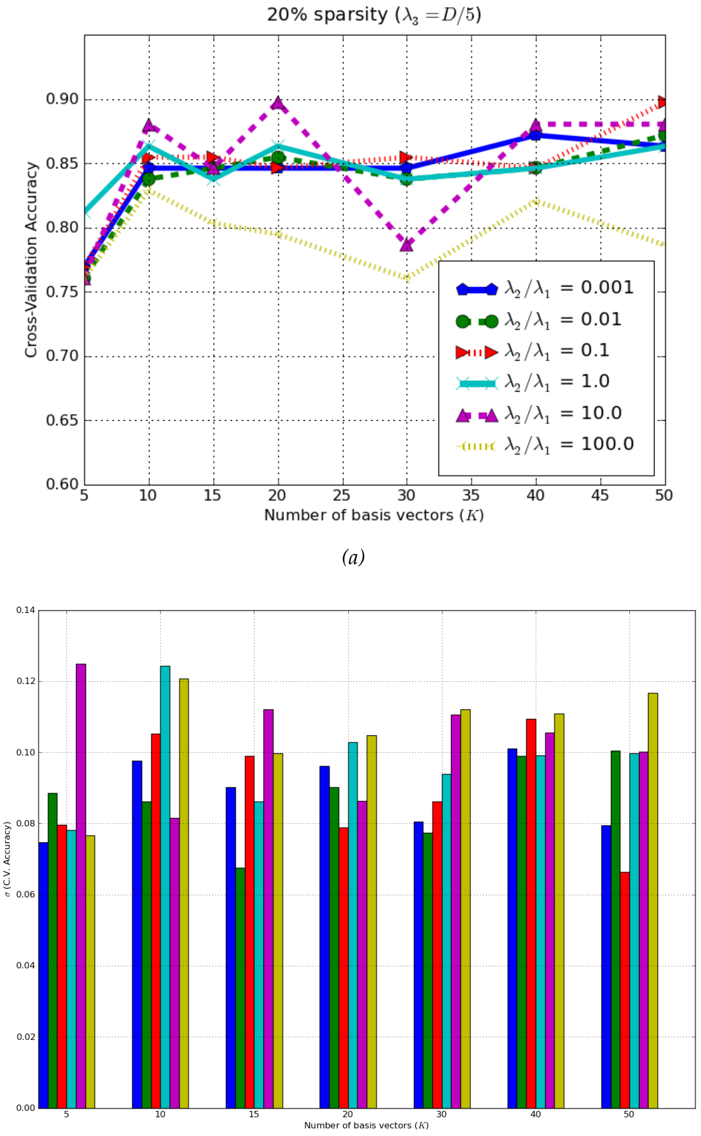
violet corresponding to $\lambda_2/\lambda_1 = 100$ and $\lambda_2/\lambda_1 = 10$ respectively) becomes more unstable as the number of basis vector increases while the blue graph, in which the generative term dominates, is quite stable. Increasing the number of basis vectors further, not only increases computational cost drastically but also degrades generalization of the model because of high dimensionality, since the number of samples is of the same order of magnitude (in this experiment $N = 117$), so we set the maximum number of basis vectors to 50 which is in the same order magnitude. The best performance is shown by red line ($\lambda_2/\lambda_1 = 0.1$) that maintains a balance between the generative and discriminative terms. This graph shows that having the generative term helps to create more stable classification rates. It also shows that unless the algorithm is pushed too much toward the discriminative side, it is fairly robust with respect to choice of parameters; for example for $K = 30$, perturbations in classification accuracy rates are about 6% for a reasonable range of λ_2/λ_1 (*i.e.*, around 0.01 and 0.1 for this data). Notice that in this cross validation process, every fold contains few samples (between 11 to 13 samples) and 7%-9% missclassification is about one miss classification per fold.

Figure3.10 compares basis vectors learned by the proposed algorithm with those of NMF and SVD. The basis vectors are overlaid on the corresponding anatomical template on various slices of sagittal and coronal cuts. In the cases of the proposed algorithm (Figure3.10a) and NMF (Figure3.10b), voxels of the basis vectors with values less than 0.3 are shown transparent for the sake of a better visualization; in case of SVD, values of voxels can be positive or negative, hence only values around zero are set to transparent. Figure3.10a clearly show Hippocampus and temporal lobe which are associated with memory and have been frequently reported [45], [51] and [127] to undergo significant shrinkage in course of the Alzheimer's disease. Hippocampus is also clearly depicted in the basis vector learned by NMF method (Figure3.10b); however, in the basis vector learned by SVD, almost all areas have nonzero positive and negative values and hence it does not clearly show which areas are important.

In order to further investigate the effect of K (number of basis vectors) on the classification accuracy, we chose 100 subjects consisting of two cohorts (50 for AD, and 50 for normal). The data is divided into 5-folds and K was varied over larger range. Figure3.12 shows the average accuracy rates. For most of the ratios of λ_2/λ_1, the average accuracy rates reach their peaks around $K \in (20, 40)$ and drop after that.

3.7 Conclusion and Discussion

In this chapter, we introduced our main framework. The method is formulated as a matrix factorization framework. It consists of three major terms: the generative term, the discriminative term



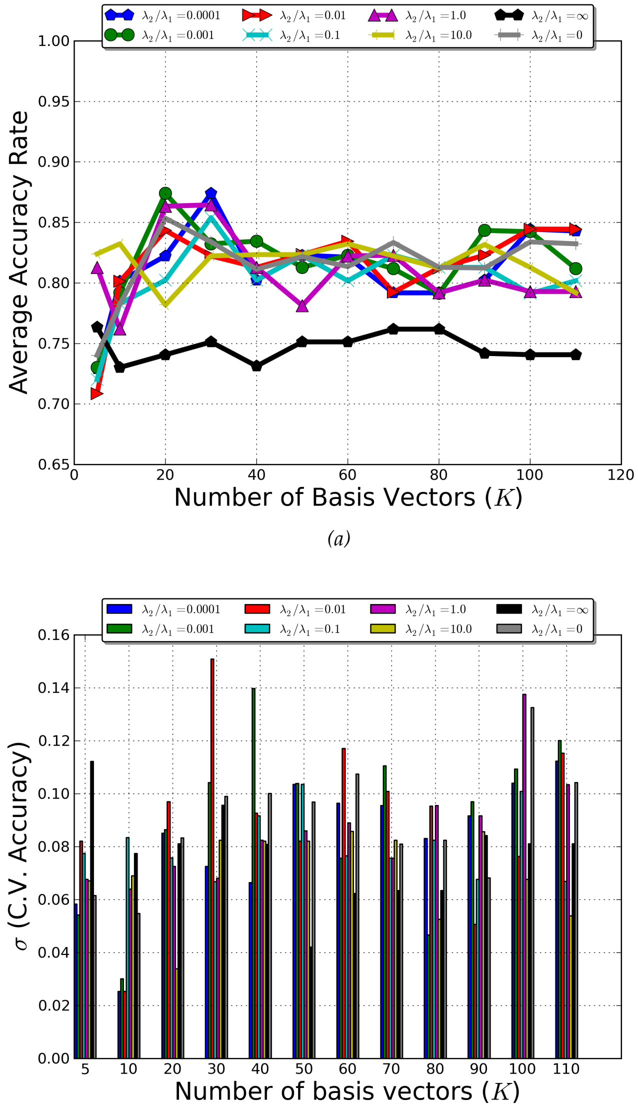
(a)

(b)

Figure 3.11: Average classification rates in 10-fold cross-validation for various ratios of $\frac{\lambda_2}{\lambda_1}$ (discriminative vs. generative) for different number of basis vectors; *i.e.*, various K. To avoid occlusion, standard deviations of the accuracy rates are added as a separate figure in (b). The y-axis, σ(C.V. Accuracy), indicates the standard deviations of the accuracy rates. The colors are the same as (a).

and the regularizer terms which can also be viewed as feasible sets (see *Eq*.3.4). We explained the feasible sets for $\mathbf{C}$ and $\mathbf{w}$ and briefly discussed $\mathbf{B}$. Since an exact definition of the feasible set for $\mathbf{B}$ depends on an application, we left elaborative discussion to Chapter 4.

It is shown in our illustrative examples in Section 2.5.1 that the generative term clusters (seg-



(a)

(b)

Figure 3.12: Average classification rates in 5-fold cross-validation for various ratios of $\frac{\lambda_2}{\lambda_1}$ (discriminative vs. generative) for longer range of K. To avoid occlusion, standard deviations of the accuracy rates are added as a separate figure in (b). The y-axis, σ(C.V. Accuracy), indicates the standard deviations of the accuracy rates. The colors are the same as (a).

ments) voxels together and the discriminative term encourages to form clusters that are discriminative. Simulation experiments in Section 3.6.1 showed that unless the *effect* size that differentiated two groups is very strong compared to the background signal, the discriminative term is required in order to recover the effect correctly. In Section 3.6.2, we showed how a balance choice of λ_2/λ_1 can help to recover areas of difference between two groups. Experiments with real data in Section 3.6.3 showed that the algorithm is robust with respect to choice of λ_2/λ_1 ratio as long as it is chosen within a reasonable range.

Chapter 4

Regularizers and Optimizers

4.1 Overview

Remember in Chapter 3, we introduced the main formulation that consists of three blocks (variables), $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{w}$:

$$\begin{aligned}(\mathbf{B}^*, \mathbf{C}^*, \mathbf{w}^*) = \arg \min_{\mathbf{B},\mathbf{C},\mathbf{w}} & \mathcal{D}(\mathbf{X}; \mathbf{B}, \mathbf{C}) + \ell(\mathbf{y}; \mathbf{X}, \mathbf{B}, \mathbf{w}) + \mathcal{R}(\mathbf{B}, \mathbf{C}, \mathbf{w}) \\ \text{subject to:} \qquad & \mathbf{B} \in \mathcal{B} \quad \mathbf{C} \in \mathcal{C} \quad \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{4.1.1}$$

All feasible sets were discussed in Chapter 3 ($\mathcal{C} : \mathbf{C} \geq 0$, $\mathcal{W} : \mathbb{R}^K$) except $\mathcal{B}$ which is to be discussed in this chapter. $\mathbf{B}$ is a matrix columns of which are the basis vectors. Each basis vector (column of $\mathbf{b}_k$) lives in $\mathbb{R}^D$ which has the same dimensionality as number of voxels of the images in the training set. We also explained in Chapter 3 that the most discriminative basis vector reveals the *effects*. Our prior knowledge about the effects is encoded in $\mathcal{B}$. Different applications call for different definitions of $\mathbf{B}$ two of which are introduced in this chapter: *Boxed-Sparsity* and *Group-Sparsity*. We also briefly discussed other possibilities and corresponding applications.

The optimization issues were also addressed in Chapter 3. We proposed an efficient first-order algorithm for optimization of $\mathbf{B}$. It was shown that different definitions of $\mathcal{B}$ only changes the proximal operator which is just a projection function:

$$\mathcal{P}_{\mathcal{B}}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathcal{B}} \|\mathbf{x} - \mathbf{y}\|_2^2 \tag{4.1.2}$$

It is also shown in Section 3.5 that the projection is the bottle-neck of the algorithm 2 because it needs to be repeated in every iterations of the algorithm. In this chapter, for each $\mathcal{B}$, an efficient algorithm for projection is proposed.

Both *Boxed-Sparsity* and *Group-Sparsity* are various definitions of the sparsity and consequently depend on a parameter specifying amount of sparsity. We investigated effect of such parameter on the classification accuracy on a real brain image dataset. Finally, we compare the classification results with the state-of-the-art algorithm on the real data.

4.2 Boxed-Sparsity

We would like to encourage basis vectors that act like indicator functions. Mathematically speaking, we would like the elements of $\mathbf{b}_k$ to be either 0 or 1, namely $\mathbf{b}_k \in \{0,1\}^D$. In addition, we are interested in finding localized basis vectors for two reasons: it increases robustness and interpretability of basis vectors. The sparsity constraint promotes the indicator functions that select subsets of voxels. The ℓ_0-norm, which counts number of nonzero entities in a vector, can be used as a regularization or constraint in order to encourage or bound sparsity. Here, we prefer to use sparsity as a constraint. Hence, a basis vector should reside in the intersection of two sets: the set of indicator functions and the set of sparse vectors, which can be written mathematically as follows:

$$\{\mathbf{b_k} \in \{0,1\}^D\} \cap \{\mathbf{b_k} \in \mathbb{R}^D : \|\mathbf{b_k}\|_0 \leq \lambda\}, \quad 0 \leq k \leq K$$

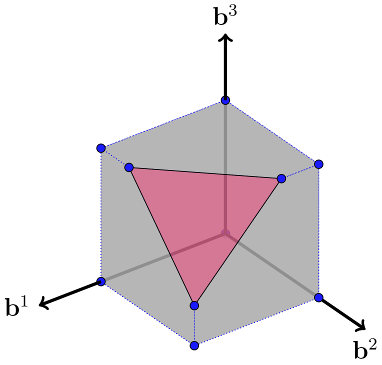


Figure 4.1: Graphical representation of *Boxed-Sparsity* ball for a hypothetical image consisting three voxels. Therefore, each basis vector lives in $\mathbb{R}^3$. The set is the intersection of ℓ_∞ and ℓ_1 norm balls in the positive orthant. The blue dots are vertices of the feasible set.

where λ is a constant that defines the level of sparseness and K is the number of basis vectors. However, this constraint is combinatorial in nature, hence difficult to optimize. In the context of machine learning [160] and optimization [35], the integer ($\{0,1\}^D$) and ℓ_0 constraints are relaxed with their convex surrogates:

$$\begin{aligned} \|\mathbf{b}\|_0 \leq \lambda &\rightsquigarrow \; \|\mathbf{b}\|_1 \leq \lambda \\ \mathbf{b} \in \{0,1\}^D &\rightsquigarrow \mathbf{0} \leq \mathbf{b} \leq \mathbf{1} \equiv \mathbf{b} \geq \mathbf{0}, \|\mathbf{b}\|_\infty \leq 1 \end{aligned} \tag{4.2.1}$$

where $\rightsquigarrow$ denotes a relaxation and $\equiv$ shows equivalence, $\|.\|_1$ is the ℓ_1-norm of a vector which is a convex relaxation of its ℓ_0-norm and $\leq$ is an element-wise inequality constraint. Geometrically, each basis vector, $\mathbf{b}_k$, dwells in the intersection of the ℓ_1-norm ball of radius λ with unit ℓ_∞-norm ball (box) in the positive orthant, which is shown graphically in Figure4.1 for $\mathbf{b} \in \mathbb{R}^3$ for sake of illustration. We call the feasible set the *Boxed-Sparsity* set, in contrast to a feasible set to be defined subsequently.

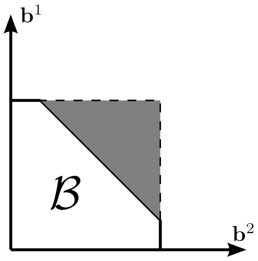


Figure 4.2: Presentation of a feasible set ($\mathcal{B}$) for $\mathbf{b} \in \mathbb{R}^2$.

4.2.1 Efficient Projections on the Boxed-Sparsity Balls

We need to repeatedly project on this set ($\mathbf{B}$) in the our optimization algorithm (alg.2). Therefore, having an efficient projection algorithm speeds up the optimization algorithm substantially. Euclidean projection operator on a feasible set can be viewed as an optimization problem:

$$\mathcal{P}(\mathbf{u}) = \arg\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|_2^2 \quad \text{s.t.} \quad \mathbf{z} \in \mathcal{B}$$

For Boxed-Sparsity, the problem is a constrained quadratic programming:

$$\begin{aligned} \min_{\mathbf{z}} \qquad & \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|_2^2 \\ \text{subject to:} & \, \mathbf{0} \le \mathbf{z} \le \mathbf{1} \\ & \mathbf{1}^T\mathbf{z} \le \lambda \end{aligned} \tag{4.2.2}$$

Geometrically, the projection point lies either on the boundary of the box in Figure4.2 or inside of the box, on the inside boundary of the shaded area in Figure4.2. To determine which one, we can simply project the point on the box:

$$\mathcal{P}_{\text{box}}(\mathbf{u}) = \min\{\mathbf{1}, [\mathbf{u}]_+\}$$

where $[\mathbf{u}]_+ = \max\{\mathbf{0}, \mathbf{u}\}$.

If $\mathcal{P}_{\text{box}}(\mathbf{u})$ still lies outside of the feasible set, it means that the projection point is on the inside boundary of the shaded area. To find the projection in this case, this problem should be solved:

$$\begin{aligned} \min_{\mathbf{z}} \quad & \frac{1}{2}\|\mathbf{u}-\mathbf{z}\|_2^2 \\ \text{subject to: } & \mathbf{0} \leq \mathbf{z} \leq \mathbf{1} \\ & \mathbf{1}^T\mathbf{z} = \lambda \end{aligned} \tag{4.2.3}$$

Lagrangian of Eqn.(4.2.3) is:

$$\begin{aligned} \mathcal{L}(\mathbf{z},\zeta,\theta,\eta) = \frac{1}{2}\|\mathbf{z}-\mathbf{u}\|_2^2 + \theta(\sum_{i=1}^{D} z_i - \lambda) \\ -\langle\zeta,\mathbf{z}\rangle + \langle\eta,\mathbf{z}-\mathbf{1}\rangle \end{aligned} \tag{4.2.4}$$

where $\theta \in \mathbb{R}$ and $\eta, \zeta \in \mathbb{R}_+^D$ are Lagrangian multipliers. Differentiating it with respect to $\mathbf{z}$ and setting it to zero, yields optimality condition: $\frac{\partial\mathcal{L}}{\partial z_i} = z_i - u_i + \theta - \zeta_i + \eta_i = 0$. By complementary slackness of KKT condition, we know whenever $z_i > 0$ then $\zeta = 0$ and whenever $z_i < 1$ then $\eta_i = 0$. Hence, if $0 < z_i < 1$ then:

$$z_i = u_i - \theta + \zeta_i - \eta_i = u_i - \theta \tag{4.2.5}$$

In order to determine optimal solution, z_i, we need to determine θ and indices for which z_i's are zero or one. If indices of ones and zeros of $\mathbf{z}$ are given, complementary slackness of KKT condition and the optimality conditions of Eqn.(4.2.3) suffices to find optimal θ:

$$\theta = \frac{1}{|\mathcal{I}|}(\sum_{i:z_i=1} 1 + \sum_{i\in\mathcal{I}} z_i - \lambda) \tag{4.2.6}$$

where $\mathcal{I} = \{i \in [n] : 0 < z_i < 1\}$ and $|\mathcal{I}|$ is cardinality of this set.

Following lemmas help us to determine the indices [1]:

Lemma 4.2.1. *[184] Let $\mathbf{z}$ be the optimal solution to the minimization in Eqn.(4.2.3). Let s and j be two indices such that $u_s > u_j$. If $z_s = 0$ then z_j must be zero as well.*

We will propose a similar lemma for the upper bound:

Lemma 4.2.2. *Let $\mathbf{z}$ be the optimal solution to the minimization in Eqn.(4.2.3). Let s and j be two indices such that $u_s > u_j$. If $z_j = 1$ then z_s must be 1 as well.*

Proof. The proof is by contradiction, similar to Lemma 4.2.1. Assume that $\mathbf{z}^*$ is optimal solution and there exist indices j and s such that $u_j < u_s$ and $z_j^* = 1$ but $z_s^* < 1$. Now, let us assume that new vector $\hat{\mathbf{z}}$ that is equal to $\mathbf{z}^*$ except in two indices j and s in which $\hat{z}_s = z_j^*$ and $\hat{z}_j = z_s^*$. It can be readily checked that $\hat{\mathbf{z}}$ is also feasible. The difference in objective value for new vector is:

$$\begin{aligned}
\|\mathbf{u} - \mathbf{z}^*\|_2^2 - \|\mathbf{u} - \hat{\mathbf{z}}\|_2^2 &= (u_j - z_j^*)^2 + (u_s - z_s^*)^2 \\
&\quad -(u_j - \hat{z}_j)^2 - (u_s - \hat{z}_s)^2 \\
&= -2u_j z_j^* - 2u_s z_s^* + 2u_j \hat{z}_j + 2u_s \hat{z}_s \\
&= 2z_s^*(u_j - u_s) + 2z_j^*(u_s - u_j) \\
&= 2(z_j^* - z_s^*)(u_s - u_j) \geq 0
\end{aligned}$$

which contradicts with optimality of $\mathbf{z}^*$. □

Given the lemmas, we can form an optimization problem similar to Eqn.(4.2.3). For a fixed θ, we solve the following optimization problem:

$$\begin{aligned}
\min_{\mathbf{z}} \quad & \frac{1}{2}\|(\mathbf{u} - \theta\mathbf{1}) - \mathbf{z}\|_2^2 \\
& \text{subject to: } \mathbf{0} \leq \mathbf{z} \leq \mathbf{1}
\end{aligned} \tag{4.2.7}$$

[1] Similar approach was adopted by Duchi *et al.* [73]

and then we search over θ such that the solution $\mathbf{z}$ satisfies the equality constraint in Eqn.(4.2.3). Observe that the term with θ in Eqn.(4.2.4) is absorbed into the quadratic term in Eqn.(4.2.7). However, Eqn.(4.2.7) has a closed form solution:

$$\mathbf{z}_\theta^* = \min\{\mathbf{1}, [\mathbf{u} - \theta\mathbf{1}]_+\} \tag{4.2.8}$$

Since we do not know the appropriate θ, we need to search for it. So far, optimization problem has simplified from D-dimensional to one dimensional problem. However, the two lemmas help us to find *exact* θ in *finite* number of iterations. The idea is to shrink $[\theta_{min}, \theta_{max}]$ with a bisection-type algorithm until number of zeros and ones stay unchanged, then θ can be found exactly with Eqn.(4.2.6). The details of the algorithm are shown in Alg.3. In Alg.3, $\mathcal{I} \leftarrow \{j \in [D] : 0 < z_j < 1\}$ and `ShiftInterval` is a function that accepts three real values arguments and returns two:

$$(\theta_3, \theta_2 + \frac{1}{2}(\theta_2 - \theta_1)) = \text{ShiftInterval}(\theta_1, \theta_2, \theta_3)$$

4.3 Group-Sparsity

Another interesting prior on $\mathbf{B}$ arises when a partition is available and needs to be taken into account. We assume a common coordinate system by warping all images to a template and an image partitioning (image segmentation) is available for the template image (*e.g.,* an anatomical parcellation in a template space). It is possible to consider sparsity constraint/regularization on the group-level rather than voxel level which promotes that a few groups (*e.g.,* brain structures) are involved in group difference rather than a few voxels. In order to encourage this property, we can enforce an ℓ_1-norm on groups instead of voxels. Before defining the idea precisely, we need a few definitions. Assuming $\mathcal{G}$ is a segmentation of an image into sets (g_i's), we can define

Algorithm 3 Efficient Projection on Boxed-Sparsity Ball

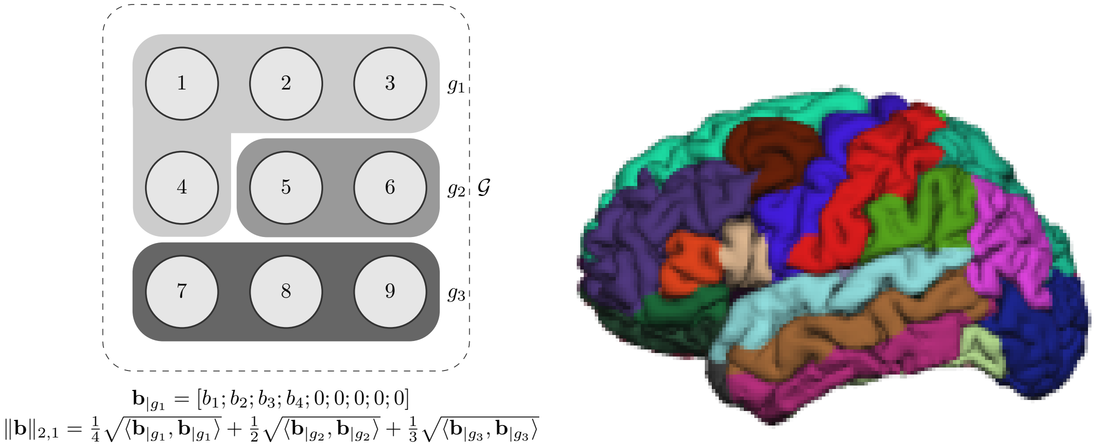
```
Require: Input $\mathbf{u}$, $\lambda$
  $\mathbf{z} \leftarrow \min\{\mathbf{1}, \max\{\mathbf{0}, \mathbf{u}\}\}$
  if $\mathbf{z}$ is infeasible then
    $\theta_1 \leftarrow 2\max_i z_i$;    $\theta_2 \leftarrow \min_i z_i$
    $\mathbf{y}_1 \leftarrow \min\{\mathbf{1}, [\mathbf{u} - \theta_1\mathbf{1}]_+\}$;    $\mathbf{y}_2 \leftarrow \min\{\mathbf{1}, [\mathbf{u} - \theta_2\mathbf{1}]_+\}$
    $\theta \leftarrow \theta_2 + \frac{1}{2}(\theta_2 - \theta_1)$
    while True do
      $\mathbf{z} \leftarrow \min\{\mathbf{1}, [\mathbf{u} - \theta\mathbf{1}]_+\}$
      if $\mathbf{1}^T\mathbf{z} > \lambda$ then
        $(\theta_2, \theta) \leftarrow$ shiftInterval$(\theta_1, \theta_2, \theta)$
        $\mathbf{y}_2 \leftarrow \mathbf{z}$
      else if $\mathbf{1}^T\mathbf{z} < \lambda$ then
        $(\theta_1, \theta) \leftarrow$ shiftInterval$(\theta_1, \theta_2, \theta)$
        $\mathbf{y}_1 \leftarrow \mathbf{z}$
      else
        return the $\mathbf{z}$
      end if
      if numbers of $\{0, 1\}$ of $\mathbf{z}$, $\mathbf{y}_1$, and $\mathbf{y}_2$ are unchanged then
        $\theta \leftarrow \frac{1}{|\mathcal{I}|}(\sum_{z=1} 1 + \sum_{i\in\mathcal{I}} z_i - \lambda)$;    $\mathbf{z} \leftarrow \min\{\mathbf{1}, [\mathbf{u} - \theta\mathbf{1}]_+\}$
        return $\mathbf{z}$
      end if
    end while
  else
    return $\mathbf{z}$
  end if
```

two *group-norms* as follows (the idea is graphically shown in Figure4.3):

$$\begin{aligned} \|\mathbf{b}\|_{1,2} &:= \sum_{g\in\mathcal{G}} \rho_g \|\mathbf{b}_{|g}\|_2 \\ \|\mathbf{b}\|_{\infty,2} &:= \max_{g\in\mathcal{G}} \{\rho_g \|\mathbf{b}_{|g}\|_2\} \end{aligned} \tag{4.3.1}$$

where $\mathbf{b}_{|g}$ is a D-dimensional vector such that its voxels not belonging to the group g are set to zero, ρ_g is a positive constant that compensates for a group-size, namely $\rho_g = \frac{1}{|g|}$ where $|\cdot|$ is cardinality of a set. Notice that in the definition of $\|\cdot\|_{1,2}$, the ℓ_2-norm is used instead of ℓ_2^2 because the squared norm does not have the sparsifying properties. This kind of regularization is called *Group* regularization or *Mixed-Norm* regularization and have received much attention in recent years in machine learning [163], [112].

Given the new norm definitions in *Eq.*(4.3.1), we can define the *Group-Sparsity* constraint


1
2
3
g_1
4
5
6
g_2
$\mathcal{G}$
7
8
9
g_3
$\mathbf{b}_{|g_1} = [b_1; b_2; b_3; b_4; 0; 0; 0; 0; 0]$
$\|\mathbf{b}\|_{2,1} = \frac{1}{4}\sqrt{\langle \mathbf{b}_{|g_1}, \mathbf{b}_{|g_1} \rangle} + \frac{1}{2}\sqrt{\langle \mathbf{b}_{|g_2}, \mathbf{b}_{|g_2} \rangle} + \frac{1}{3}\sqrt{\langle \mathbf{b}_{|g_3}, \mathbf{b}_{|g_3} \rangle}$

(a) *(b)*

Figure 4.3: (a) shows an example of a 3×3 image (hence $\mathbf{b} \in \mathbb{R}^9$) that is segmented into three regions ($\mathcal{G} = \{g_1, g_2, g_3\}$). $\mathbf{b}_{|g_1}$ and $\|b\|_{2,1}$ are shown as examples. $\langle \cdot, \cdot \rangle$ means inner product thus $\|\mathbf{b}_{|g_1}\|_2 = \sqrt{\langle \mathbf{b}_{|g_1}, \mathbf{b}_{|g_1} \rangle}$; (b) shows an example of grouping (*i.e.*, segmentation) for medical imaging applications.

mathematically as follows:

$$\|\mathbf{b}\|_{1,2} \leq \lambda$$
$$\mathbf{b} \geq 0, \|\mathbf{b}\|_{\infty,2} \leq 1 \tag{4.3.2}$$

For the rest of the chapter, we will refer to $\|\mathbf{b}\|_{1,2}$ subject to the constraints as *Group-Sparsity*. Observe the correspondence between Boxed- and Group-Sparsity: comparing *Eq.*4.2.1 and *Eq.*(4.3.2), $\| \cdot \|_{1,2}$ replaced $\| \cdot \|_1$ and $\| \cdot \|_{\infty,2}$ exchanged for $\| \cdot \|_\infty$.

4.3.1 Efficient Projection on Group-Sparsity Ball

Given Alg.(3), efficient projection on a Group-Sparsity ball is very simple because it uses Alg.(3) as a submodule. An algorithm for efficient projection on a Group-Sparsity ball is shown in Alg.(4). In this case, the following optimization problem should be solved:

$$\min_{\mathbf{z}} \qquad \frac{1}{2}\|\mathbf{u} - \mathbf{z}\|_2^2$$

$$\text{subject to:} \quad \mathbf{1}^T\mathbf{t} \leq \lambda$$

$$\rho_g \|\mathbf{z}_{|g}\|_2 \leq t_g, \forall g \in \mathcal{G}$$

$$\mathbf{z} \geq \mathbf{0}, \mathbf{t} \geq \mathbf{1} \tag{4.3.3}$$

where $\mathbf{t}$ is a positive $|\mathcal{G}|$-dimensional vector and t_g is g'th element of that and ρ_g is a constant. Eqn.(4.3.3) ia a Second Order Cone Programming (SOCP) and may look significantly different from Eqn.(4.2.2) but a careful inspection reveals that an efficient algorithm to solve Eqn.(4.2.2) (Alg.(3)) can help us to solve Eqn.(4.3.3) by defining:

$$\mathbf{v} \in \mathbb{R}^{|\mathcal{G}|}, \quad v_g = \rho_g \|[\mathbf{u}_{|g}]_+\|_2$$

The defined $\mathbf{v}$ can be provided as input to Alg.(3) to find a projection in $\mathbb{R}^{|\mathcal{G}|}$ space. Given the projected point, simple rescaling yields optimal $\mathbf{z}$. The procedure is explained in Alg.(4).

Algorithm 4 Efficient Projection on Group-Sparsity Ball

```
Require: Input $\mathbf{u}$, $\lambda$
  if $\|[\mathbf{u}]_+\|_{1,2} > \lambda$ then
    Form vector $\mathbf{v}$ as follows: $v_g = \rho_g \|[\mathbf{u}_{|g}]_+\|_2$
    $\mathbf{t} \leftarrow$ ProjectBoxedSparsity($\mathbf{v}, \lambda$) (Alg.(3))
    for all $g \in \mathcal{G}$ do
      $\mathbf{z}_{|g} \leftarrow \frac{t_g}{v_g}\mathbf{u}_{|g}$
    end for
    return $\mathbf{z}$
  else
    return $\mathbf{z}$
  end if
```

Recently there have been a few research papers about efficient projection on the group-sparsity ball for arbitrary definition of the groups. Although it has been shown that projection on group-sparsity ball for arbitrary group is possible [118], it is an expensive operation unless some special structures are assumes for the groups [128] (*e.g.,* tree structure).

$$\|\mathbf{b}\|_{1,\infty} \leq \lambda$$

$$\mathbf{b} \geq 0, \|\mathbf{b}\|_{\infty,2} \leq 1$$

4.4 Other Possibilities for the Feasible Set

In section, we address other possibilities for $\mathcal{B}$ which are not explored in this thesis but can be used depending on application. In Section 4.3, we chose ℓ_2 norm to group pixel together. Obviously, it is not the only option. For example, ℓ_∞ can be used to group voxel together, namely:

$$\|\mathbf{b}\|_{1,\infty} := \sum_{g \in \mathcal{G}} \rho_g \|\mathbf{b}_{|g}\|_2$$
$$\|\mathbf{b}\|_{\infty,\infty} := \max_{g \in \mathcal{G}} \{\rho_g \|\mathbf{b}_{|g}\|_\infty\}$$

Comparing to *Eq.*4.3.1, using this definition of group sparsity drives the maximum value of $\mathbf{b}_{|g}$ to zero. Within g'th group, $\mathbf{b}_{|g}$ tend to choose values close the maximum because only the maximum value is penalized. One potential advantage of using such definition is that resultant optimization problem is a constrained Quadratic Programming (QP) which is computationally less expensive than to SOCP in Section 4.3. Nevertheless, we showed in Section 4.3.1, for non-overlapping groups the SOCP can be computed efficiently.

Regardless of choice of the norm for grouping voxel, some application may demand overlapping groups. An example is when priors are provided as regions of interest (ROI's) that are connected through fiber tracking algorithm. A conceivable prior is that areas connected with white matter fiber track are more likely to fire simultaneously during resting-state fMRI experiments Figure4.4. If ℓ_2 norm is chosen to group the voxels, overlapping groups makes solving the proximal operator computationally expensive. Assuming that ℓ_∞ is chosen for grouping, Marial *et al.* [153] has recently shown that the proximal operator can be solved efficiently using network flow algorithm.

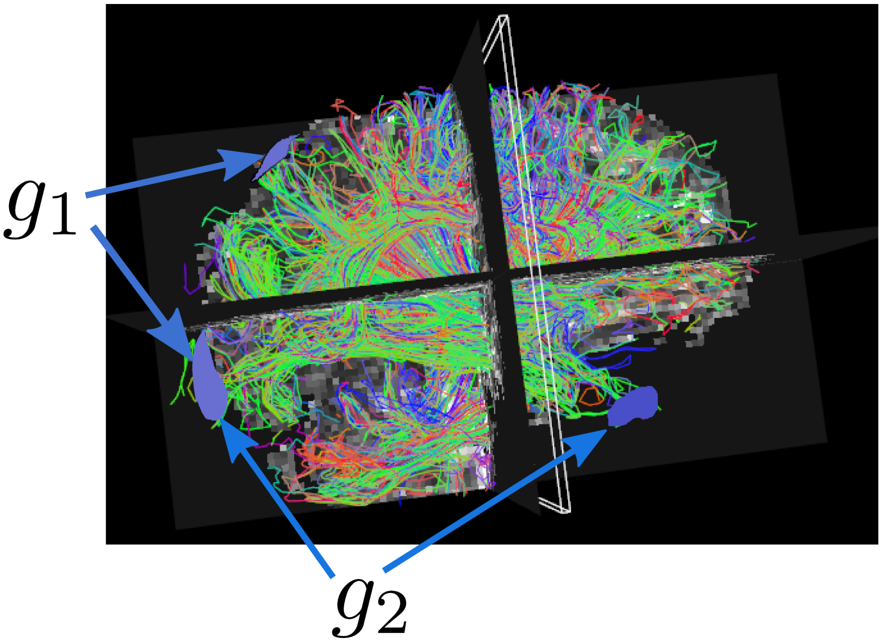


Figure 4.4: An example of applications of overlapping groups for definition of group norm: groups are defined by areas of brain that are connected through white matter fiber tracks.

4.5 On Selection of the Regularization Parameters

To set values of the parameters (*i.e.,* λ's and r), two strategies are available: first, to embed searching for the best parameters as a part of the training of the algorithm. This strategy is chosen to show the results in this chapter; second, to set values of the parameters to pre-defined values which are presumed to perform well. Ideally, the first option is preferred because it potentially yields better performance than setting parameters to pre-defined values, however, the large optimization with respect to $(\mathbf{B}, \mathbf{C}, \mathbf{w})$ renders searching an expensive task. Although the latter strategy is not investigated in this chapter, we will give intuition on how to select parameters to some fixed values.

Parameters of the proposed algorithm are as follows: K number of basis vectors; λ_1, the weight for the generative term; λ_2, the weight for the discriminative term; λ_3, the sparsity ratio for the basis vectors. We propose to choose the parameters in the following order:

1. λ_2: Given *Eq*.3.5.2 and *Eq*.3.3.3, it can be readily derived that $\frac{N}{\lambda_2}$ defines the weight for the

second term in *Eq.*3.5.2 ($\|\mathbf{w}\|_2^2$). One suggestion is to run the algorithm for a small-scale dataset for a few iterations and choose λ_2 such that it produces a reasonable classification rate. One can even run the algorithm for a few iterations without the discriminative term and extracts feature (*i.e.,* $\mathbf{B}^T\mathbf{x}_i$) in order to have a sense of an appropriate range for λ_2.

2. K and λ_3: Selection of λ_3 can be inspired by our clinical hypothesis; $\frac{\lambda_3}{D}$ approximately sets the non-zero ratio of each basis vector. Depending on our clinical expectations regarding portion of an anatomy (*e.g.,* brain) affected by the disease of interest, we can choose a range for λ_3. However, if sparseness is set to a high value (low λ_3/D), the generative term may not be able to represent the data well because it may not be able to cover the whole domain of images; hence, optimal basis vectors may stay away from the boundaries of the feasible set (where basis vectors achieve 0-1 values) while the model may try to compensate with $\mathbf{C}$ to reconstruct the data. In fact, there is a limited *budget* to reconstruct the data. In order to increase the budget, one can increase the number of basis vectors (K). However, a very large value of K increases the computational cost significantly, so one needs to trade off between excessive sparsity and computational cost. There are also other factors involved in choosing the sparsity ratio that will be discussed in Section 4.6.

3. λ_1: Once other parameters are set, we can set a value for λ_1. The ratio λ_2/λ_1 decides the balance between the generative and the discriminative terms; since λ_2 is already set, one needs to choose the ratio of λ_2/λ_1. As it will be shown in Section 3.6.3, the algorithm is relatively robust with respect to ratio of λ_1/λ_2 as long as λ_1 is in a reasonable range; hence the value of λ_1 should be chosen such that the first and second terms in in the objective of the optimization have similar magnitudes.

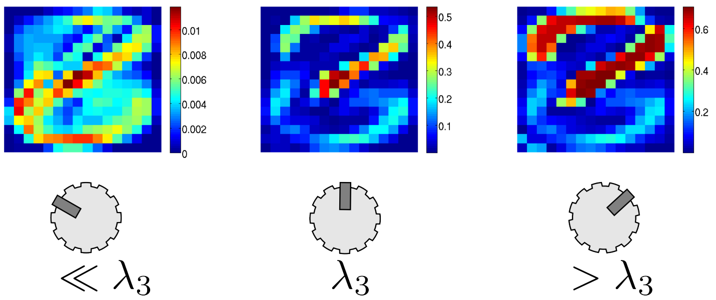


Figure 4.5: The figure shows $\mathbf{B}|\mathbf{w}|$ as a qualitative measure of how well the algorithm can delineate area of difference between "6" and "8"; pixels with hot colors are presumably more discriminative. The knobs on the bottom of figures symbolize the sparsity parameter (λ_3). If $\lambda_3 \to 0$, it yields $\mathbf{B}|\mathbf{w}| \to 0$ but less sparse result. Increasing λ_3 beyond some level does not change the patter significantly.

4.6 Experiments

4.6.1 Sparsity and Detecting Discriminative Area

In order to illustrate the effect of sparsity parameter λ_3, 100 images of "6" and "8" were selected randomly to form the USPS handwritten dataset [106], [113]. The discriminative to the generative ratio (λ_2/λ_1) is set a reasonable value (see Section 3.6.2 for discussion), number of basis vectors is set to 16 ($K = 16$) and we vary the sparsity parameter (λ_3) over a wide range ($\lambda_3 \in (0, D)$, where $D = 16$ is number of pixels) to study the effect of sparsity parameter.

Figure4.5 shows three examples of $\mathbf{B}|\mathbf{w}|$ computed for three different values of λ_3. $\mathbf{B}|\mathbf{w}| = \sum_{k=1}^{K} \mathbf{b}_k |w_k|$ where $|w_k|$ denotes the absolute value of the k'th element of $\mathbf{w}$. This measure can be viewed as a qualitative measure for how well the algorithm can delineate area of difference between two characters; if the algorithm dedicates some of basis vectors to minimize $\ell(\cdot, \cdot)$, they should high contribution in the loss function and hence have values for $|w_k|$. The figure in the middle corresponds to $\lambda_3 = 0.2D$ which is a reasonable value. Increasing λ_3 (less sparsity) does not change $\mathbf{B}|\mathbf{w}|$ significantly but it affects classification accuracy. Counter-intuitively, making λ_3 too small does not make $\mathbf{B}|\mathbf{w}|$ too sparse. In fact, as $\lambda \to 0$, $(\mathbf{B}|\mathbf{w}|) \to 0$ as expected but since columns of $\mathbf{B}$ have limited budget to represent images, $\mathbf{b}_k$ stay away from the boundaries of the feasible set which result in non-sparse basis vectors. The figure shows that there is a lower bound for λ_3 in term of finding the correct discriminative area. We will see that classification accuracy decides about the upper bound on λ_3.

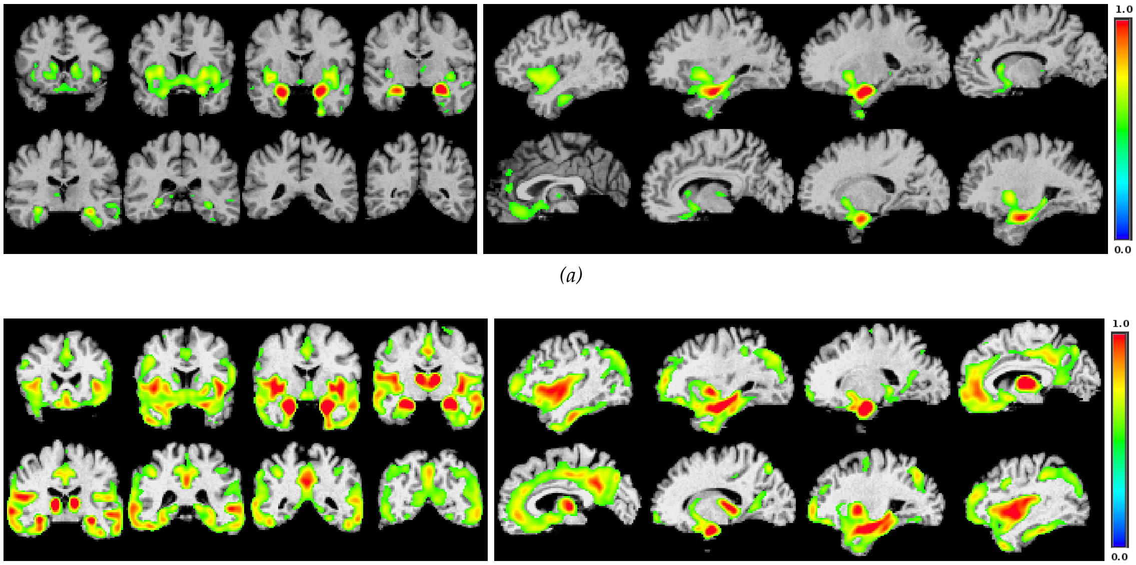
4.6.2 Sparsity and Classification Accuracy

In this section, we study how the sparsity parameter, λ_3, affects classification accuracy for Boxed- and Group-Sparsity feasible sets. The images used in this experiment are structural MR brain images (`T1` image) obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI[2]). 63 normal control (NC) individuals and 54 AD patients were pre-processed. The pre-processing pipline is exactly the same as what explained in Section 3.6.3. For experiments in this section, we set $K = 30$ that shows reasonable performance in Section 3.6.3 (see Figure3.11) and we changed λ_3 over a wide range for various ratios of λ_2/λ_1.

Figure4.6 compares examples of basis vectors for two different sparsity ratios. Increasing λ_3 from 20% to 10% of voxels yields sparser and more localized basis vectors. As discussed in Section 4.6.1, decreasing λ_3 which enforces stricter sparsity constraint (say $\lambda_3/D = 0.1\%$) may not be helpful for better representation because as λ_3 decreases, the algorithm has a limited budget of voxels (*i.e.,* few voxels can be selected) to satisfy the generative term ($\mathcal{D}(\cdot;\cdot)$); therefore it prefers to push values of the voxels away from boundaries (*i.e.,* $\{0, 1\}$) to satisfy the generative term. Nevertheless, we changed λ_3/D in range of $[0.1..0.6]$ to examine its effect on the classification accuracy (Figure4.7). The experiment elaborated in Section 3.6.3 is repeated but for different values of λ_3/D and λ_2/λ_1. The settings of the experiment in term of number of samples and pre-processing is identical with those of the experiments in Section 3.6.3.

Figure4.7 shows comparison of different ratios of λ_3/D for the Boxed-Sparsity for different rates of λ_2/λ_1. Since two types of behaviors are observed, they are shown in two separate graphs for a sake of illustration. Figure4.7a shows cases in which the generative term is dominant or moderate while Figure4.7b shows graphs in which the discriminative term is dominant.

[2] www.loni.ucla.edu/ADNI



(a)

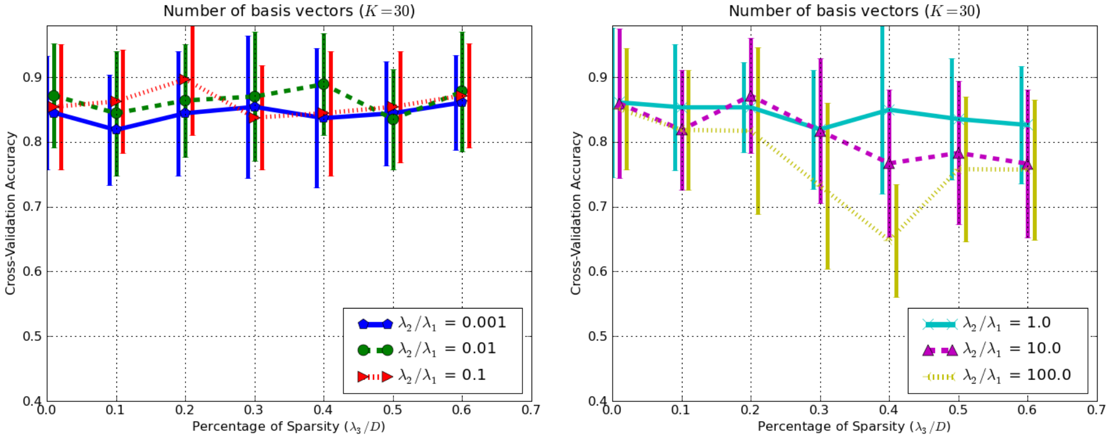
(b)

Figure 4.6: The figure shows examples of basis vectors for two different values of the sparsity parameter: (a) ($\lambda_3 = 0.1D$) on coronal and sagital views; (b) ($\lambda_3 = 0.2D$) on coronal and sagital views.

In Figure4.7a, increasing λ_3 (less sparse) slightly improves level of classification accuracy up to a certain point ($\lambda_3/D \in [0.2, 0.4]$ depending on the ratio $\frac{\lambda_2}{\lambda_1}$) because it yields better reconstruction. However from that point on, it decreases because it means less regularization on the model. Nevertheless, if the generative term is dominant, the algorithm is relatively robust.

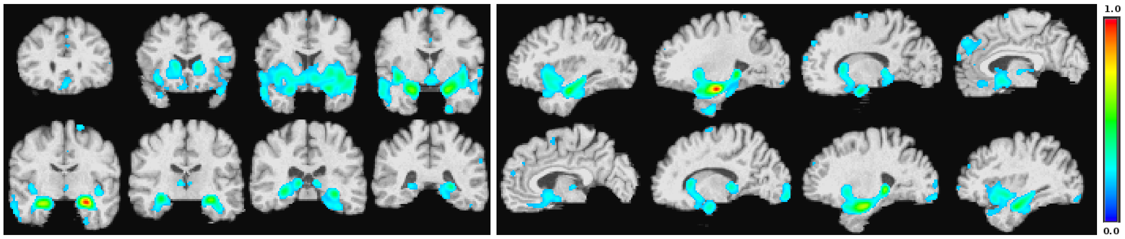
Figure4.7b shows similar graph for the cases in which the discriminative term is dominant or has relatively higher weight than those of Figure4.7a. In this case, increasing λ_3 (decreasing sparsity) deteriorates the classification accuracy. When the discriminative term is dominant, reducing sparsity can approximately be compared to ℓ_1-SVM with small regularization weight; excessive reduction of the regularization weight in ℓ_1-SVM can worsen generalization of the classifier.

Figure 4.8 shows an example of a basis vector when Group-Sparsity is used. The feasible set of the Group-Sparsity is smoother than that of the Boxed-Sparsity (Figure4.6); in other words, it has fewer sharp corners than the Boxed-Sparsity one. This encourages solutions that are smooth, *i.e.*, voxel values are likely to be in $(0, 1)$ rather than 0 or 1. Nevertheless such behavior is also affected by ℓ_2-norm of the samples (*i.e.*, normalization of samples) that are not discussed in this



(a) *(b)*

Figure 4.7: Investigation of sparsity level on the classification accuracy for the Boxed-Sparsity when: (a) the generative term is dominant; (b) the discriminative term is dominant. Standard deviations of the accuracy rates are added as the bars to the figures.

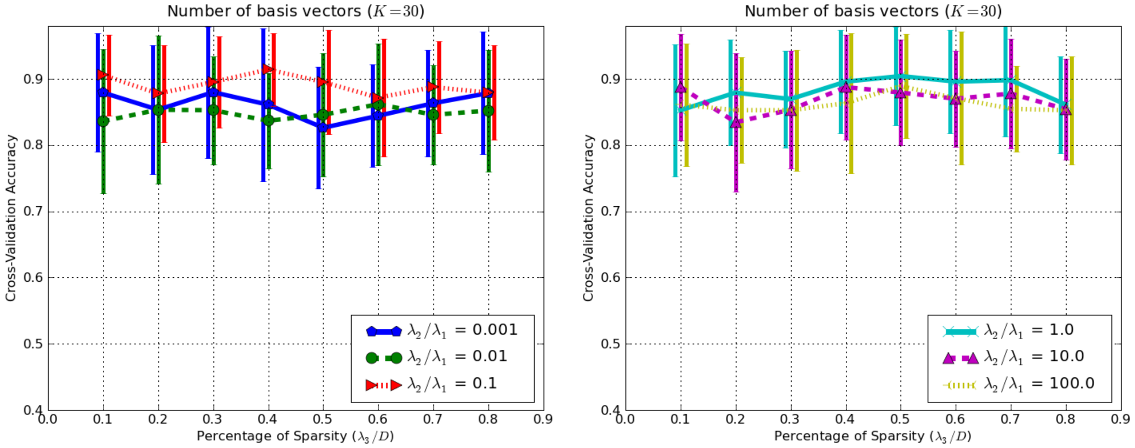


(a) *(b)*

Figure 4.8: An example of a basis vector for a case in which Group-Sparsity constraint is used. (a) coronal cuts; (b) sagittal cuts.

chapter in interest of space.

Figure4.9, depicts the same graphs as Figure4.7 but for Group-Sparsity regularization. As in Figure4.7, the graphs are divided into two (generative- or discriminative- dominant) sub-graphs for a sake of better illustration. In term of maximum accuracy, the Group-Sparsity is comparable with the Boxed-Sparsity (about 3% improvement) but it is more robust with respect to change of parameters; Figure 4.8a shows perturbation is accuracy that is about 5% across different settings. In Figure4.9b, the Group-Sparsity shows significantly more robust behavior when the discriminative term is dominant comparing to Figure4.7b. Such robustness can be explained by definition of the Group-Sparsity regularization. Due to the non-linear relationship within each group,



(a)

(b)

Figure 4.9: Investigation of sparsity level on the classification accuracy for the Group-Sparsity when: (a) the generative term is dominant; (b) the discriminative term is dominant. Standard deviations of the accuracy rates are shown as error bars.

Group-Sparsity imposes fewer degrees of freedom than those of Boxed-Sparsity, therefore it regularizes the objective further. Figure4.9b also shows that a reasonable range for Group-sparsity is around $\frac{\lambda_3}{D} \in [0.4, 0.7]$ which is different that that of the Boxed-Sparsity; the accuracy rates slightly degrade after this range.

4.6.3 Comparison with Other Methods

In this section, we compare performance of the proposed algorithm with other methods but first we need to clarify some points about parameter selection (λ's). The dataset is divided into 20 splits, 18 splits are used to learn $(\mathbf{B}, \mathbf{C}, \mathbf{w})$ and the testing accuracy on one of the two left-out splits is used to search for the best λ's and finally the classification accuracy is reported on the other left-out split.

Table 4.1 compares the accuracy rates between five different methods (two of them are variants of the proposed method) on two dataset. `Bx` and `Grp` stand for the proposed for Boxed- and Group-Sparsity constraints respectively. Singular Value Decomposition (SVD) and Non-negative Matrix Factorization were added to the table in order to have baseline comparisons. In order to

Table 4.1: Comparison of the classification accuracy rate of the proposed method using two different constraints Boxed-(Bx) and Group-(Grp) with other methods: Singular Value Decomposition (*SVD*), Non-negative Matrix Factorization (*NMF*) and COMPARE [80]. AD vs NC is Alzheimer's disease verse Normal Control from ADNI dataset and Lie vs Truth is β-maps of fMRI study for lie detection. The values inside of the parentheses are the standard deviations of the accuracy rates.

	AD vs NC	Lie vs Truth
Bx	86.6%(±14.3%)	84.1%(±20%)
Grp	**89.0%(±13.3%)**	N/A
SVD	74.2%(±19.3%)	72.5%(±21%)
NMF	62.1%(±16.3%)	55.0%(±10%)
COMPARE	86.7%(±15.3%)	**88.3%(±16.3%)**

have a fair comparison, number of basis vectors for NMF, SVD, and both variants of the proposed method are set to the same number which is 30. COMPARE is a method proposed by Fan *et al.* [80] and has shown to perform well on ADNI dataset [79].

While features extracted from NMF and SVD methods were fed to the same procedure as the proposed method to find the best classifier, COMPARE has it own routine to find an optimal classifier. AD vs NC dataset is already explained in the beginning of this section. Lie vs Truth contains 22 subjects performing a forced-choice deception and their brain activations were acquired using BOLD imaging (fMRI). SPM2 software [2] is used to calculate Parameter Estimate Images (PEIs), *i.e.,* regression coefficients or β, of the HRF regressors for each of the 50 conditions from the least mean square fit of the model to the time series. The 50 conditions include forty-eight regressors modeled "lie" and "truth" events individually while two additional regressors modeled the variant distracter and recurrent distracter conditions.

In the Table 4.1, while the Group-sparsity regularization outperforms COMPARE, the Boxed-sparsity performs almost as well as COMPARE on the AD vs NC dataset. On the Lie vs Truth dataset, COMPARE outperforms our method although the Boxed-sparsity is in a reasonable range of the best performance. The Group-Sparsity result for fMRI dataset is shown as "N/A" because fMRI images which are pre-processed with SPM2 are registered to SPM2 atlas with *affine* transformation. Therefore, structural brain regions of the atlas do not match well with the corresponding

Table 4.2: Comparison of the proposed method using two different constraints, *i.e.*, the Boxed-(Bx) and Group-(Grp) Sparsity with other methods: Singular Value Decomposition (*SVD*), Non-negative Matrix Factorization (*NMF*) and COMPARE [80]. AD vs NC is Alzheimer's disease verse Normal Control from ADNI dataset and converter versus non-converter MCI subjects (MCI-C vs MCI-NC). The values inside of the parenthesis are the standard deviations of the accuracy rates.

	AD vs NC	MCI-C vs MCI-NC
Bx	**84.2%(±8.3%)**	60.7%(±9.4%)
Grp	83.7%(±8.6%)	**61.5%(±8.3%)**
SVD	70.9%(±14.1%)	57.3%(±2.9%)
NMF	71.8%(±14.7%)	53.5%(±7.8%)
COMPARE	82.2%(±7.4%)	59.4%(±10.5%)

regions on the individual subjects that makes the definition of the groups in the Group-Sparsity inaccurate.

The values reported in the Table 4.1 for the AD vs NC dataset are in the same range as the accuracy rates reported in [58]; Nevertheless the conditions of the experiments (including pre-processing, features extraction, samples in the training and testing lists, *etc.*) are different, which make the results not one-to-one comparable.

4.6.4 Sensitivity Analysis of the Parameters

In this section, we perform a few experiments to investigate the effect of parameter selection (λ's) on the classification accuracy rates. In this section, instead of optimizing λ's, we set λ's to the most frequently chosen ones in the Section 4.6.3. The MCI subjects were not involved in the experiments of the Section 4.6.3. In addition, we held out 205 AD and NC subjects (89 AD and 114 NC) from the ADNI dataset. Therefore, optimizing λ's in the Section 4.6.3 is oblivious with respect to the samples used in this section. In addition to the AD versus NC classification, we have included classification between converter and non-converter MCI subjects to the Table 4.2 which is known to be a difficult classification problem [58]. In fact, this experiment shows conservative results for the proposed methods.

As the Table 4.2 shows, the proposed method outperforms other methods on both datasets.

The classification rates are relatively low on the `MCI-C vs MCI-NC` dataset as reported in the literature [58] yet the proposed method shows slightly better performance comparing to other methods in the Table. This experiment shows that as long as the datasets are similar, one can reduce the computational cost of optimizing λ's by removing the extra nested loop for parameter selection (*i.e.,* searching for the best λ's inside of training sets) without significant degradation in the performance of the classifiers.

4.7 Conclusion and Discussion

In this chapter, we introduced various possible feasible sets for basis vectors. Different applications may impose various priors which lead to different definition of feasible sets, two of which were discussed here namely Boxed-Sparsity and Group-Sparsity. Boxed-Sparsity simply enforces sparsity in the voxel level neglecting relationship between voxels in the image domain. Group-Sparsity assumes a segmentation (partitioning) exists and enforces sparsity on groups of voxels and implicitly considers relationship between voxels in the image domain. Mathematical definition of the feasible set defined by Boxed-Sparsity was reduced to intersection of ℓ_∞-ball and ℓ_1-norm ball in the non-negative orthant. For Group-Sparsity, ℓ_∞ and ℓ_1 were replaced with their group-norm counterparts. The proposed optimization in Section 3.5 requires to project on the feasible set in each iteration; therefore if the projection is time consuming, it will render the whole algorithm very inefficient. We first proposed an efficient procedure for Boxed-Sparsity projection in Section 4.2.1. This procedure was used as subroutine to project on the Group-Sparsity set in Section 4.3.1.

We have also experimented with other types of regularizers in order to incorporate relationship between voxels [20]. We realized that for TV_2^1-norm (see 2.1), there is no significant difference in results if the images are pre-smoothed. Nevertheless, TV_1^1- or $TV_2^{1/2}$-norms (see 2.1) are not equivalent to pre-smoothing operation but projection algorithms on feasible sets defined by such norms are computationally expensive. Recently Fadili *et al.* [77] proposed a first order method for projection on $TV_2^{1/2}$-norm but given that the projection should be repeated in every iteration of SPG, it renders the optimization algorithm very slow.

In the experiment section, it was show that the algorithm is robust with respect to choice of parameters as long as they are chosen within a reasonable range. It also shows that the generative term is helpful; indeed we have observed in our experiments that in the process of searching for the best λ's, those settings biased toward the generative terms are selected quite frequently. The experiments shows that discriminative term is also essential because in its absence, the formulation becomes more or less similar to NMF [141] formulation which is shown to underperform in Table 4.1. Nevertheless, for very large sample size experiments finding optimal parameters might be computationally expensive. Therefore, in Section 4.5, we analyzed the role of each parameter in well-possessedness of the objective function and introduced an intuitive sequence to pick λ's within a reasonable range. In addition, we empirically showed in the Section 4.6.4 that as long as datasets are similar one can avoid parameter selection without significant degradation in the accuracy rate.

In Section 4.6.3, we also compared the proposed method with PCA and NMF as baseline methods and COMPARE [80] as the state-of-the-art algorithm. Both variants of the proposed method outperformed the baseline methods (*i.e.,* NMF and PCA) and performed better or almost as well as COMPARE. The Group-sparsity achieved the best performance in `AD vs NC` but it was not applicable to `Lie vs Truth` because we defined the groups for the Group-sparsity based on a segmentation of an atlas and all fMRI subjects are brought to the atlas space using only affine registration; it yields inaccurate brain segmentation for each subject and consequently inaccurate definition for the groups. It is also worth mentioning that COMPARE achieves such level of accuracy using 150-250 features while our algorithm uses only 30 basis vectors (*i.e.,* number of features). There is no clear winner between the Group- and the Box-sparsity.

Part II

Extensions

Chapter 5

Application for Multi-Channel Imaging

5.1 Introduction

This chapter presents a general discriminative dimensionality reduction framework for multi-channel image-based classification in medical imaging datasets. The major goal is to use all channels simultaneously to transform very high dimensional images to a lower dimensional representation in a discriminative way. In addition to being discriminative, the proposed approach has the advantage of being clinically interpretable.

We propose a framework based on regularized tensor decomposition. We will show that different variants of tensor factorization imply various hypothesis about data. Inspired by the idea of multi-view dimensionality reduction in machine learning community, two different kinds of decomposition will be presented and their implications will be discussed in this chapter. We have validated our method on different datasets including a multi-channel longitudinal brain imaging study. We compared this method with a state-of-the-art classification software based on purely discriminative feature reduction (COMPARE [80]).

Over recent years, emphasis of modern medical image analysis in context of diagnosis has shifted toward developing new biomarkers. Recently, various structural (*e.g.,* MRI, DTI, *etc.*) and functional (*e.g.,* PET, resting state fMRI, *etc.*) imaging channels have been utilized to develop new biomarkers for diagnosis. Multiple image channels can provide a rich multi-parametric signature that can be used to design more sensitive biomarkers [136], [108]. For example, while structural MR images provide sensitive measurements for detection of atrophy in brain regions [83], recent studies [66] have shown FDG-PET[1] can quantify reduction of glucose metabolism in parietal lobes, the posterior cingulate, and other brain regions [66]; combination of both channels can be very instrumental in early diagnosis of Alzheimer's disease [82].

An immediate solution to exploit multiple channels is to concatenate all image channels into a long vector, but learning a classifier that generalizes well in such a high dimensional space is even harder than in the uni-channel case because multi-channel datasets tend to be small. Therefore, dimensionality reduction plays an even more important role here. Most existing studies extract features from a few predefined areas [136]. Zhang [235] suggested extracting features from a few pre-defined regions of interest (ROIs) and combining them into one kernel that then input to a kernel-SVM classifier. However, predefined regions might not be optimal for diagnosis on the individual level, *i.e.,* classification of subjects into normal and abnormal groups. Ideally, the whole image (*e.g.,* brain scan) should be viewed as a large dimensional observation and relevant regions to the target variable of interest (class labels, here) should be derived from such high dimensional observation. High-dimensional pattern classification methods have been proposed for morphological analysis [80], [92] which aim to capture multivariate nonlinear relationships in the data. A critical step underlying the success of such methods is effective feature extraction and selection, *i.e.,* dimensionality reduction. In Chapter 3, we proposed a constrained matrix factorization framework for dimensionality reduction while simultaneously being discriminative

[1] fluorodeoxyglucose positron emission tomography

and representative; however, that method only works for uni-channel cases.

One could concatenate all image channels of a subject into long columns of a matrix and simply apply the method in Chapter 3 or a similar method. However, the straightforward approach is limited with respect to its ability to model different hypotheses regarding the data. In this chapter, we extend the formulation proposed in Chapters 3 and 4 by viewing the data matrix as tensor. In our matrix notation, the first and the second indices enumerate voxels and subjects respectively; we introduce new index that enumerates channels; it extends the data matrix to a tensor. The proposed method here is not exactly tensor factorization as defined in [47], [130]. However, the advantage of viewing the data as a tensor is that the tensor structure allows us different decompositions which imply various hypotheses about data. The proposed method is inspired by the *multi-view* setting in the machine learning community [126], [8]. In the multi-view setting, there are various views, sometimes in a rather abstract sense, of the data which co-occur; here views are multiple channels. There are also target variables of interest (*e.g.,* class labels). The goal is to learn the target via the relationship between different views [126]. In this chapter, we introduce two factorizations and explain their connotations. One of the variants is more appropriate for a setting that all channels focus on the same tissue type; for example PET and T1 which both focus on gray matter tissue. The other variant is more applicable for channels characterizing different tissue types; for example DTI and T1 which characterize white and gray matter tissues respectively. We also view fMRI resting-state data an an instances of multi-channel image; in this case, each time snapshot of brains, which is a volumetric image, is viewed as a channel. We derive the factorization by solving a large scale optimization problem.

In the section 5.2.1, we briefly review the framework introduced in the Chapters 3 and 4. In the section 5.2.2, two variants of the extensions for multi-channel cases are presented for medical image classification purposes. The section 5.2.3 shows how resting-state fMRI can be viewed as multi-channel image. We explore the applicability of the method for discovery of the so-called *default-mode-network* (DMN). By extending the notion of group sparsity introduced in the

Chapter 4, we show how structural connectivity can be used a prior to guide inference of the DMN. In the section 5.3.1, we apply the methods introduced in the sections 5.2.2 on real datasets for classification. Since we do not have a solid ground-truth for DMN, the method is tested on a synthesized data in the section 5.3.2. Finally, we explore the applicability of the method on a real resting-state data in the section 5.2.3 and compare the results with what is reported in the clinical literature.

5.2 Method

5.2.1 General Framework

The novel method proposed in this chapter is based on an extension of the previously proposed framework for uni-channel in Chapters 3 and 4, which we briefly present here for perspective. Similar to Chapters 3, the proposed method reduces the dimensionality in a discriminative way while preserving the semantics of images; hence it is clinically interpretable and produces good classification accuracy. We use regularized matrix factorization formalism for dimensionality reduction. Regularized matrix factorization decomposes a matrix into two or more matrices such that the decomposition describes the matrix as accurately as possible. Such a decomposition could be subjected to some constraints or priors. Let us assume that the columns of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]$ represent observations (*i.e.*, sample images that are vectorized), and $\mathbf{B} \in \mathbb{R}^{D \times K}$ and $\mathbf{C} \in \mathbb{R}^{K \times N}$ decompose the matrix such that $\mathbf{X} \approx \mathbf{BC}$. K is the number of basis vectors, which is a parameter of the algorithm, D is the number of voxels in images and N is the number of samples. The columns of matrix $\mathbf{B}$ (called $\mathbf{b}_k$) can then be viewed as basis vectors and the n^{th} column of $\mathbf{C}$ (called $\mathbf{c}_n$) contains corresponding loading coefficients or weights of the basis vectors for the n^{th} observation. The columns $\mathbf{b}_k \in \mathcal{B}$ and $\mathbf{c}_n \in \mathcal{C}$ are subjected to some constraints which define the feasible sets $\mathcal{B}$ and $\mathcal{C}$. We use variable $y_n \in \{-1(\text{abnormal}), 1(\text{healthy})\}$ to denote labels of the subjects. Healthy subjects are denoted by 1 and abnormal ones by -1.

An optimal basis vector ($\mathbf{b}_k$) operates as a region selector; therefore its entries (b_{jk}) must be either *on* or *off* (*i.e.*, $b_{jk} \in \{0, 1\}$). Since optimizing integer values is computationally expensive, particularly for the large dimensionality characteristic of medical images, we relax this constraint to $0 \leq b_{jk} \leq 1$ which can be encoded mathematically by a combination of ℓ_∞ norm and non-negativity ($\mathbf{b} \geq 0$). Assuming that only certain structures of an anatomy are affected (*e.g.*, atrophy of hippocampus in Alzheimer's disease), we can impose sparsity on the basis vectors which also makes them more interpretable. The sparsity constraint can be enforced by an inequality constraint over the ℓ_1 norm of the basis vectors. These two properties constitute the feasible set for the basis vectors ($\mathcal{B}$) as follows (see Chapter 4 for more details):

$$\mathcal{B} := \{\mathbf{b} \in \mathbb{R}^D : \mathbf{b} \geq \mathbf{0}, \|\mathbf{b}\|_\infty \leq 1, \|\mathbf{b}\|_1 \leq \lambda_3\}, \tag{5.2.1}$$

where the ratio of λ_3/D encodes the ratio of sparsity of the basis vectors.

For the feasible set of coefficients ($\mathcal{C}$), we only assume non-negativity (*i.e.*, $\mathcal{C} := \{\mathbf{c} : \mathbf{c} \geq \mathbf{0}\}$) because our images are usually non-negative however this is not a limitation for the model, and this constraint can be relaxed in the case of negative values in image (see Section 5.2.2).

In order to find optimal $\mathbf{B}$ and $\mathbf{C}$ matrices, we define the following constrained optimization problem:

$$\begin{aligned}
\min_{\mathbf{B},\mathbf{C},\mathbf{w}\in\mathbb{R}^K} \quad & \lambda_1 \mathcal{D}(\mathbf{X}; \mathbf{BC}) + \lambda_2 \sum_{n=1}^{N} \ell(y_n; f(\mathbf{x}_n; \mathbf{B}, \mathbf{w})) + \|\mathbf{w}\|_2 \\
\text{subject to: } & f(\mathbf{x}_n; \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T \mathbf{x}_n, \mathbf{w} \rangle \\
& \mathbf{b}_k \in \mathcal{B}, \quad \mathbf{C} \geq 0
\end{aligned} \tag{5.2.2}$$

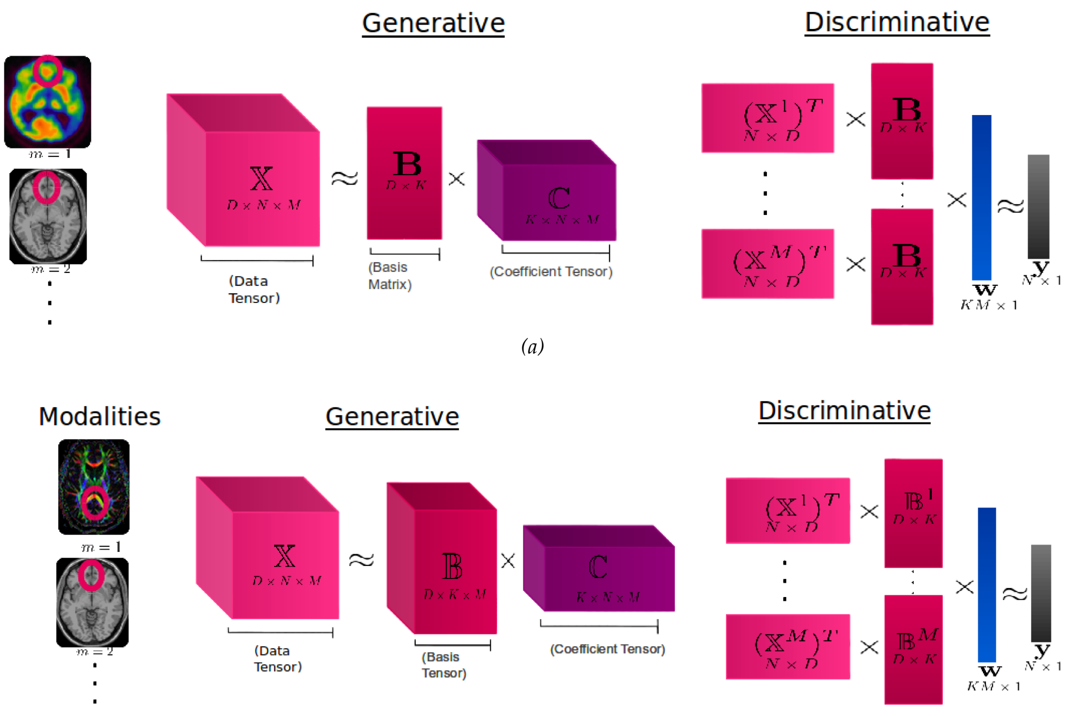
The cost function of the optimization problem consists of two terms: 1) the *generative* term ($\mathcal{D}(\cdot;\cdot)$) encourages the decomposition ($\mathbf{BC}$) to be close to the data matrix ($\mathbf{X}$); 2) the *discriminative* term ($\ell(y_n; f(\mathbf{x}_n, \mathbf{B}, \mathbf{w}))$) is a *loss* function that encourages a classifier $f(\cdot)$ to produce class labels that are consistent with available labels ($\mathbf{y}$). The classifier parametrized by $\mathbf{w}$ produces a label given the new feature ($\mathbf{v}_n = \mathbf{B}^T\mathbf{x}_n$) which is the projection an image ($\mathbf{x}_n$) on the basis vectors. We use a linear classifier, hence $f(\mathbf{x}_n, \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T\mathbf{x}_n, \mathbf{w}\rangle$. Similar to Chapter 3, we set $\mathcal{D}(\mathbf{X}; \mathbf{BC}) = \|\mathbf{X} - \mathbf{BC}\|_F^2$ and λ_1 is a constant. For the loss function, we choose a hinge squared loss function: $\ell(y, \tilde{y}) = (\max\{0, 1 - y\tilde{y}\})^2$, that is a common choice in machine learning (see Chapter 3).

There are three blocks in the optimization problem in *Eq.*(5.2.2): $\mathbf{w}, \mathbf{B}$, and $\mathbf{C}$. The problem is not jointly convex with respect to all blocks however it is block-wise convex. In other words, if any two pairs of blocks are fixed, the problem is convex with respect to the remaining block. The optimization scheme starts from a random initialization of blocks, fixes two blocks, optimizes with respect to the remaining one, and repeats this process for each block. The whole process is repeated till convergence. Optimization with respect to $\mathbf{C}$ and $\mathbf{w}$ is not challenging but, due to the large-scale dimensionality of a medical image, optimization with respect to $\mathbf{B}$ requires a specialized method (see Chapter 4 for details).

5.2.2 Extension to Multi-Modality: Classification Problem

Unlike the uni-channel case, in which each voxel stores a scalar value, in the multi-channel case, each voxel of an image is associated with an array of values. In Section 5.2.1, we stored the training data into a matrix ($\mathbf{X}$); while in the multi-channel case, we need to structure the data into a tensor ($\mathbb{X}$). In fact, in the general framework (Section 5.2.1), the matrix $\mathbf{X}$ can be viewed as an order-2 tensor[2] in which the first index (rows) enumerates voxels and the second index (columns) enumerates subjects. We simply extend this matrix to an order-3 tensor in which the third index (faces) enumerates channels. One can simply concatenate all image channels of a subject into long columns of a matrix and then apply the method presented in Chapters 3 and 4, or a similar method. However, the advantage of viewing the matrix data to a tensor data is that various factorizations can be proposed, each of which implies different hypotheses about

[2] The order of a tensor is the number of indices necessary to refer unambiguously to an individual component of a tensor.


Generative
Discriminative
$m=1$
$m=2$
$\mathbb{X}$ $D \times N \times M$
$\approx$
$\mathbf{B}$ $D \times K$
$\times$
$\mathbb{C}$ $K \times N \times M$
(Data Tensor)
(Basis Matrix)
(Coefficient Tensor)
$(\mathbb{X}^1)^T$ $N \times D$
$(\mathbb{X}^M)^T$ $N \times D$
$\mathbf{B}$ $D \times K$
$\mathbf{w}$ $KM \times 1$
$\mathbf{y}$ $N \times 1$
Modalities
Generative
Discriminative
$\mathbb{X}$ $D \times N \times M$
$\mathbb{B}$ $D \times K \times M$
$\mathbb{C}$ $K \times N \times M$
(Data Tensor)
(Basis Tensor)
(Coefficient Tensor)
$\mathbb{B}^1$ $D \times K$
$\mathbb{B}^M$ $D \times K$

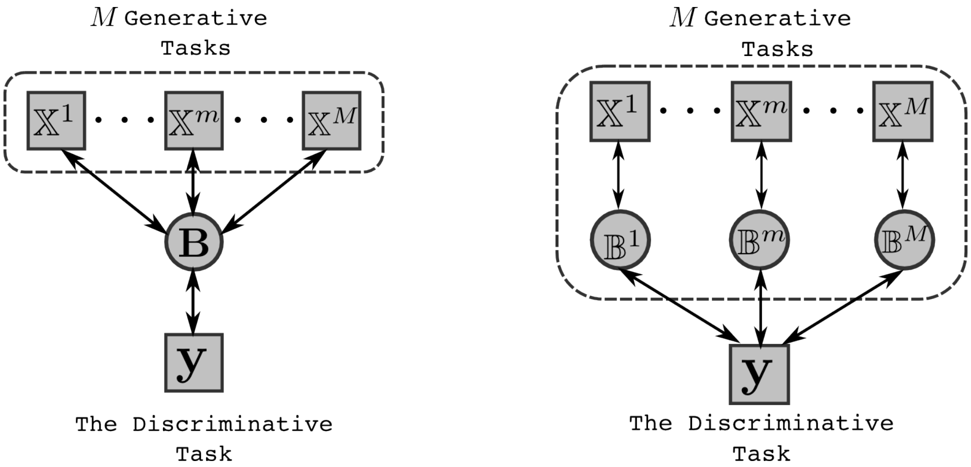
(a)

(b)

Figure 5.1: The difference between the two proposed factorizations: (a) `multi-View`$(\mathbb{X}, \mathbf{y})$, (b) `multi-View`$(\mathbf{y})$. There are M channels stored in the data tensor $(\mathbb{X})$; in (b) for `multi-View`$(\mathbf{y})$, we need to have M sets of basis vectors $(\mathbb{B}^{(1)}, \cdots, \mathbb{B}^{(M)})$ and corresponding coefficients $(\mathbb{C}^{(1)}, ..., \mathbb{C}^{(M)})$, while for `multi-View`$(\mathbb{X}, \mathbf{y})$ (in (a)), there is one set of basis vectors $(\mathbf{B})$ shared across channels. The method in (a) is more proper for channels focusing on the same tissue type while (b) can be applied for channels focusing on different tissue types.

the data because of the structure of a tensor. In this section, we introduce two factorizations and explain their connotations (pictorially represented in Figure 5.1).

Our method can be viewed as *multi-view* learning [126]. In the multi-view setting, the goal is to implicitly learn about the target via the relationship between different views [126]. The goal is to learn the target (here, class labels) via the relationship between different views (here, different channels) [126]. In this chapter, we introduce two factorizations and explain their connotations. One of the variants is more appropriate for a setting that all channels focus on the same tissue type; for example PET and T1 which both focus on gray matter tissue. The other variant is more applicable for channels characterizing different tissue types; for example DTI and T1 which characterize white and gray matter tissues respectively.



(a)

(b)

Figure 5.2: This figure shows how we can viewed the proposed methods as *Multi-Task* learning. (a) shows a schematic representation of multi-View($\mathbb{X}, \mathbf{y}$) (*i.e., Eq.*5.2.3). There are M generative tasks and one discriminative task; $\mathbf{B}$ is shared across them. (b) shows schematic representation of multi-View($\mathbf{y}$) (*i.e., Eq.*5.2.4). There are M generative tasks each of which has it own $\mathbb{B}^m$ but they share their parameters with the discriminative task.

multi-View($\mathbb{X}, \mathbf{y}$):

One assumption could be that there is one hidden variable (here basis vectors: $\mathbf{B}$) that is shared across image channels and class labels. This mostly makes sense for the cases that the multiple channels measure various quantities of *the same tissue*; for example Fractional Anisotropy (FA) and Trace both characterize white the matter tissue of a brain. Different channels indicate different signatures of an abnormality at the same region of the anatomy (*e.g.,* brain); therefore they share the location ($\mathbf{B}$) but with different coefficients ($\mathbb{C}^m$). In this case, both class labels ($\mathbf{y}$) and data ($\mathbb{X}$) are the targets; we will refer to the method as multi-View($\mathbb{X}, \mathbf{y}$) (see Figure 5.1a). It can also be viewed as a Multi-task learning process [163]. Here, we have $M + 1$ tasks: M generative tasks to reconstruct the data and 1 task to reconstruct the class label (*i.e.,* classification). In multi-View($\mathbb{X}, \mathbf{y}$), all $M + 1$ tasks share the same parameters, namely $\mathbf{B}$ (see Figure 5.2a).

We can modify *Eq.*5.2.2 as follows:

$$\min_{\mathbf{B}, \mathbb{C}, \mathbf{w} \in \mathbb{R}^K} \quad \lambda_1 \sum_{m=1}^{M} \|\mathbb{X}^m - \mathbf{B}\mathbb{C}^m\|_F^2 + \lambda_2 \sum_{n=1}^{N} \ell(y_n; f(\mathbf{x}_n; \mathbf{B}, \mathbf{W})) + \|\mathbf{W}\|_F^2$$

$$
\begin{aligned}
\text{subject to: } & f(\mathbb{X}_n; \mathbf{W}, \mathbf{B}) = \sum_{m=1}^{M} \langle \mathbf{w}_m, \mathbf{B}^T \mathbb{X}_n^m \rangle \\
& \mathbf{b}_k \in \mathcal{B}, \quad \mathbb{C} \geq 0
\end{aligned}
\tag{5.2.3}
$$

where the generative term $\mathcal{D}(\cdot;\cdot)$ is augmented to reconstruct different channels however $\mathbf{B}$ is shared across the channels and they differ by their coefficients ($\mathbb{C}^m$). The classifier ($f(\cdot)$) is also augmented. It is parametrized by a matrix ($\mathbf{W} \in \mathbb{R}^{K \times M}$) instead of $\mathbf{w} \in \mathbb{R}^K$ but this extension can also be viewed as extension of of $\mathbf{w}$ to longer vector as shown in Figure 5.1a. The regularizer of $\mathbf{w}$ in *Eq.*5.2.2 is simply augmented to the Frobenius norm, namely $\|\mathbf{W}\|_F^2 = \sum_{m=1}^{M} \|\mathbf{w}_m\|_2^2$.

`multi-View`($\mathbf{y}$):

Unlike `multi-View`($\mathbb{X}, \mathbf{y}$), an alternative assumption could be that there is no hidden variable shared across channels, hence every channel has its own basis vectors ($\mathbb{B}^{(m)}$), but projection on these basis vectors collaborate to predict class labels. For example, different channels may measure quantities on *non-overlapping regions* of a brain (*e.g.,* white matter and gray matter) each quantifying complementary features about the class labels. We refer to this variation as `multi-View`($\mathbf{y}$). Since $\mathbb{B}^{(m)}$'s need to collaborate on the discriminative term, this assumption is still different than applying the uni-channel method separately. `multi-View`($\mathbf{y}$) can also be viewed as Multi-task learning. Similar to `multi-View`($\mathbb{X}$,$\mathbf{y}$), there are $M+1$ tasks, M generative tasks that are independent from each other because they have their own parameters ($\mathbb{B}^{(m)}$) and 1 discriminative task which shares the parameter with each of the M generative tasks (see Figure 5.2b).

We can modify *Eq.*5.2.2 as follows:

$$
\begin{aligned}
\min_{\mathbb{B},\mathbb{C},\mathbf{w}\in\mathbb{R}^K} \quad & \lambda_1 \sum_{m=1}^{M} \sum_{m=1}^{M} \|\mathbb{X}^m - \mathbb{B}^m \mathbb{C}^m\|_F^2 + \lambda_2 \sum_{n=1}^{N} \ell(y_n; f(\mathbf{x}_n; \mathbf{B}, \mathbf{W})) + \|\mathbf{W}\|_F^2 \\
\text{subject to: } & f(\mathbb{X}_n; \mathbf{W}, \mathbb{B}) = \sum_{m=1}^{M} \langle \mathbf{w}_m, (\mathbb{B}^m)^T \mathbb{X}_n^m \rangle
\end{aligned}
$$

$$\mathbf{B}(:,n,m) \in \mathcal{B}, \quad \mathbb{C} \geq 0 \tag{5.2.4}$$

where the generative term $\mathcal{D}(\cdot;\cdot)$ is augmented to reconstruct different channels which is basically sum of M independent reconstruction with their own $\mathbb{B}^m$ and $\mathbb{C}^m$. The classifier $(f(\cdot))$ is also augmented with respect to *Eq.*5.2.2. It is parametrized by a matrix ($\mathbf{W} \in \mathbb{R}^{K \times M}$) instead of $\mathbf{w} \in \mathbb{R}^K$ but this extension can also be viewed as extension of $\mathbf{w}$ to longer vector as shown in Figure 5.1b. Notice that the generative term is separable for each channel but basis matrices ($\mathbb{B}^m$'s) are coupled together through the loss function ($\ell(\cdot,\cdot)$) in Eq.(5.2.4); therefore, it is different than applying the uni-channel algorithm (Section 5.2.1) separately and concatenating extracted features later for a classifier.

5.2.3 Resting-state fMRI: Network Detection

Over recent years, there has been a growing interest in studying brain connectivity using resting-state fMRI (rs-fMRI) [30]. By discovering which regions are functionally connected, we can learn more about the functional organization of the brain [4] and potentially identify bio-markers for diseases such Alzheimer's [211]. However, unlike task-based fMRI, there is no external variable to fit a model against which renders discovering brain networks challenging.

A common approach is to calculate temporal correlations between the mean signals of pre-defined regions of interest (ROI's) [203]; if two regions are highly correlated, they are considered connected. Although such a model-based method may produce interpretable results, the outcomes are highly dependent on the ROIs chosen [203]. At the other end of the spectrum are data-driven approaches that do not require pre-defined seeds, such as Independent Component Analysis (ICA) [39]. To improve spatial localization of ICA and in turn the clinical interpretability of the results, many researchers have suggested using a sparsity prior for the spatial term; for example, Varoquaux *et al.* [206] suggests smooth-Lasso penalty as a regularizer. Another data-driven method is clustering [52], [91]. For example, Golland *et al.* [91] suggests optimally partitioning the volume into a set of disjoint networks. However, a shortcoming of this approach is that the clusters typically do not overlap whereas one region of a brain may be involved in multiple networks. Furthermore, the clusters are not necessarily biologically plausible. By reducing degrees of freedom, a structural prior may alleviate the lack of controlled experimental design in rs-fMRI and potentially improve clinical interpret-ability.

The relationship between functional and structural connectivity is not fully understood, but combining the two may improve our understanding of brain networks. A few methods have been recently proposed to embed functional and structural connectivity into a common framework [208], [63], [236]. In this section, we propose a method that bridges between user-driven and data-driven approaches. We reinterpret clustering as matrix factorization that decomposes data into two sets of latent variables: spatial maps of brain activity and corresponding time courses. Subjects share the activation maps but every subject has its own time signature. We model functional activity maps as sparse combinations of structurally connected parcels that we refer here as *groups*. The *groups* can simply be set of voxels in an ROI or connected set of voxels through fiber tracks (see Figure 5.4). We suggest imposing sparsity on the *union of the groups* rather than at the voxel level; *i.e.,* we would like encourage few groups to co-activate instead of voxels.

We can view rs-fMRI as an instance of multi-channel image. Each time sample, which is an image, can be viewed as a channel. The proposed method in Section 5.2.2 can be applied to identify networks. In this case, identifying a functional network can be viewed as a generative problem (unless there are two or more cohorts of subjects for whom a class labels exist), and we suggest using *Eq.*5.2.3 ($\lambda_2 = 0$) because we would like to find a common area across channels (or time points). It is shown in Chapter 3 that the method can be viewed as a clustering approach except that it allows clusters to overlap. Figure 5.3 shows the concepts pictorially. Assume our dataset contains T time points for each of N subjects, and that each time point is an image containing D voxels. The data is stored in a tensor $\mathbb{X} \in \mathbb{R}^{D \times N \times T}$. Using *Eq.*5.2.3, the columns of $\mathbf{B}$ are representative of the clusters that can be viewed as regions in the brain. Assuming that

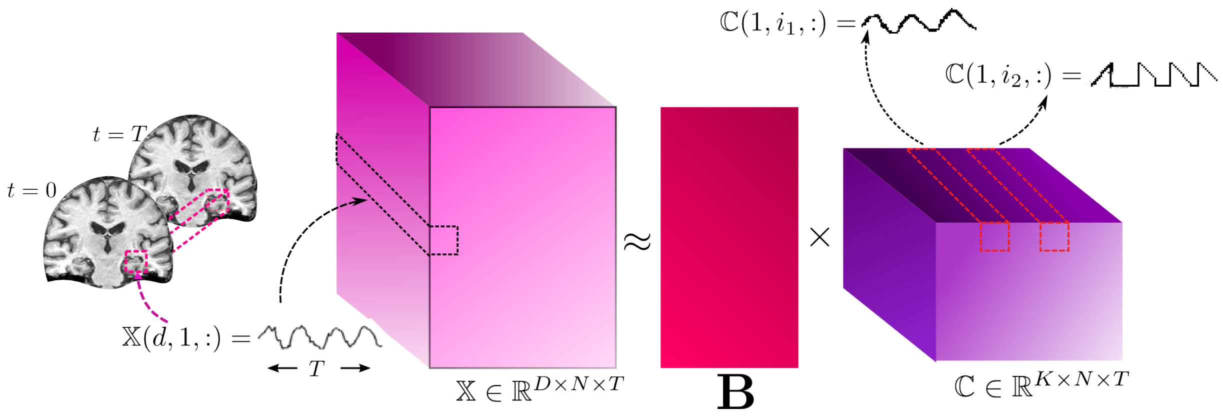

$\mathbb{C}(1, i_1, :) =$
$\mathbb{C}(1, i_2, :) =$
$t = T$
$t = 0$
$\mathbb{X}(d, 1, :) =$
T
$\mathbb{X} \in \mathbb{R}^{D \times N \times T}$
$\approx$
$\times$
$\mathbf{B}$
$\mathbb{C} \in \mathbb{R}^{K \times N \times T}$

Figure 5.3: The figure shows T time series of all N subjects collected in $\mathbb{X}$. $\mathbb{X}(d, 1, :)$ denotes time signal of the d'th voxel of the first subject. $\mathbb{C}(1, i_1, :)$ and $\mathbb{C}(1, i_2, :)$ are time-series *centroids* of the first cluster (basis) corresponding to i_1'th and i_2'th subjects respectively. Notice that the algorithm clusters time-series; subjects share common basis (spatial pattern) but their time-series may differ.

all subjects are aligned, $\mathbf{B}$ finds common areas. Subjects shares the activation maps but every subject has its own time signature (see Figure 5.3).

Setting $\lambda_2 = 0$, for rs-fMRI, the following optimization problem needs to be solved:

$$\begin{aligned} &\min_{\mathbf{B}, \mathbb{C}} \quad \sum_{t=1}^{T} \|\mathbb{X}^t - \mathbf{B}\mathbb{C}^t\|_F^2 \\ &\text{subject to: } \|\mathbf{b}_k\|_1 \leq \lambda_3, \quad \|\mathbf{b}_k\|_\infty \leq 1, \quad \mathbf{b}_k \geq 0 \end{aligned} \tag{5.2.5}$$

Here, T is number of time-points and $\mathbb{X}^t$ is a matrix (t'th face of $\mathbb{X}$) holding images of all N subjects in the t'th time point. Assuming that each image has D voxels, $\mathbb{X}^t \in \mathbb{R}^{D \times N}$. Notice that the *non-negativity* term is dropped on $\mathbb{C}$ because after de-trending and other pre-processing steps on rs-fMRI signal, the time series signal is not non-negative. However, the non-negativity on $\mathbf{B}$ is kept because it contributes in clustering properties of the formulation.

Structural Connectivity Prior

The relationship between functional and structural connectivity is not fully understood, but combining the two may improve our understanding of brain networks. A few methods have been recently proposed to embed functional and structural connectivity into a common frame-

$$\mathcal{G} = \{g_1, g_2, (g_1 \cup g_2), \cdots\}$$

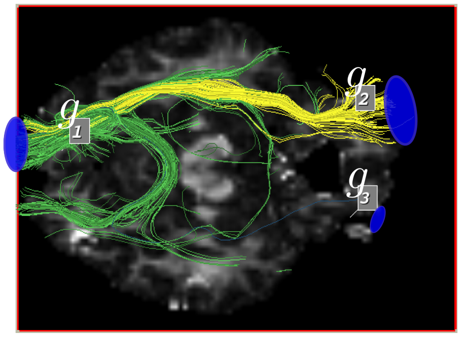
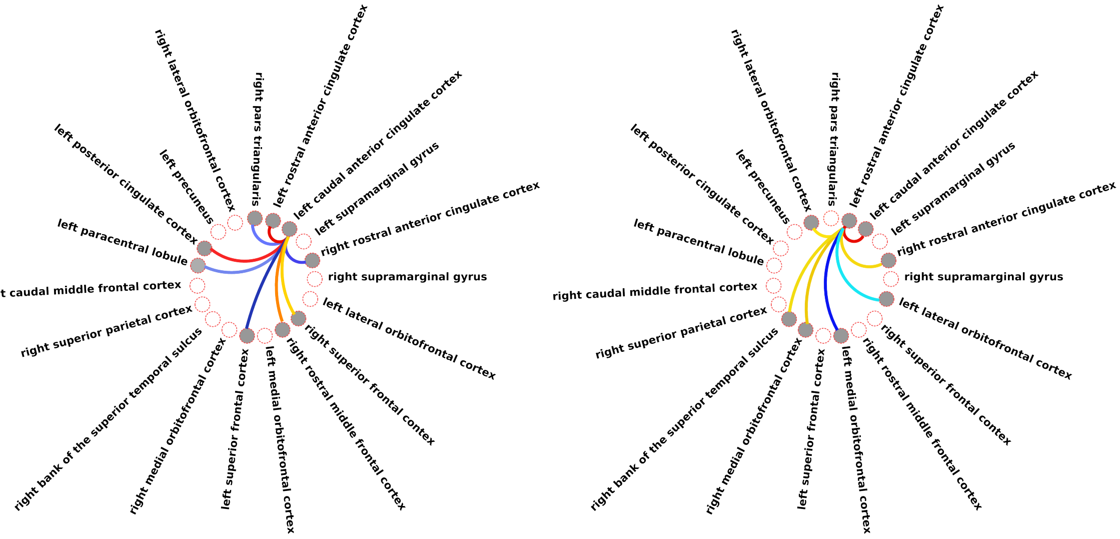


Figure 5.4: An example on how groups can be constructed in real data: each of the 3 regions g_1,g_2,g_3 can be member of sets of groups ($\mathcal{G}$). There are a lot of fibers connecting g_1 and g_2; hence $(g_1 \cup g_2)$ can form another group.

work [208], [63], [236].

The formulation in *Eq*.5.2.5 does not account for prior knowledge about the underlying data (*i.e.*, 3D image). To illustrate the value of domain knowledge, notice: 1) the columns of $\mathbb{X}^t$ concatenate all voxels into a long vector and ignore that voxels are structured in a specific order within an image, 2) it is not obvious how to incorporate other types of prior knowledge about brain structures such as connectivity into the formulation. For example, we might know a priori that two regions are connected through white-matter fiber tracks (see Figure 5.4) or there is correlation between the their gray-matter thickness [123]. A possible scenario is shown in Figure 5.5; warmer colors indicate stronger connections between areas. The strength of the connections can be measured via different methods such as white matter tractography (see Figure 5.4) or any other method [123]. It is not immediately obvious how to incorporate such information into the formulation.

We propose to form groups based on structural connectivity. The idea is that members of a group are more *similar* (*e.g.*, connected) to each other. Instead of imposing sparsity on voxel-level, we suggest to impose sparsity on the *union of groups*; *i.e.*, instead of few voxels, we would like to encourage few groups to co-activate. Before introducing the notion, recall from Section 4.3, $\mathcal{G} =$ contains set of groups. Each group $g \in \mathcal{G}$ is set of indices of voxels belonging to that group. In


right lateral orbitofrontal cortex
right pars triangularis
left rostral anterior cingulate cortex
left caudal anterior cingulate cortex
left supramarginal gyrus
right rostral anterior cingulate cortex
right supramarginal gyrus
left lateral orbitofrontal cortex
right superior frontal cortex
right rostral middle frontal cortex
left medial orbitofrontal cortex
left superior frontal cortex
right medial orbitofrontal cortex
right bank of the superior temporal sulcus
right superior parietal cortex
right caudal middle frontal cortex
left paracentral lobule
left posterior cingulate cortex
left precuneus

(a) (b)

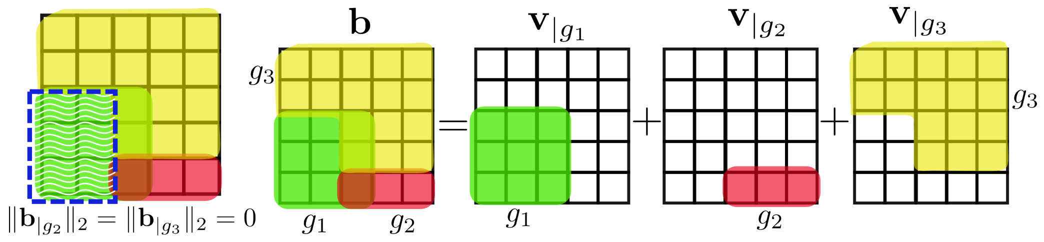
Figure 5.5: (a) and (b) shows connectivity examples in human brain: warmer colors mean stronger connections [123]. It shows that "left caudal anterior cingulate cortex" is strongly connected to "left rostral anterior cingulate cortex" and "left posterior cingulate cortex" can be considered as a group g_1. On the other hand, "left rostral anterior cingulate cortex" is strongly connected to "left caudal anterior cingulate cortex" which can be considered as a group g_2.

Chapter 4, defined two variants of the group-norms as follow:

$$\begin{aligned} \|\mathbf{b}\|_{1,2} &:= \sum_{g \in \mathcal{G}} \rho_g \|\mathbf{b}_{|g}\|_2 \\ \|\mathbf{b}\|_{\infty,2} &:= \max_{g \in \mathcal{G}} \{\rho_g \|\mathbf{b}_{|g}\|_2\} \end{aligned} \tag{5.2.6}$$

where $\mathbf{b}_{|g}$ is a D-dimensional vector such that its voxels not belonging to the group g are set to zero and ρ_g is a positive constant. There are two major differences between our objectives here and Section 4.3:

- We emphasized in Section 4.3 that the groups should not overlap. However, in order to detect connectivity, groups may or in some cases should overlap. For example, in Figure 5.5, it shows that "left caudal anterior cingulate cortex" is strongly connected to "left rostral anterior cingulate cortex" and "left posterior cingulate cortex" which can form group g_1. On the other hand, "left rostral anterior cingulate cortex" is strongly connected to "left



(a) *(b)*

Figure 5.6: In (a), if $\|\mathbf{b}_{|g_2}\|_2 = \|\mathbf{b}_{|g_3}\|_2 = 0$, set of possible non-zero voxels belong to intersection of complement of $g_2 \cup g_3$ (inside of dashed blue line) which might not be meaningful. However the model in (b) does not have the same problem: both g_1 and g_3 can shrunk to zero but since $\mathbf{b}$ is combination of the groups and support region of $\mathbf{b}$ is still a valid group.

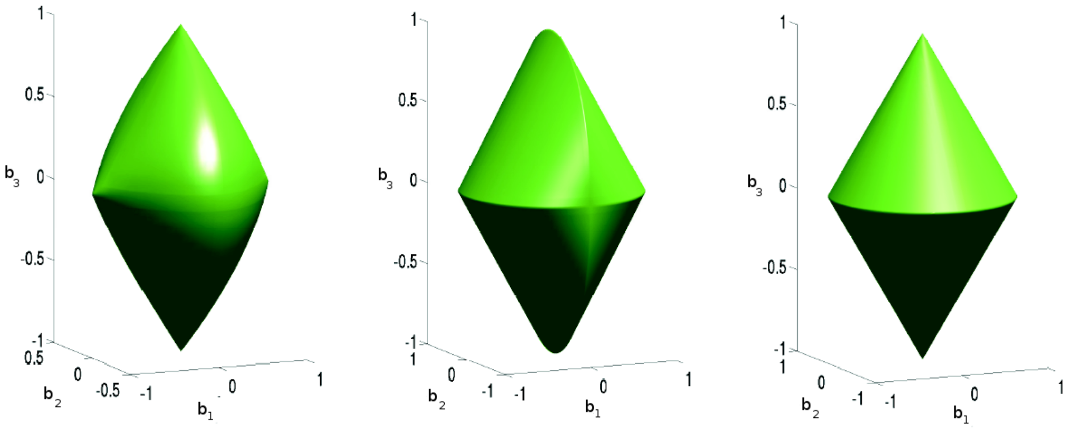
caudal anterior cingulate cortex" which can form g_2.

- Even if groups overlap, imposing sparsity on the group level does not mean that the result would be the union of the groups but it might be their intersection which may not meaningful. The idea is presented in Figure 5.6a: if in *Eq.*5.2.6 with overlapping groups, $\|\mathbf{b}_{|g_2}\|_2 = \|\mathbf{b}_{|g_3}\|_2 = 0$, then the pixels which are allowed to be non-zero belong to $(g_2 \cup g_3)^c$. If g_1 happens to be important (hence $\|\mathbf{b}_{g_1}\|_2 \neq 0$), only pixels of g_1 which are also in $(g_2 \cup g_3)^c$ can be non-zero; however $g_1 \bigcap (g_2 \cup g_3)^c$ may no be meaningful.

Obozinski and Jacob [118] suggested a regularizer to select entire variables in a union of selected groups. The idea is to introduce a new set of variables, $\mathbf{v}_{|g} \in \mathbb{R}^D$, that is non-zero only inside of the group (*i.e., supp*$(\mathbf{v}_{|g}) \subset g$) and add an extra equality constraint, namely $\mathbf{b} = \sum_{g \in \mathcal{G}} \mathbf{v}_{|g}$. Inspired by this idea, we can change *Eq.*5.2.6 to define a new regularization:

$$\Omega^{\mathcal{G}}(\mathbf{b}) := \min_{\forall g \in \mathcal{G}, \mathbf{v}_g} \sum_{g \in \mathcal{G}} \rho_g \|\mathbf{v}_{|g}\|_2, \quad \text{s.t.} \sum_{g \in \mathcal{G}} \mathbf{v}_{|g} = \mathbf{b}, \quad \mathbf{v}_{|g} \geq 0, \|\mathbf{v}_{|g}\|_2 \leq 1 \tag{5.2.7}$$

The equality constraint decomposes $\mathbf{b}$ as sum of $\mathbf{v}_{|g}$'s whose support are included in each group as in Figure 5.6b. Since the equality constraint is enforced, an i'th entry of $\mathbf{b}$ can be non-zero as long as it belongs to at least one non-shrunk group. Figure 5.7 illustrates the unit balls of the group-norms for three different definitions of the group-norm introduced in this thesis



(a) *(b)* *(c)*

Figure 5.7: (b) and (c) are unit balls of group norms with and without union property respectively. Singularities exist in both cases, but occur at different positions: for (c) they correspond to situations where only b_1 or only b_2 is nonzero, *i.e.,* where all covariates of one group are shrunk to 0. (b) corresponds to new regularization in *Eq.*5.2.7, singularities correspond to situations where only b_1 or only b_3 is equal to 0, *i.e.,* where all covariates of one group are nonzero. For comparison, the unit ball of non-overlapping group is also shown for $\mathcal{G} = \{\{1\}, \{2, 3\}\}$. Figures are adopted from [162].

so far. Figure 5.7c shows the unit ball of the group norm defined in Chapter 4 (*i.e., Eq.*5.2.6) for $\mathcal{G} = \{\{1\}, \{2, 3\}\}$. Figure 5.7a and Figure 5.7b show the unit balls for $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$. They show the differences between the new (*i.e., Eq.*5.2.7) and the former definition of the group norm (*i.e., Eq.*5.2.6) when groups overlap. While Figure 5.7a has only four singularities corresponding to either $b_1 = 0$ or $b_3 = 0$, Figure 5.7b has two circular sets of singularities corresponding to $(b_2, b_3) = 0$ and $(b_1, b_2) = 0$ [162]. The unit ball of non-overlapping group norm is shown just for comparison.

Replacing the regularization term in *Eq.*5.2.5 with the new regularizer, we have:

$$\begin{aligned} &\min_{\mathbf{B},\mathbb{C}\in\mathbb{R}^K} \quad \sum_{t=1}^{T} \|\mathbb{X}^t - \mathbf{B}\mathbb{C}^t\|_F^2 \\ &\text{subject to:} \quad \Omega^{\mathcal{G}}(\mathbf{b}_k) \leq \lambda_3, \quad 1 \leq k \leq K \end{aligned} \tag{5.2.8}$$

the problem is not jointly convex but it is convex fixing each block and optimizing with respect to the other.

We use a block-wise optimization scheme to solve *Eq.*5.2.8. Solving for $\mathbb{C}$ is straightforward but due to the large dimensionality of the images, solving with respect to $\mathbf{B}$ is challenging. As discussed in Chapter 4, the bottle-neck is on solving the projection or proximity operator efficiently, namely

$$\mathcal{P}(\mathbf{u}) = \arg \min_{\Omega^{\mathcal{G}}(\mathbf{b}) \leq \lambda_3} \|\mathbf{u} - \mathbf{b}\|_2, \tag{5.2.9}$$

In Algorithm 2 in Chapter 4, we suggested to use SPG to solve for $\mathbf{B}$. Algorithm 3 or Algorithm 4 were suggested for the proximity operator depending on the type of regularizer. For the new group norm introduced here, we suggest to use SPG to solve *Eq.*5.2.9. Therefore, the optimizer for $\mathbf{B}$ constitutes of two nested SPG; one for the proximity operator and the other one for updating $\mathbf{B}$. The idea is to rewrite *Eq.*5.2.9 as another equivalent optimization problem with non-overlapping groups and use the algorithms proposed in Chapter 4 as a sub-module. To do so, we need to rewrite *Eq.*5.2.9 as follows:

$$\begin{aligned} &\min_{\mathbf{z}} \|\mathbf{A}\mathbf{z} - \mathbf{u}\|_2 \\ &\text{s.t. } \|\mathbf{z}\|_{1,2} \leq \lambda_3, 0 \leq \mathbf{z} \leq 1 \end{aligned} \tag{5.2.10}$$

where $\mathbf{z}$ is constituted by concatenating elements of each group therefore it has $\sum_{g \in \mathcal{G}} |g|$ elements; $\| \cdot \|_{1,2}$ is a non-overlapping norm defined similar to *Eq.*4.3.1. $\mathbf{A} \in \mathbb{R}^{D \times (\sum_{g \in \mathcal{G}} |g|)}$ is a membership matrix; its entries are either 0 or 1 depending on i'th row being a member of corresponding group in $\mathbf{z}$ or not. This reformulation casts the problem to a non-overlapping group on $\mathbf{z}$ that let us use the efficient Algorithm 4 proposed in Chapter 4. The algorithm is summarized in Algorithm 5.

Algorithm 5 SPG solver for the proximity operator (*Eq*.5.2.10)

```
Require: Initial point, step-length bounds $0 < \alpha_{\min} < \alpha_{\max}$, $\nu$, $M$
  repeat
    $\mathbf{d} \leftarrow \mathcal{P}_{\mathcal{B}}(\mathbf{z}^k - \alpha_k(\mathbf{A}^T\mathbf{A}\mathbf{z}^k - 2\mathbf{A}^T\mathbf{u})) - \mathbf{z}^k$     (using Alg.4)
    $\gamma \leftarrow 1$
    $M \leftarrow \max_{k-M \leq i \leq k}\{\|\mathbf{A}\mathbf{z}^i - \mathbf{u}\|_2\}$
    while $\|\mathbf{A}(\mathbf{z}^k + \gamma\mathbf{d}) - \mathbf{u}\|_2 > M + \nu\gamma\langle(\mathbf{A}^T\mathbf{A}\mathbf{z}^k - 2\mathbf{A}^T\mathbf{u}), \mathbf{d}\rangle$ do
      Choose $\gamma \in (0, 1)$ with quadratic interpolation [96]
    end while
    $\mathbf{z}^k \leftarrow \mathbf{z}^k + \gamma\mathbf{d}$
    compute step-length: $\alpha_k \leftarrow \min\{\alpha_{\max}, \max\{\alpha_{\min}, \alpha_{bb}\}\}$ ($\alpha_{bb}$ in Eq.(3.5.9))
    $k \leftarrow k + 1$
  until some convergence criteria satisfied
```

5.3 Experiments

In this section, we show the results from two sets of experiments. In Section 5.3.1, we apply our method on real multi-channel data for classification purposes. In Section 5.3.2, we simulate synthetic images and examine the applicability of the proposed method in detecting a network in various settings as well as exploring its utility in the analysis of fMRI data.

5.3.1 Classification with Real Data

For this section, we acquired a subset of images from a longitudinal brain imaging study for validation of our method. The objective of this choice was to investigate the longitudinal progression of changes in brain structure (MRI) and brain function ($[^{15}O]$-water PET-CBF) in relation to cognitive change in cognitively normal older adults. We used slopes of CVLT[3] score over the follow-up period as a measure of cognitive function to subdivide the entire cohort into two groups: top 20% (25 subjects) showing the highest cognitive stability (CN: cognitively normal), and bottom 20% (25 subjects) showing the most pronounced cognitive decline (CD: cognitively declining).

All T1-MR images used in this study were pre-processed according to [80] and registered to a template. Two volumetric tissue density maps [187] were formed for white matter (WM), gray matter (GM) regions. These maps quantify an expansion (or contraction) to the tissue applied by

[3] California Verbal Learning Test [64]

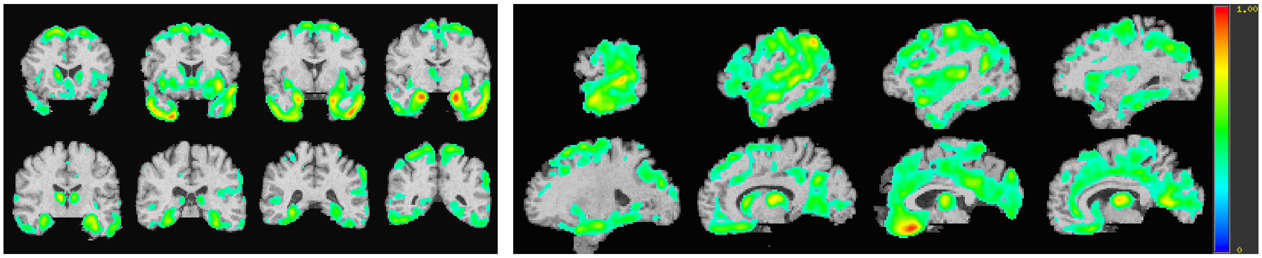


Figure 5.8: Two examples of the basis vectors shown in different cuts. Left: Multi-View$(\mathbb{X}, \mathbf{y})$, Right: Multi-View$(\mathbf{y})$ ($\gamma^* = 100$; number of basis vectors is 60).

the transformation to warp the image to the template space.

Samples are divided into five folds and $4/5$ of samples are used for training basis vectors (an example of which is shown in Figure 5.8); projections on these basis vectors are used as features and are fed to an SVM classifier.

In uni-parametric dataset, the algorithm is relatively stable as long as λ's are chosen within reasonable ranges (see [17]). We set the parameters to the most frequently chosen parameters used for the uni-channel case on a totally different dataset. Numbers reported in Table 5.1 are produced using such parameters. Nevertheless, we performed sensitivity analysis with respect to ratio of λ_1/λ_2 and number of basis vectors, K (see Figure 5.9). For notational brevity, in γ^* for ratio of λ_1/λ_2 we used for Table 5.1. Different curves in Figure 5.9 denote different ratios of λ_1/λ_2. While Multi-View$(\mathbf{y})$ is relatively stable with respect to K and different ratios, performance of Multi-View$(\mathbb{X}, \mathbf{y})$ improves as K increases. Although parameters that are more inclined toward the unsupervised setting (*e.g.*, $\lambda_1/\lambda_2 = 10\gamma^*$) under-perform settings that are excessively discriminative (*e.g.*, $\lambda_1/\lambda_2 = 0.001\gamma^*$), are more stable. This observation can be explained by the fact that a weak regularization was imposed on the discriminative term (*i.e.*, there is almost no $\|\mathbf{W}\|_F^2$) making the algorithm vulnerable to over-fitting.

Table 5.1 reports the average classification rates on the left-out folds for different scenarios and methods. We used a publicly available software, called COMPARE [80], for comparison. The COMPARE method has been applied to many problems and has been claimed to per-

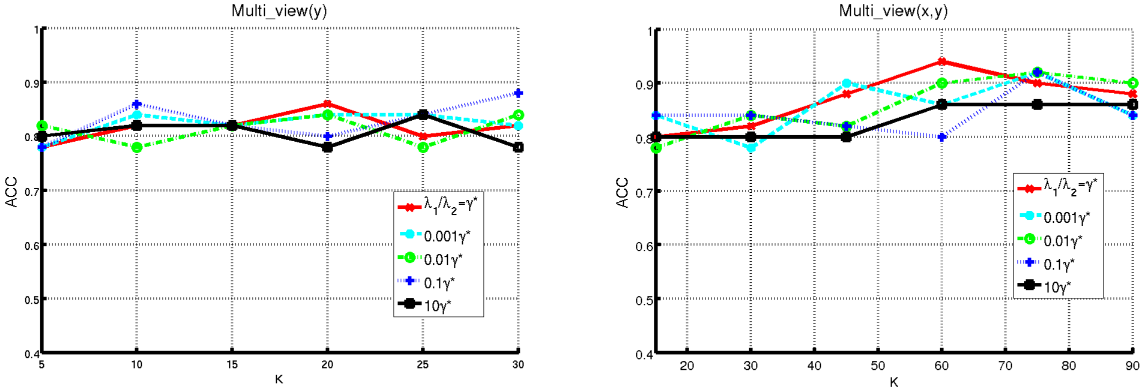


Figure 5.9: Sensitivity Analysis: accuracy rates with respect to different number of basis vectors (K) for various ratios of λ_1/λ_2. Left: `Multi-View`($\mathbf{y}$). Right: `Multi-View`($\mathbb{X}, \mathbf{y}$).

Table 5.1: Comparison of classification accuracy rates for different scenarios and different methods on "cognitively normal" (NC) versus "cognitively declining" (CD) subjects. Results are reported in the format: accuracy (sensitivity,specificity); with $\gamma^* = 100$; total number of basis vectors in each experiment is 60.

	NC vs. CD			
	(WM,PET)	(WM,GM)	(GM,PET)	(GM, WM, PET)
`Multi-View`($\mathbb{X}, \mathbf{y}$)	0.82 (0.84,0.8)	0.76 (0.72,0.8)	**0.84** (0.88,0.8)	**0.94** (0.88,1.0)
`Multi-View`($\mathbf{y}$)	0.86 (0.84,0.88)	0.84 (0.8,0.88)	0.78 (0.8,0.76)	0.84 (0.84,0.84)
`m-COMPARE`	**0.88** (0.8,0.96)	**0.86** (0.88,0.84)	0.8 (0.8,0.8)	0.86 (0.84,0.88)
`COMPARE`	0.78 (0.68,0.88)	0.82 (0.76,0.88)	0.82 (0.84,0.8)	0.82 (0.76,0.88)
`Single-View`	0.84 (0.8,0.88)	0.84 (0.8,0.88)	0.82 (0.84,0.8)	0.8 (0.76,0.84)

form very well. Its variants, *i.e.*, `COMPARE` and `m-COMPARE`, are similar to `Multi-View`($\mathbf{y}$) and `Multi-View`($\mathbb{X}, \mathbf{y}$) respectively. For comparison, we have included `Single-View` results for each scenario in which basis vectors are extracted independently and features are concatenated and fed to the same procedure to find the best parameters for a classifier as the multi-view methods. Since results shown in the table are column-wise comparable, the highest values in the column are magnified with a bold font in each column. In general, `Multi-View`($\mathbb{X}, \mathbf{y}$) or its counterpart `m-COMPARE` perform better. In all columns, at least one of the multi-view methods outperforms the single view equivalent and the best performance is achieved by `Multi-View`($\mathbb{X}, \mathbf{y}$).

5.3.2 Network Recovery: Synthetic Data

In this set of experiments, parametric consistency of the algorithm is studied empirically. In other words, under certain generative assumptions to generate samples (*i.e.,* images), we empirically study if the algorithm can successfully recover correct parameters (*i.e.,* basis vectors). The idea is to synthesize basis vectors resembling default network in human brain. As reported in literature [95], there is a so-called *default network* in a brain that involves approximately similar areas in all individuals. The three basis vectors in Figure 5.10a ($\mathbf{b}_1$ to $\mathbf{b}_3$) mimic this effect. This network operates with it own synchronicity but each individual has her own frequency. Parts of this network may positively ($\mathbf{b}_1$ and $\mathbf{b}_2$) or negatively correlated with each other ($\mathbf{b}_1$ and $\mathbf{b}_3$). In addition to the default network, each individual may have her own activation pattern. This activation may even overlap with default network. In order to mimic this property, we randomly select three basis vectors (from $\mathbf{b}_4, \cdots$) with their own synchronicity.

$$
\begin{array}{lll}
f_i^1 \sim \mathcal{U}[f_{min}, f_{max}], & f_i^2 \sim \mathcal{U}[f_{min}, f_{max}] & \\
\varepsilon_{i1}(t), \varepsilon_{i2}(t), \varepsilon_{i3}(t) \sim \mathcal{N}(0, \sigma), & a_1 \sim \mathcal{N}(\eta, \sigma), a_2 \sim \mathcal{N}(-\eta, \sigma) \;\; (\eta > 0) & \\
c_{i1}(t) = sin(f_i^1 t) + \varepsilon_{i1}(t), & c_{i2}(t) = a_1 sin(f_i^1 t) + \varepsilon_{i2}(t), & c_{i3}(t) = a_2 sin(f_i^1 t) + \varepsilon_{i3}(t), \\
s_4, s_5, s_6 \in \{4, \cdots, 12\} & s_4 \neq s_5 \neq s_6 & \mathbb{P}(4) = \cdots = \mathbb{P}(12) \\
c_{is_4}(t) = sin(f_i^2 t) + \varepsilon_{i4}(t), & c_{is_5}(t) = sin(f_i^2 t) + \varepsilon_{i5}(t), & c_{is_6}(t) = sin(f_i^2 t) + \varepsilon_{i6}(t), \\
\mathbf{x}_i(t) = \sum_{k=1}^{3} \mathbf{b}_k c_{ik}(t) + \sum_{k=4}^{6} \mathbf{b}_k (s_k c_{is_k}(t)) + \varepsilon_i & & \varepsilon_i \sim \mathcal{N}(0, \sigma)
\end{array}
\tag{5.3.1}
$$

Equations in *Eq.*5.3.1 summarize the procedure we used to generate data:

- First, two frequencies f_i^1 and f_i^2 are sampled from a uniform distribution between f_{min} and f_{max}, *i.e.,* $\mathcal{U}[f_{min}, f_{max}]$. f_i^1 is used to synchronize the common basis and f_i^2 is used to synchronize the individual basis.

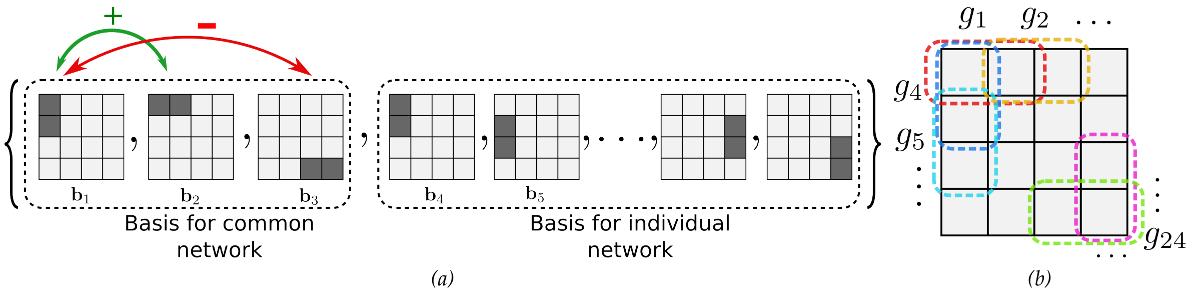


Figure 5.10: (a) shows common basis vectors that are shared across population and basis that can be chosen for each individual. Loading coefficients between $\mathbf{b}_1$ and $\mathbf{b}_2$ are positively correlated and loading coefficients between $\mathbf{b}_2$ and $\mathbf{b}_3$ are negatively correlated. (b) shows the definition of the groups for Group-Sparsity regularization in *Eq.*5.2.7.

- Two correlation values a_{1i} and a_{2i} are sampled from two normal distributions. η is the average positive correlation between $\mathbf{b}_1$ and $\mathbf{b}_2$ and $-\eta$ is the average negative correlation between $\mathbf{b}_1$ and $\mathbf{b_3}$ (see Figure 5.10a).
- From $\mathbf{b}_4$ to $\mathbf{b}_1$2, two basis vectors are chosen without replacement. s_4, s_5, and s_6 indicate indices of the basis vectors. Corresponding coefficients c_{is_4}, c_{is_5}, and c_{is_6} are synchronized with f_i^2.
- All basis and coefficients are mixed together to form the image $\mathbf{x}_i \in \mathbb{R}^{256}$[4] with noise ε_i.

100 images are generated with the procedure in *Eq.*5.3.1. The number of time samples is set to 30 (*i.e.*, $1 \leq t \leq 30$). It renders $\mathbb{X} \in \mathbb{R}^{256 \times 100 \times 30}$.

In order to evaluate the success of the algorithm, we have two criteria: 1) finding correct basis (*i.e.*, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$), 2) getting correct correlation sign between ($\mathbf{b}_1$,$\mathbf{b}_2$) and ($\mathbf{b}_2$,$\mathbf{b}_3$). Those criteria are encoded into three terms: 1) d_1 that is ℓ_2 distance between ground-truth $\mathbf{b}_1$ and our closest estimate to it, say $\hat{\mathbf{b}}_1$: $d_1 = \|\hat{\mathbf{b}}_1 - \mathbf{b}_1\|$; 2) d_2 that is the ℓ_2 distance between basis with positive correlation with $\hat{\mathbf{b}}_1$ that is the most similar to $\mathbf{b}_2$: $d_2 = \|\hat{\mathbf{b}}_2 - \mathbf{b}_2\|$. If there is no basis with positive correlation $d_2 = \|\mathbf{b}_2\|$. 3) Similar definition for $\mathbf{b}_3$ except with negative correlation. Finally $d = d_1 + d_2 + d_3$. We also investigated how an informative group-sparsity prior can affect the consistency. Figure 5.10b shows that groups are moving vertical and horizontal patches of voxels.

[4] Each image is 4×4 blocks and each block is 4×4 pixels, hence the image is 16×16 pixels, hence $\mathbf{x}_i \in \mathbb{R}^{256}$.

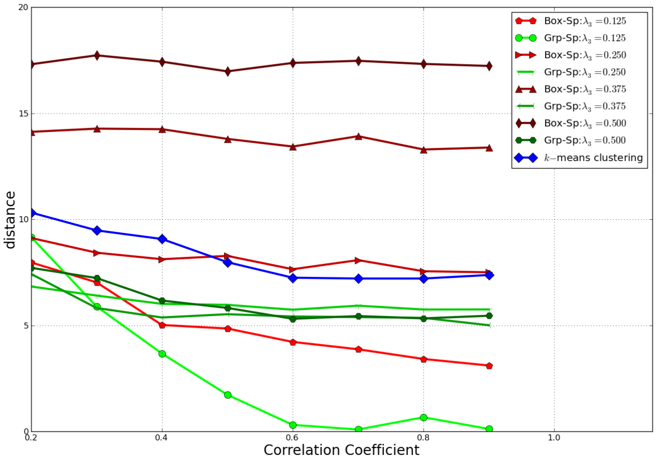


Figure 5.11: The $y-$axis denotes the distance between ground-truth and estimated network in synthetic data. $x-$axis denotes different correlation ratio. The blue curve corresponds to $k-$means clustering algorithm is considered as baseline. Different shades of red are Boxed-Sparsity algorithm for various values of the sparsity constraint (λ_3) and different shades of green are Group-Sparsity for various values of λ_3. Group sparsity robustly outperforms $k-$means.

The experiment was repeated 10 times and averages of d are shown in Figure 5.11 for different ratios of correlation (*i.e.,* η) on the x-axis. Results of the $k-$means are also reported as a base-line. To study the robustness of the algorithm with respect to different values of sparseness constraint, λ_3 is chosen from $\{\frac{D}{8}, \frac{D}{4}, \frac{3D}{8}, \frac{D}{2}\}$; the ground-truth for λ_3 is $\frac{D}{8}$ as shown in Figure 5.10a. Increasing correlation improves the the consistency (decreases d) as expected. Figure 5.11 shows while Boxed-sparsity (*Eq.*5.2.1) outperforms the base-line (*i.e.,* $k-$means) only if λ_3 is chosen close to the ground-truth, group-sparsity is very robust and it outperforms the $k-$means for relatively large values of λ_3.

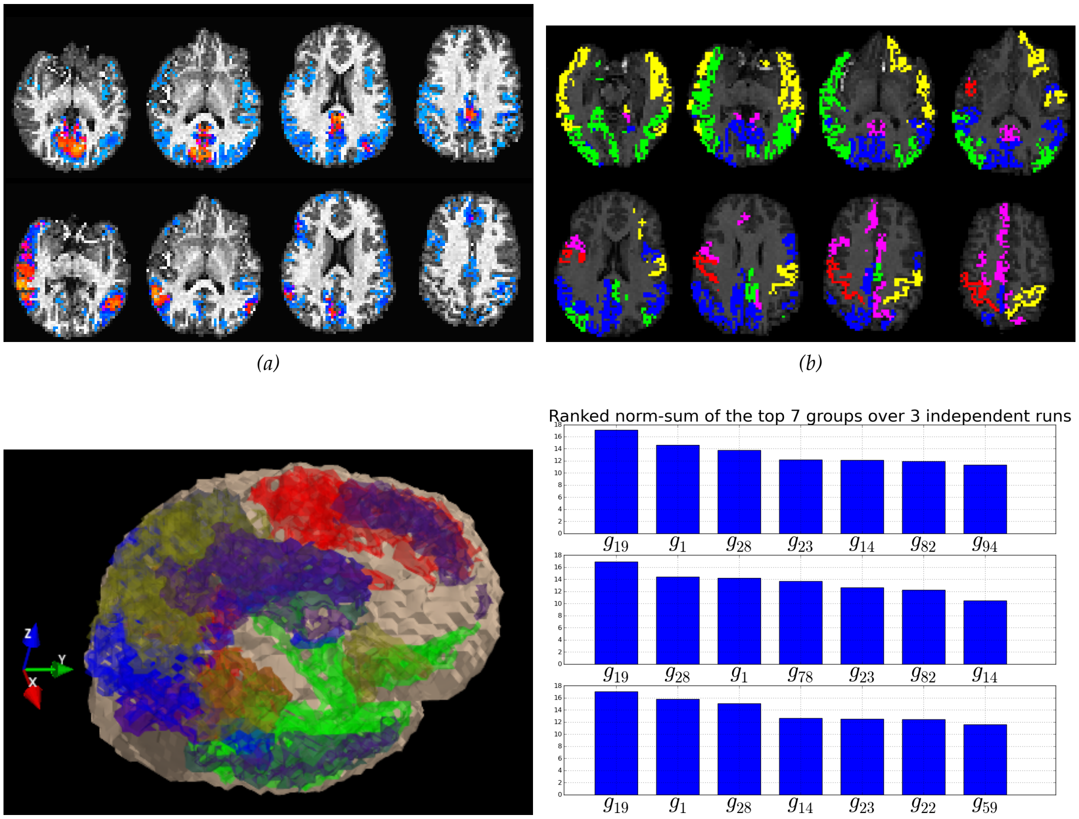
5.3.3 Network Recovery: fMRI Data

For *Default-Mode-Network* (DMN) in rs-fMRI, we have neither quantitative ground-truth nor we have two cohorts of subjects (*i.e.,* normal vs. abnormal) for classification. Therefore, to evaluate our results, we compare them what is reported in clinical literature and anatomical knowledge. For qualitative assessment of the algorithm, we selected 50 controls (23 female, 27 male, mean age 13.4 ± 3.38) from the open-access ADHD-200 dataset released by NITRC[5]. The resting BOLD fMRI scan for each subject ($360s$ length, TR = $2s$) first underwent standard fMRI preprocessing (motion correction, de-trending, smoothing). We then took the residual series of each scan after nuisance regression on motion and mean WM and CSF signals, masked it to include only gray matter (GM), and non-rigidly registered it to a template subject. We constructed 99 groups for our experiments; 25 of which are defined based on strong structural connectivity as described in the Human Connectome project[6] ([123] explains how the structural connectivities are inferred), while the rest are individual regions defined by Freesurfer parcellation. The areas of brain represented by the groups are listed in Table 5.2. Notice that multiple parcellation methods or different approaches such as fiber-tracking could be used to specify the groups.

To reduce computational cost, each experiment was limited to 30 basis vectors. We tested λ values of 5, 10, and 15 and did not observe significant differences, though with $\lambda = 15$ the basis vectors have higher budget and group norms can achieve values closer to 1. Figure 5.12a illustrates a few cuts of the top basis vectors, ranked according to maximum value. We focused on the top five basis vectors, sorted the groups comprising each of these basis according to their norm, and focused on the top 50% of non-zero groups. A common observation was that these top groups tended to be either single areas considered to be involved in the DMN (Default-Mode-Network), *e.g.,* superior parietal, or combinations defined on DMN seed regions. More significantly, multiple DMN areas tended to be ranked highly in the same basis; for example, combinations based on the posterior cingulate cortex (PCC) and bilateral middle temporal gyri repeatedly appeared together in the top 5 basis vectors. Some basis vectors contained groups which are not known to be part of the DMN, such as the rostral middle frontal gyrus ROI, but these were typically ranked lower than DMN-associated groups. To summarize the experiment, we add $\|\mathbf{v}_{|g}\|_2$ of the groups in all basis vectors of the experiment and rank the groups in descending order based on the values in order to find the most selected groups. Figure 5.12b shows

[5] http://fcon_1000.projects.nitrc.org/indi/adhd200/

[6] http://www.humanconnectomeproject.org


Z
Y
X
Ranked norm-sum of the top 7 groups over 3 independent runs
0
2
4
6
8
10
12
14
16
18
g_{19} g_1 g_{28} g_{23} g_{14} g_{82} g_{94}
g_{19} g_{28} g_1 g_{78} g_{23} g_{82} g_{14}
g_{19} g_1 g_{28} g_{14} g_{23} g_{22} g_{59}

(a) *(b)*

(c) *(d)*

Figure 5.12: (a) shows example of two of the basis vectors on different axial cuts. (b) shows the top 5 groups (different colors) in an experiment. (c) is 3D visualization of (b). To show the repeatability, (d) summarizes the index of top 7 groups based on sum of $\|\mathbf{v}_{|g}\|_2$ in all basis vectors in each of the 3-folds.

the top 5 groups denoted in the different colors. We used a similar measure to evaluate the repeatability of the algorithm. The 90 subjects were divided into three cohorts (3-folds) and the results of group ranking were compared across the three runs. Figure 5.12d shows only the top 7 groups for the three runs for each fold. The figure shows that the algorithm is very consistent, the first group is always g_{19}, which contains regions connected to the left PCC, and is followed by g_1 and g_{28}, which are based on the left middle temporal and left precuneus respectively. The rest of the groups are also consistent although the order may vary slightly in the three runs.

Table 5.2: Name of the areas in the brain used to define the groups for the fMRI experiment. **LH**,**RH**, and **BH** stand for left hemisphere, right hemisphere, and both hemispheres respectively. **WM** stands for white matter.

g_{1}	BH left middle temporal	g_{34}	LH parahippocampal	g_{67}	RH parahippocampal
g_{2}	BH right inferior parietal	g_{35}	LH precuneus	g_{68}	LH lateralorbitofrontal
g_{3}	LH right PCC	g_{36}	RH entorhinal	g_{69}	RH temporalpole
g_{4}	LH right middle temporal	g_{37}	LH frontalpole	g_{70}	RH fusiform
g_{5}	RH leftparahippocampal	g_{38}	LH precentral	g_{71}	LH temporal pole
g_{6}	RH right inferior parietal	g_{39}	RH isthmus cingulate	g_{72}	RH rostral middle frontal
g_{7}	LH right inferior parietal	g_{40}	RH cuneus	g_{73}	LH rostral anterior cingulate
g_{8}	RH rightprecuneus	g_{41}	RH middle temporal	g_{74}	LH inferior temporal
g_{9}	LH left inferior parietal	g_{42}	RH inferior temporal	g_{75}	RH lateral occipital
g_{10}	RH right middle temporal	g_{43}	rh rostral anterior cingulate	g_{76}	RH caudal anterior cingulate
g_{11}	LH left parahippocampal	g_{44}	RH inferior parietal	g_{77}	LH insula
g_{12}	RH left precuneus	g_{45}	LH paracentral	g_{78}	LH superior parietal
g_{13}	BH right precuneus	g_{46}	RH transversetemporal	g_{79}	RH bankssts
g_{14}	BH right middle temporal	g_{47}	LH supramarginal	g_{80}	RH caudalmiddlefrontal
g_{15}	RH right PCC	g_{48}	LH fusiform	g_{81}	LH postcentral
g_{16}	RH left PCC	g_{49}	RH precuneus	g_{82}	LH rostralmiddlefrontal
g_{17}	LH left PCC	g_{50}	LH middletemporal	g_{83}	RH postcentral
g_{18}	RH left middle temporal	g_{51}	RH pericalcarine	g_{84}	LH parsopercularis
g_{19}	BH left PCC	g_{52}	LH entorhinal	g_{85}	RH superior parietal
g_{20}	BH right PCC	g_{53}	RH paracentral	g_{86}	LH lingual
g_{21}	RH right parahippocampal	g_{54}	LH cuneus	g_{87}	LH parsorbitalis
g_{22}	LH left middle temporal	g_{55}	LH bankssts	g_{88}	LH transverse temporal
g_{23}	BH left inferior parietal	g_{56}	LH medial orbitofrontal	g_{89}	RH frontalpole
g_{24}	LH right parahippocampal	g_{57}	LH superior temporal	g_{90}	LH pericalcarine
g_{25}	BH left parahippocampal	g_{58}	RH pars triangularis	g_{91}	LH isthmus cingulate
g_{26}	LH right precuneus	g_{59}	RH precentral	g_{92}	LH superiorfrontal
g_{27}	LH left precuneus	g_{60}	RH lingual	g_{93}	RH supramarginal
g_{28}	BH left precuneus	g_{61}	LH inferior parietal	g_{94}	LH lateraloccipital
g_{29}	RH left inferiorparietal	g_{62}	RH parsorbitalis	g_{95}	RH posteriorcingulate
g_{30}	BH right parahippocampal	g_{63}	RH insula	g_{96}	LH caudal middle frontal
g_{31}	RH parsopercularis	g_{64}	LH caudal anterior cingulate	g_{97}	RH medial orbitofrontal
g_{32}	LH parstriangularis	g_{65}	RH superior temporal	g_{98}	LH posterior cingulate
g_{33}	RH lateral orbitofrontal	g_{66}	RH superior frontal	g_{99}	Background and WM

5.4 Conclusion

We proposed a framework that exploits all channels in a dataset simultaneously to reduce dimensionality in a discriminative yet interpretable way. Inspired by multi-view learning, two variants of constrained tensor factorization are suggested each of which implies different hypothesis about the data. We showed that the algorithm is relatively robust with respect to choice of parameters and achieves good classification results.

In the fMRI experiment, we proposed a method that bridges user- and data-driven approaches to infer functional connectivity. It allows prior knowledge about brain structures (*e.g.,* fiber-tracking) to be incorporated to guide this inference. It was shown that the method improves robustness compared to the $k-$means on the synthetic data and finds areas reported frequently in the clinical literature as belonging to the default-mode-network.

Chapter 6

Semi-Supervised Learning

6.1 Introduction

Medical imaging community frequently relies on voxel-wise image analysis to define areas of difference between groups [10] or to extract features for classification. However, this approach is not well suited for identifying complex population differences because it does not take into account the multivariate relationships in data [20, 40]. Moreover, regions showing significant group difference are not necessarily discriminative for classifying individuals. In order to overcome these limitations, high-dimensional pattern classification methods have been proposed in the recent literature [80, 92]. A fundamental limitation of these methods with respect to medical imaging is their need for large training sets of labeled data. One way to address this issue is to train the methods using unlabeled data, which may exist in large quantities. However, it is not clear how to exploit unlabeled data for dimensionality reduction. We will explore these topics in the subsequent sections.

Semi-supervised learning refers to a class of machine learning techniques that simultaneously use both labeled and unlabeled data for training in settings in which a small amount of labeled data and a large amount of unlabeled data are available. Semi-supervised learning combines elements of unsupervised and supervised learning. In many medical imaging applications, such situations arise either due to the availability of abundant sample images with no labels, or more importantly due to uncertainty about the labels. For example, subjects may deviate from the normal population and may be diagnosed with a certain disease in future follow-up scans; class labels of such subjects are not very well-defined. This is the case for subjects diagnosed as Mild Cognitive Impairment (MCI) who show some impairment in their cognitive scores and have high risk to develop Alzheimer's disease (AD) in near future [99]. One may be interested to predict future follow-up labels (converging to AD or not) of the MCI subjects by considering them as unlabeled data. Considering MCI subjects as unlabeled data allows an algorithm to locate unlabeled subjects in the spectrum of normal vs. abnormal. Recently, several methods have been proposed to address this issue. Sabuncu *et al.* [180] and Blezek *et al.* [33] proposed different frameworks for joint image registration and clustering that can exploit unlabeled images. Ribbens *et al.* [174] suggested a probabilistic method that can incorporate prior clinical information.

Our proposed method is based on techniques proposed in the previous chapters. As explained in Chapter 3, our method has two building blocks: Generative and Discriminative. The Generative block attempts to find a low rank decomposition of the data and in effect, it clusters voxels together given the constraint defined in Chapter 4. The discriminative part of the method seeks to classify subjects given the the decomposition of the generative block. This framework can be readily extended to the semi-supervised setting. The unlabeled data can contribute in the generative term and help the discriminative task indirectly by imposing a better regularization.

Section 6.2 briefly sums up our general framework that we expand upon. The section also presents the extension for semi-supervised setting and finally in Section 6.3 the applicability of the method is investigated on the bench-mark and real data in a few experiments.

6.2 Method

Dimensionality reduction is typically applied to achieve a generalizable classification rate when number of samples is less than dimensionality of features. We propose to use regularized matrix factorization formalism for dimensionality reduction. This framework allows to keep the semantics of images; hence it produces interpretable results.

In this section, we lay out the general framework. Regularized matrix factorization decomposes a matrix into two or more matrices such that the decomposition describes the matrix as accurate as possible. Such a decomposition could be subjected to some constraints or priors. Let assume that columns of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]$ represent observations (*i.e.,* sample images that are vectorized), and $\mathbf{B} \in \mathbb{R}^{D\times r}$ and $\mathbf{C} \in \mathbb{R}^{r\times N}$ decompose the matrix such that $\mathbf{X} \approx \mathbf{BC}$. r is number of basis vectors which is a parameter of the algorithm, D is number of voxels of images and N is number of samples. The columns of matrix $\mathbf{B}$ (called $\mathbf{b}_k$) can then be viewed as basis vectors and the i'th column of $\mathbf{C}$ (called $\mathbf{c}_i$) contains corresponding loading coefficients or weights of the basis vectors for the i'th observation. The columns $b_k \in \mathcal{B}$ and $c_i \in \mathcal{C}$ are subjected to some constraints, which we denote with the feasibility sets $\mathcal{B}$ and $\mathcal{C}$. We use variable $y_i \in \{-1, 0, 1\}$ to denote labels of the subjects. Healthy subjects are denoted by 1 and abnormal ones by -1; 0 is used simply for unlabeled subject indicating that labels are not decided for them.

In order to define the feasible sets ($\mathcal{B}$), we need to elaborate the requirements that our algorithm should satisfy: 1) The basis vectors must be anatomically meaningful; this means that a constructed basis vector should correspond to contiguous anatomical regions preferably in areas which are biologically related to a pathology of interest. In other words, the basis vectors should not resemble spread disjoint voxels. Sparsity of the basis vectors, *i.e.,* a relatively small number of voxels with non-zeros values, encourages it to be more spatially localized. 2) The basis must be discriminative: we are interested in finding features, *i.e.,* projections onto the basis vectors, that construct spatial patterns that best differentiate groups. 3) The decomposition ($\mathbf{BC}$) should be a good representative of data without compromising the previous properties.

In this chapter, we assume that images are non-negative hence it is reasonable to impose non-negativity on $\mathbf{B}$ and $\mathbf{C}$. Thus, our proposed method can be viewed as a variant of non-negative matrix factorization (NMF). NMF [141], [68] is an additive model that is known to decompose images into meaningful parts that are shared across subjects; this property is favorable for our application. Such part-based decomposition property encourages basis vectors to be similar to anatomically meaningful parts of images (*i.e.,* Hippocampus, Caudate, *etc.*for brain images). We also assume that certain structures (*e.g.,* Hippocampus) of an anatomy of interest (*e.g.,* brain) are affected by the abnormality (*e.g.,* shrinkage of Hippocampus Alzheimer's disease); this property can be viewed as a sparsity constraint on the basis vectors which also help the basis vectors to be more interpretable. We encode these properties via non-negativity on the coefficients and combination of non-negativity and ℓ_1 and ℓ_∞ norms on the basis vectors. The ℓ_1-norm encourages the sparsity property and combination of ℓ_∞ and non-negativity promotes part-based decomposition:

$$\begin{aligned} &\mathcal{C} := \{\mathbf{c} \in \mathbb{R}^r : \mathbf{c} \geq 0\} \\ &\mathcal{B} := \{\mathbf{b} \in \mathbb{R}^D : \mathbf{b} \geq 0, \|\mathbf{b}\|_\infty \leq 1, \|\mathbf{b}\|_1 \leq \lambda_3\} \end{aligned} \tag{6.2.1}$$

where ratio of λ_3/D encodes ratio of sparsity of the basis vectors. In order to find optimal $\mathbf{B}$ and $\mathbf{C}$, we define the following constrained optimization problem:

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{C},\mathbf{w}\in\mathbb{R}^r} \; \mathcal{D}(\mathbf{X};\mathbf{BC}) + \sum_{i\in\mathcal{L}} \ell(y_i; \langle \mathbf{B}^T\mathbf{x}_i, \mathbf{w}\rangle) + \|\mathbf{w}\|_2 \\ &\text{subject to:} \qquad \mathbf{b}_k \in \mathcal{B}, \quad \mathbf{c}_i \in \mathcal{C} \end{aligned} \tag{6.2.2}$$

The cost function of the optimization problem consists of two terms: 1) *Generative* term ($\mathcal{D}(\cdot;\cdot)$) that encourages the decomposition, $\mathbf{BC}$, to be close to the data matrix ($\mathbf{X}$); both labeled and unlabeled data contribute to this term. 2) *Discriminative* term ($\ell(y_i; f(\mathbf{x}_i, \mathbf{B}, \mathbf{w}))$) is a *loss* function that encourages a classifier $f(\cdot)$ to produce class labels that are consistent with available labels ($\mathbf{y}$). The classifier parametrized by $\mathbf{w}$, projects each image ($\mathbf{x}_i$) on the basis vectors to produce new features ($\mathbf{v}_i = \mathbf{B}^T\mathbf{x}_i$) and produces a labels. In this chapter, we use a linear classifier, hence $f(\mathbf{x}_i, \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T\mathbf{x}_i, \mathbf{w}\rangle$. Only labeled data contribute to the discriminative term. Various choices are possible for $\mathcal{D}(\cdot;\cdot)$ and $\ell(\cdot;\cdot)$. Here, we set $\mathcal{D}(\mathbf{X};\mathbf{BC}) = \lambda_1\|\mathbf{X} - \mathbf{BC}\|_F^2$ where λ_1 is a constant. For the loss function, we choose a hinge squared loss function: $\ell(y, \tilde{y}) = (\max\{0, 1 - y\tilde{y}\})^2$ which is a common choice in Support Vector Machine literature. Summing over $\mathcal{L}$ for the loss function simply indicates that the labeled subjects participate in this term.

In case of semi-supervised learning in our method, some subjects have certain labels (denoted by $\mathbf{X}_L$) and some subjects do not have labels (denoted by $\mathbf{X}_U$). In other words, the data matrix ($\mathbf{X}$) can be partitioned into two sub-matrices, namely $\mathbf{X} = [\mathbf{X}_L \quad \mathbf{X}_U]$. Our generative-discriminative framework can easily handle such cases. Recall the objective function of the optimization problem in *Eq.*(6.2.2); it was decomposed into three terms: generative term ($\mathcal{D}(\cdot;\cdot)$), discriminative term ($\ell(\cdot;\cdot)$), and regularization term. $\mathbf{X}_L$ contributes in both generative and discriminative terms while $\mathbf{X}_U$ only contributes in the generative term, namely:

$$\begin{aligned} &\boldsymbol{\Theta} = \{\mathbf{B}, \mathbf{C}, \mathbf{w}\} \\ &\mathcal{J}(\boldsymbol{\Theta}) = \mathcal{D}([\mathbf{X}_L, \mathbf{X}_U]; \boldsymbol{\Theta}) + \ell(\mathbf{y}; \mathbf{X}_L; \boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta}) \end{aligned} \tag{6.2.3}$$

in which $\boldsymbol{\Theta}$ is introduced to simplify the notation by grouping all parameters into $\boldsymbol{\Theta}$, $\mathcal{J}(\cdot)$ denotes the objective function, $\mathcal{R}(\cdot)$ stands for the regularization term. *Eq.*(6.2.3) shows that unlabeled samples are not penalized in the discriminative term (the second term) because the true labels are not available for them. This setting will be validated in Section 6.3.

6.3 Experiments

In this section, we investigate the performance of the extension of our method to semi-supervised learning. In order to examine the effectiveness of the proposed method for semi-supervised learning, we performed two sets of experiments. In the first set of experiments, the proposed method is compared with well-established semi-supervised methods on a benchmark data published earlier by Schölkopf *et al.* [42]. In the second sets of experiments, we apply the method on real medical images acquired from the ADNI dataset.

6.3.1 Experiment with Benchmark Datasets

Table 6.1 compares accuracy rates of the proposed method with those of three well-established semi-supervised learning methods on three datasets of a publicly available benchmark [42]. Although the setting in [42] is not in favor of our method and the proposed method is designed to address semi-supervised learning for medical image data, the results can show the soundness of the method in a very general context. Full descriptions of the datasets and pre-processing steps are elaborated in [42] but briefly:

- `USPS` : It is a dataset consisting of 150 images of each of the ten digits randomly drawn from the USPS set of handwritten digits. The digits “2” and “5” were assigned to the class +1, and all the others formed class -1. The images were obscured by application of algorithm 21.1 in [42] to prevent people from exploiting spatial relationship of features in the images [42]; more specifically for this dataset: $D = 241$ and $N = 1500$.
- `Text` : This is the `5 comp.*` groups from the `Newsgroups` dataset and the goal is to classify the `ibm` category versus the rest (provided by Tong *et al.* [200]); more specifically for this dataset: $D = 11,960$ and $N = 1500$.
- `BCI` : This dataset originates from research toward the development of a brain computer interface (BCI) (Lal *et al.* [135]). In each trial, EEG (electroencephalography) was acquired

Table 6.1: Comparison of classification error rates on a semi-supervised benchmark [42] between the semi-supervised extension of the proposed method and a few well-established methods. SSL-Bx stands for Boxed-Sparsity constrained formulation in the semi-supervised setting (Section 6.2)

	USPS	Text	BCI	
SSL-Bx	21.6	35.5	**47.23**	($N_{\text{label}} = 10$)
Linear TSVM	30.66	**28.6**	50.04	
non-Linear TSVM	25.20	31.21	49.15	
lapSVM	**19.05**	37.28	49.25	
SSL-Bx	13.1	24.8	**29.19**	($N_{\text{label}} = 100$)
Linear TSVM	21.12	**22.31**	42.67	
non-Linear TSVM	9.77	24.52	33.25	
lapSVM	**4.7**	23.86	32.39	

from a single subject from 39 electrodes. An autoregressive model of order 3 was fitted to each of the resulting 39 time series. The trail was represented by the total of $117 = 39 \times 3$ fitted parameters; more specifically for this dataset: $D = 117$ and $N = 400$.

In Table 6.1, in the first four rows, number of label samples (N_{label}) are set to 10 and in the second four rows, it is set to 100. The Table reports error rates for non/linear Transductive Support Vector Machine (TSVM) [121], Laplacian SVM (lapSVM) [188], which are chosen due to their good performance on the three datasets, in addition to the error rate for the proposed method. Entries of the table for lapSVM and non/linear-TSVM are adopted from [42]. According to [42], hyper-parameters of each of the algorithms are chosen by minimizing the test error, which is not possible in real applications; however, the results of this procedure can be useful to judge the potential of a method. To be comparable, similar procedure was applied to find λ_1/λ_2, λ_3/D and K for our algorithm.

Table 6.1 shows that no method consistently outperforms other methods across datasets; however, the results are consistent on each dataset. It shows that although our method outperforms others only on the BCI dataset, it is within a reasonable range of the best performance. This result motivates us to employ the semi-supervised extension of our method on a real medical image data.

6.3.2 Semi-Supervised Learning on a Brain Image Dataset

In this experiment 238 structural MRI images of MCI subjects were acquired from the ADNI dataset and used as unlabeled data. All 238 MCI subjects have at least 2 scans corresponding to 24-36 months follow-ups. Among 238 subjects, 99 patients have converted to AD at some point by their third year follow ups (MCI-C) and 139 did not convert after three years (MCI-NC). AD and NC subjects explained in the Chapter 4 were used as labeled data and the MCI subjects (MCI-C/MCI-NC) were used an unlabeled data. RAVENS maps of the images were computed by the same pre-processing pipeline as those of AD and NC subjects explained in the Section 4.6. Labeled data (AD/NC) is divided to 20 folds; data from 19 folds plus unlabeled data (MCI subjects) is used to learn the basis vectors. One fold out of 20 folds of the labeled data plus the unlabeled data were used for testing. In order to avoid searching for the best parameters, the most frequently selected parameters in the Section 4.6.3 were used as the parameters. Both variants of the regularizers introduced in Chapter 4: the Boxed-Sparsity (*Eq.*4.2.1), and the Group-Sparsity (*Eq.*4.3.2). For Group-Sparsity, similar to the Chapter 4, all images are registered to a template and an image partitioning (image segmentation) is available for the template image (*e.g.,* an anatomical parcellation in a template space). We used the support of each segmentation (*i.e.,* brain area) to define the groups.

To evaluate the performance of the algorithm, accuracy rates on the labeled data (AD/NC) and recall rates on the unlabeled data are reported in Table 6.2 for both regularization types. Since unlabeled data is shared between 20 folds, the recall rates (true positive and true negative rates depending on the class label) are averaged among 20 folds.

Table 6.2 shows the results for the semi-supervised learning, `SSL-Bx/Grp` represent semi-supervised learning for the Boxed- and Group-Sparsity constraints respectively. The classification accuracy rates for the labeled data have been improved slightly for the Boxed-Sparsity compared to the Table 4.1 meaning that unlabeled data can help improving the classification ac-

Table 6.2: This table shows application of the algorithm in a semi-supervised setting on the ADNI. The accuracy and recall rates (True-Positive and True-Negative rates) for labeled (AD/NC) and unlabeled data (MCI-C/MCI-NC) are shown in the table. *ssl-Bx* and *ssl-Grp* indicate semi-supervised setting of the proposed algorithm with the Boxed-Sparsity and Group-Sparsity constraints respectively.

	Accuracy	Recall	
	`AD vs NC`	`MCI-C`	`MCI-NC`
SSL-Bx	87.2%(±14.9%)	79.3%(±6.5%)	44.6%(±5.8%)
SSL-Grp	88.9%(±12.3%)	85.4%(±3.6%)	39.9%(±5.9%)

curacy for the labeled data. While the recall rates show high values for the MCI-C group, they demonstrate low recall rates for the MCI-NC group. Such low values can partly be justified by the fact that the patients in the MCI-NC group have not converted to the AD group yet but they may convert in the future. In addition, the labeled data anchored the classifiers to produce valid results for the AD/NC groups and avoid a case in which all data are assigned to one class. Therefore, Area Under Curve (AUC) of the classifiers should be investigated for further evaluation of the method.

Note that for all values reported in Table 6.2, basis vectors (hence features) extracted in the semi-supervised way but the classifiers are supervised (Logistic Model Trees [137]). One question would be whether a semi-supervised classifier can improve the results. Therefore, we designed an experiment to answer multiple questions: 1) Whether it is helpful to feed the features extracted using semi-supervised basis learning to a semi-supervised classifier instead of a supervised classifier; 2) Whether our semi-supervised basis learning is useful when there are few labeled samples; 3) How the number of labeled samples and different configurations of (semi-)supervised basis learning and (semi-)supervised classifiers affect AUC for MCI subjects.

For computational efficiency, the basis vectors $\mathbf{B}$ were learned only from 79 MCI subjects (as unlabeled data), and 20 AD and 20 NC subjects (as labeled data). The labeled subjects were divided into five folds for cross validation ($4/5$ for training and $1/5$ for testing) and the 79 MCI subjects were shared as unlabeled data across folds. In order to investigate the effect of number of labeled data, we performed four basis learning experiments by increasing number of revealed labels from 4 to 32; each fold has $4/5 \times (20 + 20) = 32$ AD/NC subjects and we revealed labels of AD/NC subjects as: $\{(2,2),(4,4),(8,8),(16,16)\}$. Rest of MCI subjects (*i.e.,* $238 - 79 = 159$) and AD/NC subjects that do not contribute in the basis learning are added to the testing lists for each fold.

After basis learning, features are extracted by projecting all images on the learned basis vectors. These features were fed into a supervised-classifier (Logistic Model Trees [137]) and a semi-supervised classifier (linear Laplacian SVM [24]) to produces labels. To have a reference point for comparison, we also learned the basis without unlabeled data (supervised basis learning). Figure 6.1 plots accuracy rates of AD/NC with respect to the number labeled data in different settings. The accuracy rates were computed on the left-out labeled data and the rest of the labeled data that was not introduced during the basis learning or training of the classifier. For brevity, **SF** in Figure 6.1 indicates Supervised Features, *i.e.,* using only labeled data to learn the basis vectors, and **SSF** denotes Semi-Supervised Features, *i.e.,* using the labeled and the unlabeled data to learn the basis vectors. The figure shows different scenarios for classification: supervised features fed into a supervised classifier (**SF** + **SC**) and a semi-supervised classifier (**SF** + **SSF**) and compares them with with semi-supervised features fed into a supervised classifier (**SSF** + **SC**) and a semi-supervised classifier (**SSF** + **SSF**).Figure 6.1a and Figure 6.1b show accuracy rates and AUC for the MCI respectively when the Boxed-sparsity is used for regularization and Figure 6.1c and Figure 6.1d represent the same quantifies when the Group-sparsity is applied as the sparsity regularization.

The results shown in Figure 6.1 can be summarized as follows:

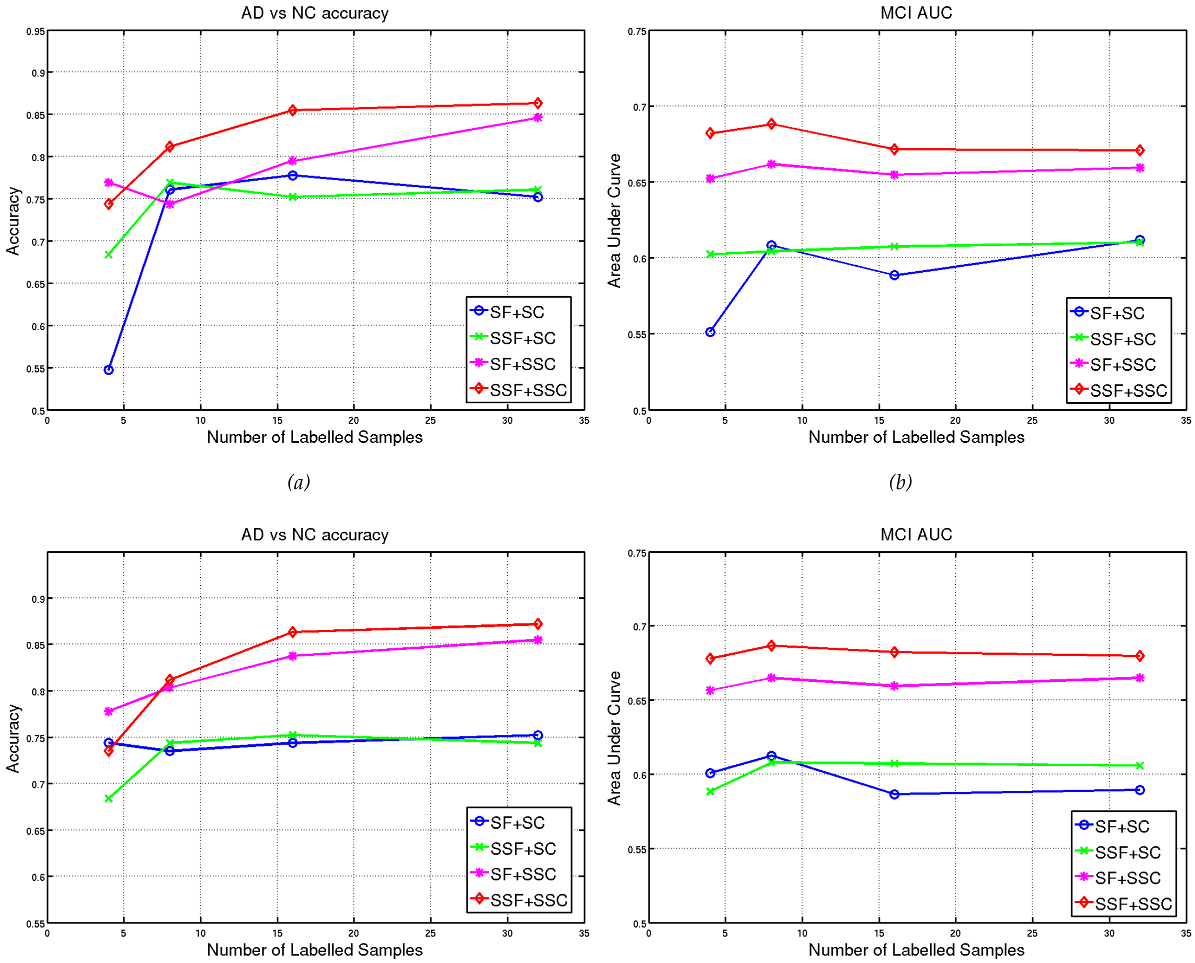
- *semi-supervised basis learning helps:* in all scenarios semi-supervised features (**SSF**) which are extracted by basis vectors learned in presence of unlabeled data outperform their corresponding supervised features (**SF**). Significant difference can be seen when the semi-supervised features are fed into semi-supervised classifier (*i.e.,* **SSF+SSC**) which achieves the best performance for both measures particularly for the Boxed-Sparsity.

- *semi-supervised classifier helps:* in all scenarios in Figure 6.1 semi-supervised classifiers (*i.e.,* **SF+SSC** and **SSF+SSC**) outperform their corresponding supervised classifiers for both types of regularizations (Boxed-Sparsity: Figure 6.1a-6.1b, Group-Sparsity: Figure 6.1c-6.1d) and both measures (*i.e.,* accuracy and AUC).

Note that semi-supervised features are more stable in terms of performance even if they are fed into a supervised classifier; for example, compare **SF+SC** and **SSF+SC** in Figure 6.1b and Figure 6.1d. Also note that AUC measures are computed for MCI-NC/MCI-C subjects because there is no real ground truth for them; hence AUC might be a better measure to show that the classifiers are not biased toward one of the classes although good performances on the labeled data (*i.e.,* AD vs NC) already show this fact.

6.4 Conclusion

We presented a framework to reduce the dimension of image features in the presence of unlabeled data. Constrained matrix decomposition problem was adapted for generative and discriminative basis learning and extended to semi-supervised formulation. Semi-supervised dimensionality reduction method outperforms supervised dimensionality reduction for both classification tasks considered in our experiments, both in terms of classifier accuracy and area under curves (AUC).



(a)

(b)

(c)

(d)

Figure 6.1: The accuracy rates and Area Under Curve (AUC) versus different number of labeled samples for different regularizations. **SF** and **SSF** stand for supervised and semi-supervised features respectively *i.e.*, supervised basis learning with or without unlabeled data; **SC** and **SSC** denote supervised classifier (Logistic Model Trees [137]) or semi-supervised classifier (linear lapSVM) respectively. (a) The accuracy rates of AD/NC when the Boxed-Sparsity is used as regularization. (b) AUC for MCI-NC/MCI-C subjects when the Boxed-Sparsity is used as regularization. (c) The accuracy rates of AD/NC when the Group-Sparsity is used as regularization. (d) AUC for MCI-NC/MCI-C subjects when the Group-Sparsity is used as regularization.

Chapter 7

Conclusion and Future Research

In this thesis, we combined two learning paradigms, generative and discriminative, to address the curse of dimensionality for medical imaging classification applications. While in most methods for medical image classification feature extraction and classification are performed separately, in this thesis, we combined those two steps into one framework. We proposed a novel formulation that cast the problem as a large-scale constrained matrix factorization. The formulation, in effect, chooses a subset of the rows (voxels of the images) and classifies the columns (subjects). The method allowed us to reduce the dimension without compromising classification or to produce clinically meaningless results (Chapter 3). The experiments with the synthetic and the real images in Chapter 3 showed how a balanced choice between generative and discriminative terms can help us to recover areas of difference between two classes of images. The extension of the method to the multi-channel case was also straightforward by changing our view from the data matrix to the data tensor. Such a change in the perspective allowed us to account for two scenarios: modalities characterizing the same tissue type ($\texttt{Multi-View}(\mathbb{X}, \mathbf{y})$) and different tissue type ($\texttt{Multi-View}(\mathbf{y})$). Experiments in the Chapter 5 on a few multi-channel datasets showed a superior classification performance with respect to the state-of-the method [80].

There are also several avenues for improvement which are left for future work:

- **Automatic Parameter Selection:** There are a few parameters (*i.e.,* $\lambda_1, \lambda_2, \lambda_3$) that need to be tuned using cross-validation. Although we have showed in Chapter 3 (see Section 3.6.3) that the algorithm is relatively robust with respect to a wide range of parameters, in order to achieve a high classification rate, one needs to do cross-validation inside of the training set. We have also suggested an ad hoc method to set the parameters in the Section 4.5. Another method is to estimate λ's from the data using a similar method to the Bayesian approaches such as Relevance-Vector-Machine (RVR) [199].
- **Incorporating Orthogonality as a Constraint:** We observed in our experiments on real brain images that increasing the weight on the discriminative term encourages many of the basis vectors to be similar to each other. Currently, we do not have any term to push the basis vectors away from each other. Given that $\mathbf{b}_i$'s are all non-negative, pushing $\mathbf{b}_i$'s to be dissimilar can be viewed as an orthogonality constraint which is difficult to impose in our current block-wise convex formulation. In addition, imposing a strict orthogonality is not favourable for our problem (we would like basis to have some level of overlap) but we would prefer a soft version of the orthogonality constraint. There are a few works that address similar problem [229], [164] but they break the block-wise convexity of the formulation which is essential for its computational efficiency.
- **Other Variants of the Groups Sparsity:** The Boxed-Sparsity regularizer ignores the underlying structure of an image because it simply concatenates voxels of an image into a vector. One approach to account for image structure could be to incorporate smoothness on the $\mathbf{b}_i$'s (*e.g.,* different variants of the TV-norm in Chapter 2) to encourage smoothness. In [20], we used TV_2^1-norm as a regularizer. We empirically realized that the algorithm yields similar results if the images are smoothed before being fed to the algorithm. This is not the case for $TV_2^{1/2}$- or TV_1^1-norms however they impose significant computational cost on the optimization algorithm. For this reason, we always pre-smooth the images before applying the method. Another remedy to this problem is to use other variants of the sparsity norm

for the feasible set (*e.g.,* Group-Sparsity in Chapters 4 and 5). We showed in Section 5.2.3 that by allowing the groups to overlap, we can go beyond just image structure and consider long-range connection between areas of a brain. Nevertheless, the experimental results in the Section 5.3.3 are limited and further evaluation is needed.

- **A Better Regularization for Semi-Supervised Learning:** The modular nature of the method makes it readily extensible for semi-supervised, and unsupervised learning cases. The semi-supervised learning is important in medical imaging datasets when there are large sets of subjects not classified as normal but lacking fully confident disease labels (*e.g.,* MCI cases). The experiments in the Chapter 6 showed that the semi-supervised basis learning helps in term of predicting follow-up labels of the MCI subjects. It is also possible to add extra regularization to incorporate relationship between samples (neighborhood information) similar to the Laplacian SVM ($\ell-$SVM) [189]. In $\ell-$SVM the idea is as follows: the samples with similar features (*e.g.,* close to each other in ℓ_2 sense) should have similar class labels. Neighborhood information can be encoded via a graph Laplacian and can be added as regularization to the objective function. We have explored this idea in [17]. The graph Laplacian was build by measuring the amount of deformation to register every pair of the images. Comparing the results of the experiments in [17] and those of Chapter 6, we realized that the generative term has more impact in the semi-supervised learning than the Laplacian term; nevertheless further investigation is required to find the right balance between the generative term and a better Laplacian term to exploit unlabeled data.
- **Tightening the Relaxation:** Finding the optimal basis vectors requires solving a large-scale optimization problem. We relaxed the selection constraints for the voxels (*i.e.,* $\{0, 1\}$ was converted to $[0, 1]$) for the computational purposes. A novel technique based on the proximal method [49] was proposed to gain computational efficiency. Nevertheless, due to the relaxation, the entries of each basis vector are not necessarily $0/1$ but rather between 0 and 1; this makes an ambiguity for choosing a threshold. A remedy is to avoid the relaxation

and solve the combinatorial optimization problem. Since off-the-shelf discrete optimization solvers cannot tackle this problem due to its large-dimensionality, one may resort to a greedy methods to solve it. One way is to use matching pursuit methods [75], [76] although the scalability of such approaches for a large-scale problem is questionable. Since each $\mathbf{b}_i$ can be viewed as a subset, finding optimal K subsets (K is the number of $\mathbf{b}_i$'s) is even more difficult because number of possibilities to choose from (even with a greedily method) is exponential. An intelligent sampling method can potentially outperform simple greedy algorithm, nevertheless special attention is required to construct an efficient sampling algorithm [134].

- **Unifying with Registration:** Through out this thesis, we assumed that all images are registered to a common template and the images in the dataset are reconstructed using a linear combination of basis vectors and coefficients (the $\mathbf{BC}$ in the generative term). We can view the registration step as a generative process that generated the images by deforming the template image. This perspective allows us to extend the linear reconstruction to a non-linear one. We have partially studied this idea in [18] without the discriminative term. The idea was to linearly reconstruct the stationary velocity fields of the diffeomorphic registration that reside in the tangent space of the identity map (no deformation) in the template space. This approach can potentially unify the registration step within the framework nevertheless adding the discriminative term to the formulation imposes computational difficulties that need to be addressed in the future.

Bibliography

[1] The mosek optimization software.

[2] Wellcome department of imaging neuroscience.

[3] *The Journal of the American Medical Association*, 305(3):223–319, January 2011.

[4] Sophie Achard, Raymond Salvador, Brandon Whitcher, John Suckling, and Ed Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J Neurosci*, 26(1):63–72, January 2006.

[5] T. Acharya and A.K. Ray. *Image processing: principles and applications*. Wiley-Interscience, 2005.

[6] O. Acosta, P. Bourgeat, M.A. Zuluaga, J. Fripp, O. Salvado, and S. Ourselin. Automated voxel-based 3d cortical thickness measurement in a combined lagrangian–eulerian pde approach using partial volume maps. *Medical image analysis*, 13(5):730–743, 2009.

[7] M. Aharon, M. Elad, and A. Bruckstein. K-svd: Design of dictionaries for sparse representation. *Proceedings of SPARS*, 5:9–12, 2005.

[8] Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 25–32, New York, NY, USA, 2007. ACM.

[9] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[10] J. Ashburner and K. J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6 Pt 1):805–821, Jun 2000.

[11] A. Averbuch, S. Dekel, and S. Deutsch. Adaptive compressed image sensing using dictionaries.

[12] Suyash P Awate, Tolga Tasdizen, Norman Foster, and Ross T Whitaker. Adaptive markov modeling for mutual-information-based, unsupervised mri brain-tissue classification. *Med Image Anal*, 10(5):726–39, October 2006.

[13] R. Bakshi, S. Ariyaratana, R.H.B. Benedict, and L. Jacobs. Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions. *Archives of neurology*, 58(5):742, 2001.

[14] Martina Ballmaier, Arthur W Toga, Rebecca E Blanton, Elizabeth R Sowell, Helen Lavretsky, Jeffrey Peterson, Daniel Pham, and Anand Kumar. Anterior cingulate, gyrus rectus, and orbitofrontal abnormalities in elderly depressed patients: an mri-based parcellation of the prefrontal cortex. *Am J Psychiatry*, 161(1):99–108, January 2004.

[15] Richard Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, July 2007.

[16] Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA J Numer Anal*, 8(1):141–148, January 1988.

[17] Kayhan N. Batmanghelich, Dong Hye Ye, Kilian M. Pohl, Ben Taskar, and Christos Davatzikos. Disease classification and prediction via semi-supervised dimensionality reduction. In *ISBI*, pages 1086–1090. IEEE, 2011.

[18] N. Batmanghelich, A. Gooya, S. Kanterakis, B. Taskar, and C. Davatzikos. Application of trace-norm and low-rank matrix decomposition for computational anatomy. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 146–153. IEEE, 2010.

[19] Nematollah Batmanghelich, Aoyan Dong, Ben Taskar, and Christos Davatzikos. Regularized tensor factorization for multi-modality medical image classification. *Med Image Comput Comput Assist Interv*, 14(Pt 3):17–24, 2011.

[20] Nematollah Batmanghelich, Ben Taskar, and Christos Davatzikos. A general and unifying framework for feature construction, in image-based pattern classification. *Inf Process Med Imaging*, 21:423–434, 2009.

[21] Nematollah K Batmanghelich, Ben Taskar, and Christos Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Trans Med Imaging*, 31(1):51–69, January 2012.

[22] A. Baune, F.T. Sommer, M. Erb, D. Wildgruber, B. Kardatzki, G. Palm, and W. Grodd. Dynamical cluster analysis of cortical fmri activation. *NeuroImage*, 9(5):477–489, 1999.

[23] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[24] Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research*, 7:2399–2434, November 2006.

[25] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[26] Dimitri P. Bertsekas and Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999.

[27] Michael Biggs, Ali Ghodsi, and Stephen Vavasis. Nonnegative matrix factorization via rank-one downdate. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 64–71, New York, NY, USA, 2008. ACM.

[28] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan. Nonmonotone spectral projected gradient methods on convex sets. volume 10, pages 1196–1211, Philadelphia, PA, USA, 2000. Society for Industrial and Applied Mathematics.

[29] C.M. Bishop. Neural networks for pattern recognition. 1995.

[30] B.B. Biswal, M. Mennes, X.N. Zuo, S. Gohel, C. Kelly, S.M. Smith, C.F. Beckmann, J.S. Adelstein, R.L. Buckner, S. Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734, 2010.

[31] M.B. Blaschko, J.A. Shelton, A. Bartels, C.H. Lampert, and A. Gretton. Semi-supervised kernel canonical correlation analysis with application to human fmri. *Pattern Recognition Letters*, 2011.

[32] D.M. Blei and J.D. McAuliffe. Supervised topic models. *Arxiv preprint arXiv:1003.0783*, 2010.

[33] Daniel J Blezek and James V Miller. Atlas stratification. *Med Image Anal*, 11(5):443–457, Oct 2007.

[34] S. Bouix, J. Pruessner, D. Collins, and K. Siddiqi. Hippocampal shape analysis using medial surfaces. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*, pages 33–40. Springer, 2001.

[35] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

[36] R.G. Boyes, J.L. Gunter, C. Frost, A.L. Janke, T. Yeatman, D.L.G. Hill, M.A. Bernstein, P.M. Thompson, M.W. Weiner, N. Schuff, et al. Intensity non-uniformity correction using n3 on 3-t scanners with multichannel phased array coils. *Neuroimage*, 39(4):1752–1762, 2008.

[37] ET Bullmore, S. Rabe-Hesketh, RG Morris, SCR Williams, L. Gregory, JA Gray, MJ Brammer, et al. Functional magnetic resonance image analysis of a large-scale neurocognitive network. *Neuroimage*, 4(1):16–33, 1996.

[38] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[39] V.D. Calhoun, J. Liu, and T. AdalI. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

[40] Melissa K Carroll, Guillermo A Cecchi, Irina Rish, Rahul Garg, and A. Ravishankar Rao. Prediction and interpretation of distributed neural activity with sparse models. *Neuroimage*, 44(1):112–122, Jan 2009.

[41] E Ceyhan, M Hosakere, T Nishino, J Alexopoulos, R D Todd, K N Botteron, M I Miller, and J T Ratnanather. Statistical analysis of cortical morphometrics using pooled distances based on labeled cortical distance maps. *J Math Imaging Vis*, 40(1):20–35, May 2011.

[42] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[43] R. Chen and E.H. Herskovits. Graphical-model-based morphometric analysis. *Medical Imaging, IEEE Transactions on*, 24(10):1237–1248, 2005.

[44] Yong-Sheng Chen, Li-Fen Chen, Ya-Ting Chang, Yung-Tien Huang, Tong-Ping Su, and Jen-Chuen Hsieh. Quantitative evaluation of brain magnetic resonance images using voxel-based morphometry and bayesian theorem for patients with bipolar disorder. *Journal of Medical and Biological Engineering*, 3(28):127–133, 2008.

[45] Gael Chetelat, Beatrice Desgranges, Vincent De La Sayette, Fausto Viader, Francis Eustache, and Jean-Claude Baron. Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. *Neuroreport*, 13(15):1939–1943, Oct 2002.

[46] Moo K Chung, Keith J Worsley, Steve Robbins, TomÃ¡s Paus, Jonathan Taylor, Jay N Giedd, Judith L Rapoport, and Alan C Evans. Deformation-based surface morphometry applied to gray matter deformation. *Neuroimage*, 18(2):198–213, February 2003.

[47] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[48] Mayfield clinic for brain and spine. Spect (single photon emission computed tomography) scan, 2011. [Online; accessed 7-May-2012].

[49] P.L. Combettes and J.C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

[50] P.L. Combettes, V.R. Wajs, et al. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2006.

[51] A. Convit, J. de Asis, M. J. de Leon, C. Y. Tarshish, S. De Santi, and H. Rusinek. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to alzheimer's disease. *Neurobiol Aging*, 21(1):19–26, 2000.

[52] D. Cordes, V. Haughton, J.D. Carew, K. Arfanakis, and K. Maravilla. Hierarchical clustering to measure connectivity in fmri resting-state data. *Magn Reson Imaging*, 20:305–317, May 2002.

[53] Robert G. Cowell. *Probabilistic networks and expert systems*. Springer Verlag, 1999.

[54] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri) ”brain reading”: detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2 Pt 1):261–270, Jun 2003.

[55] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2:265–292, 2002.

[56] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2):201–233, 2002.

[57] W.R. Crum, T. Hartkens, and DLG Hill. Non-rigid image registration: theory and practice. *British journal of radiology*, 77(Special Issue 2):S140, 2004.

[58] Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali, Olivier Colliot, and The Alzheimer's Disease Neuroimaging Initiative. Automatic classification of patients with alzheimer's disease from structural mri: A comparison of ten methods using the adni database. *Neuroimage*, Jun 2010.

[59] Yu-Hong Dai, William W. Hager, Klaus Schittkowski, and Hongchao Zhang. The cyclic barzilai–borwein method for unconstrained optimization. *IMA J Numer Anal*, 26(3):604–627, July 2006.

[60] A. d'Aspremont d'Aspremont d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. *A direct formulation for sparse PCA using semidefinite programming*. Computer Science Division, University of California, 2004.

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

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

[63] Fani Deligianni, Gael Varoquaux, Bertrand Thirion, Emma Robinson, David J Sharp, A David Edwards, and Daniel Rueckert. A probabilistic framework to infer brain functional connectivity from anatomical connections. *Inf Process Med Imaging*, 22:296–307, 2011.

[64] D. Delis, J. Kramer, E. Kaplan, and B. Ober. *California Verbal Learning Test-Research Edition*. The Psychological Corporation, New York, NY, 1987.

[65] Inderjit S. Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. In *In: Neural Information Proc. Systems*, pages 283–290, 2005.

[66] J. Diehl, T. Grimmer, A. Drzezga, M. Riemenschneider, H. Förstl, and A. Kurz. Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. a pet study. *Neurobiol Aging*, 25(8):1051–1056, Sep 2004.

[67] S. Dodel, J.M. Herrmann, and T. Geisel. Functional connectivity by cross-correlation clustering. *Neurocomputing*, 44:1065–1070, 2002.

[68] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts, 2003.

[69] D.L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[70] K.B. Duan and S. Keerthi. Which is the best multiclass svm method? an empirical study. *Multiple Classifier Systems*, pages 732–760, 2005.

[71] J.M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *Image Processing, IEEE Transactions on*, 18(7):1395–1408, 2009.

[72] F DuBois Bowman, Brian Caffo, Susan Spear Bassett, and Clinton Kilts. A bayesian hierarchical framework for spatial modeling of fmri data. *Neuroimage*, 39(1):146–56, January 2008.

[73] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, December 2009.

[74] D. Dueck, Q.D. Morris, and B.J. Frey. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, 21(suppl 1):i144–i151, 2005.

[75] M. Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Verlag, 2010.

[76] M. Elad and M. Aharon. Image denoising via learned dictionaries and sparse representation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 895–900. IEEE, 2006.

[77] J.M. Fadili and G. Peyré. Total variation projection with first order schemes. *Image Processing, IEEE Transactions on*, 20(3):657–669, 2011.

[78] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, August 2008.

[79] Yong Fan, Nematollah Batmanghelich, Chris M Clark, Christos Davatzikos, and Alzheimer's Disease Neuroimaging Initiative. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4):1731–1743, Feb 2008.

[80] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur, and Christos Davatzikos. Compare: classification of morphological patterns using adaptive regional elements. *IEEE Trans Med Imaging*, 26(1):93–105, Jan 2007.

[81] S.R.A. Fisher. *Statistical methods for research workers*. Number 5. Genesis Publishing Pvt Ltd, 1932.

[82] Norman L Foster, Judith L Heidebrink, Christopher M Clark, William J Jagust, Steven E Arnold, Nancy R Barbas, Charles S DeCarli, R. Scott Turner, Robert A Koeppe, Roger Higdon, and Satoshi Minoshima. Fdg-pet improves accuracy in distinguishing frontotemporal dementia and alzheimer's disease. *Brain*, 130(Pt 10):2616–2635, Oct 2007.

[83] Nick C Fox and Jonathan M Schott. Imaging cerebral atrophy: normal ageing to alzheimer's disease. *Lancet*, 363(9406):392–394, Jan 2004.

[84] K. J. Friston, A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994.

[85] Karl Friston, Carlton Chu, Janaina Mourão-Miranda, Oliver Hulme, Geraint Rees, Will Penny, and John Ashburner. Bayesian decoding of brain images. *Neuroimage*, 39(1):181–205, Jan 2008.

[86] K.J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[87] R. Garg, G.A. Cecchi, and A.R. Rao. Full-brain auto-regressive modeling (farm) using fmri. *NeuroImage*, 2011.

[88] M.N. Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1997.

[89] Rainer Goebel, Alard Roebroeck, Dae-Shik Kim, and Elia Formisano. Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping. *Magn Reson Imaging*, 21(10):1251–61, December 2003.

[90] A. F. Goldszal, C. Davatzikos, D. L. Pham, M. X. Yan, R. N. Bryan, and S. M. Resnick. An image-processing system for qualitative and quantitative volumetric analysis of brain images. *J Comput Assist Tomogr*, 22(5):827–837, 1998.

[91] Polina Golland, Yulia Golland, and Rafael Malach. Detection of spatial activation patterns as unsupervised segmentation of fmri data. *Med Image Comput Comput Assist Interv*, 10(Pt 1):110–118, 2007.

[92] Polina Golland, W. Eric L. Grimson, Martha E. Shenton, and Ron Kikinis. Deformation analysis for shape based classification. In *IN IPMI*, pages 517–530. SpringerVerlag, 2001.

[93] Polina Golland, W. Eric L Grimson, Martha E Shenton, and Ron Kikinis. Detection and analysis of statistical differences in anatomical shape. *Med Image Anal*, 9(1):69–86, Feb 2005.

[94] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*, 14(1 Pt 1):21–36, Jul 2001.

[95] M.D. Greicius, B. Krasnow, A.L. Reiss, and V. Menon. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.*, 100:253–258, Jan 2003.

[96] L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

[97] L. Grippo and M. Sciandrone. Nonmonotone globalization techniques for the barzilai-borwein gradient method. *Computational Optimization and Applications*, 23(2):143–169, 2002.

[98] R. Grosse, R. Raina, H. Kwong, and A.Y. Ng. Shift-invariant sparse coding for audio classification. *cortex*, 9:8, 2007.

[99] Michael Grundman, Ronald C Petersen, Steven H Ferris, Ronald G Thomas, Paul S Aisen, David A Bennett, Norman L Foster, Clifford R Jack, Douglas R Galasko, Rachelle Doody, Jeffrey Kaye, Mary Sano, Richard Mohs, Serge Gauthier, Hyun T Kim, Shelia Jin, Arlan N Schultz, Kimberly Schafer, Ruth Mulnard, Christopher H van Dyck, Jacobo Mintzer, Edward Y Zamrini, Deborah Cahn-Weiner, Leon J Thal, and Alzheimer's Disease Cooperative Study. Mild cognitive impairment can be distinguished from alzheimer disease and normal aging for clinical trials. *Arch Neurol*, 61(1):59–66, Jan 2004.

[100] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh, editors. *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing. Springer, 1 edition, August 2006.

[101] Patric Hagmann, Lisa Jonasson, Philippe Maeder, Jean-Philippe Thiran, Van J Wedeen, and Reto Meuli. Understanding diffusion mr imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *Radiographics*, 26 Suppl 1:S205–23, October 2006.

[102] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update; sigkdd explorations. 11, 2009.

[103] J.H. Han, S. Yang, and B.U. Lee. A novel 3-d color histogram equalization method with uniform 1-d gray scale histogram. *Image Processing, IEEE Transactions on*, 20(2):506–512, 2011.

[104] Michael Hanke, Yaroslav O Halchenko, Per B Sederberg, Stephen JosÃ© Hanson, James V Haxby, and Stefan Pollmann. Pymvpa: A python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics*, 7(1):37–53, 2009.

[105] L.K. Hansen, J. Larsen, F.Å. Nielsen, S.C. Strother, E. Rostrup, R. Savoy, N. Lange, J. Sidtis, C. Svarer, and O.B. Paulson. Generalizable patterns in neuroimaging: How many principal components? *NeuroImage*, 9(5):534–544, 1999.

[106] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. *BeiJing: Publishing House of Electronics Industry*, 2004.

[107] Y. He, Y. Zang, T. Jiang, M. Liang, and G. Gong. Detecting functional connectivity of the cerebellum using low frequency fluctuations (lffs). *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 907–915, 2004.

[108] Chris Hinrichs, Vikas Singh, Guofan Xu, and Sterling Johnson. Mkl for robust multi-modality ad classification. *Med Image Comput Comput Assist Interv*, 12(Pt 2):786–794, 2009.

[109] Thomas Hofmann. Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[110] C.W. Hsu and C.J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[111] Xue Hua, Alex D Leow, Suh Lee, Andrea D Klunder, Arthur W Toga, Natasha Lepore, Yi-Yu Chou, Caroline Brun, Ming-Chang Chiang, Marina Barysheva, Clifford R Jack, Matt A Bernstein, Paula J Britson, Chadwick P Ward, Jennifer L Whitwell, Bret Borowski, Adam S Fleisher, Nick C Fox, Richard G Boyes, Josephine Barnes, Danielle Harvey, John Kornak, Norbert Schuff, Lauren Boreta, Gene E Alexander, Michael W Weiner, Paul M Thompson, and Alzheimer's Disease Neuroimaging Initiative. 3d characterization of brain atrophy in alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *Neuroimage*, 41(1):19–34, May 2008.

[112] Junzhou Huang and Tong Zhang. The benefit of group sparsity. Mar 2009.

[113] J.J. Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, 1994.

[114] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

[115] M. Ingalhalikar, S. Kanterakis, R. Gur, T. Roberts, and R. Verma. Dti based diagnostic prediction of a disease via pattern classification. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 558–565, 2010.

[116] T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. 1999.

[117] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[118] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proc. of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

[119] T. Jebara. *Machine learning: Discriminative and generative*, volume 755. Springer Netherlands, 2004.

[120] T. Joachims, T. Finley, and C.N.J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[121] Thorsten Joachims. Transductive learning via spectral graph partitioning. In *In ICML*, pages 290–297, 2003.

[122] M.I. Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. *Computational Cognitive Science Technical Report*, 9503, 1995.

[123] AA Joshi, SH Joshi, I. Dinov, DW Shattuck, RM Leahy, and AW Toga. Anatomical structural network analysis of human brain using partial correlations of gray matter volumes. In

Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on, pages 844–847. IEEE, 2010.

[124] Sarang Joshi, Stephen Pizer, P Thomas Fletcher, Paul Yushkevich, Andrew Thall, and J S Marron. Multiscale deformable model segmentation and statistical shape analysis using medial descriptions. *IEEE Trans Med Imaging*, 21(5):538–50, May 2002.

[125] Noor Jehan Kabani, David J. MacDonald, Colin J. Holmes, and Alan C. Evans. 3d anatomical atlas of the human brain. *NeuroImage*, 7:S717, 1998.

[126] Sham Kakade and Dean Foster. Multi-view regression via canonical correlation analysis. pages 82–96. 2007.

[127] J. A. Kaye, T. Swihart, D. Howieson, A. Dame, M. M. Moore, T. Karnos, R. Camicioli, M. Ball, B. Oken, and G. Sexton. Volume loss of the hippocampus and temporal lobe in healthy elderly persons destined to develop dementia. *Neurology*, 48(5):1297–1304, May 1997.

[128] Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[129] A. Klein, J. Andersson, B.A. Ardekani, J. Ashburner, B. Avants, M.C. Chiang, G.E. Christensen, D.L. Collins, J. Gee, P. Hellier, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*, 46(3):786–802, 2009.

[130] T.G. Kolda. *Multilinear operators for higher-order decompositions*. United States. Department of Energy, 2006.

[131] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[132] Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pages 426–434, 2008.

[133] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[134] A. Kulesza and B. Taskar. Structured determinantal point processes. In *Proc. NIPS*, 2010.

[135] Thomas Navin Lal, Michael Schröder, Thilo Hinterberger, Jason Weston, Martin Bogdan, Niels Birbaumer, and Bernhard Schölkopf. Support vector channel selection in bci. *IEEE Trans Biomed Eng*, 51(6):1003–1010, Jun 2004.

[136] S. M. Landau, D. Harvey, C. M. Madison, E. M. Reiman, N. L. Foster, P. S. Aisen, R. C. Petersen, L. M. Shaw, J. Q. Trojanowski, C. R. Jack, M. W. Weiner, W. J. Jagust, and Alzheimer's Disease Neuroimaging Initiative. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–238, Jul 2010.

[137] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. In *Machine Learning*, pages 241–252, 2003.

[138] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, pages 536–543. ACM, 2008.

[139] D. Lashkari, R. Sridharan, E. Vul, P.J. Hsieh, N. Kanwisher, and P. Golland. Search for patterns of functional specificity in the brain: A nonparametric hierarchical bayesian model for group fmri data. *NeuroImage*, 2011.

[140] Steffen L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, July 1996.

[141] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999.

[142] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[143] J.P. Lerch, J. Pruessner, A.P. Zijdenbos, D.L. Collins, S.J. Teipel, H. Hampel, and A.C. Evans. Automated cortical thickness measurements from mri can accurately separate alzheimer's patients from normal elderly controls. *Neurobiology of aging*, 29(1):23–30, 2008.

[144] F.H. Lin, A.R. McIntosh, J.A. Agnew, G.F. Eden, T.A. Zeffiro, and J.W. Belliveau. Multivariate analysis of neuronal interactions in the generalized partial least squares framework: simulations and empirical studies. *NeuroImage*, 20(2):625–642, 2003.

[145] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multitask lasso, with applications to neural semantic basis discovery. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 649–656. ACM, 2009.

[146] Gabriele Lohmann, Kirsten G Volz, and Markus Ullsperger. Using non-negative matrix factorization for single-trial analysis of fmri data. *Neuroimage*, 37(4):1148–60, October 2007.

[147] Peter Lorenzen, Brad Davis, and Sarang Joshi. Unbiased atlas formation via large deformations metric mapping. *Med Image Comput Comput Assist Interv*, 8(Pt 2):411–418, 2005.

[148] A.C. Lozano, G. Swirszcz, and N. Abe. Group orthogonal matching pursuit for variable selection and prediction. *Advances in Neural Information Processing Systems*, 22:1150–1158, 2008.

[149] JB Maintz and M.A. Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.

[150] J. Mairal. Sparse coding for machine learning, image processing and computer vision. 2010.

[151] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2010.

[152] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009.

[153] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. *Arxiv preprint arXiv:1008.5209*, 2010.

[154] S.G. Mallat. *A wavelet tour of signal processing*. Academic Pr, 1999.

[155] S.G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

[156] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fmri-based prediction of behaviour. *Medical Imaging, IEEE Transactions on*, (99):1–1, 2011.

[157] Vincent Michel, Alexandre Gramfort, GaÃ"¡l Varoquaux, Evelyn Eger, and Bertrand Thirion. Total variation regularization for fmri-based prediction of behavior. *IEEE Trans Med Imaging*, 30(7):1328–40, July 2011.

[158] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[159] Y. Nesterov. Gradient methods for minimizing composite objective function. core discussion papers 2007076, université catholique de louvain. *Center for Operations Research and Econometrics (CORE)*, 2007.

[160] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, pages 78+, New York, NY, USA, 2004. ACM.

[161] TR Oakes, T. Johnstone, KS Ores Walsh, LL Greischar, AL Alexander, AS Fox, and RJ Davidson. Comparison of fmri motion correction software tools. *Neuroimage*, 28(3):529–543, 2005.

[162] G. Obozinski, L. Jacob, and J.P. Vert. Group lasso with overlaps: the latent group lasso approach. *Arxiv preprint arXiv:1110.0413*, 2011.

[163] Guillaume Obozinski and Ben Taskar. Multi-task feature selection. Technical report, 2006.

[164] E. Oja and Z. Yang. Orthogonal nonnegative learning for sparse feature extraction and approximate combinatorial optimization. *Frontiers of Electrical and Electronic Engineering in China*, 5(3):261–273, 2010.

[165] J. Pearle. Probabilistic reasoning in intelligent systems, 1988.

[166] D. L. Pham and J. L. Prince. Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Trans Med Imaging*, 18(9):737–752, Sep 1999.

[167] Anqi Qiu, Timothy Brown, Bruce Fischl, Jun Ma, and Michael I Miller. Atlas generation for subcortical and ventricular structures with its applications in shape analysis. *IEEE Trans Image Process*, 19(6):1539–47, June 2010.

[168] X. Qu, W. Zhang, D. Guo, C. Cai, S. Cai, and Z. Chen. Iterative thresholding compressed sensing mri based on contourlet transform. *Inverse Problems in Science and Engineering*, 18(6):737–758, 2010.

[169] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[170] R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[171] R.D.S. Raizada, F.M. Tsao, H.M. Liu, I.D. Holloway, D. Ansari, and P.K. Kuhl. Linking brain-wide multivoxel activation patterns to behaviour: Examples from language and math. *NeuroImage*, 51(1):462–471, 2010.

[172] M.A. Ranzato, F.J. Huang, Y.L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. Ieee, 2007.

[173] J.D.M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[174] Annemie Ribbens, Frederik Maes, Dirk Vandermeulen, and Paul Sueten. Semisupervised probabilistic clustering of brain mr images including prior clinical information. In *MCV'10 Proceedings of the 2010 international MICCAI conference on Medical computer vision: recognition techniques and applications in medical imaging*, pages 184–194, 2010.

[175] F. Rodriguez. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. Technical report, DTIC Document, 2008.

[176] Alard Roebroeck, Elia Formisano, and Rainer Goebel. Mapping directed influence over the brain using granger causality and fmri. *Neuroimage*, 25(1):230–42, March 2005.

[177] Y.D. Rubinstein and T. Hastie. Discriminative vs informative learning. In *Proc. Third Int. Conf. on Knowledge Discovery and Data Mining*, pages 49–53, 1997.

[178] S. Ryali, K. Supekar, D.A. Abrams, and V. Menon. Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, 51(2):752–764, 2010.

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

[180] Mert R Sabuncu, Serdar K Balci, Martha E Shenton, and Polina Golland. Image-driven population analysis through mixture modeling. *IEEE Trans Med Imaging*, 28(9):1473–1487, Sep 2009.

[181] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2008.

[182] E. Salmon, F. Collette, C. Degueldre, C. Lemaire, and G. Franck. Voxel-based analysis of confounding effects of age and dementia severity on cerebral metabolism in alzheimer's disease. *Hum Brain Mapp*, 10(1):39–48, May 2000.

[183] Mark Schmidt, Ewout van den Berg, Michael P. Friedlander, and Kevin Murphy. Optimizing costly functions with simple constraints:a limited-memory projected quasi-newton algorithm. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR: Workshop and Conference Proceedings series*, pages 456–463, 2009.

[184] Shai Shalev-Shwartz, Yoram Singer, P. Bennett, and Emilio Parrado-hernández. Efficient learning of label ranking by soft projections onto polyhedra. In *Journal of Machine Learning Research*, volume 7, 2006.

[185] A. Shashua, R. Zass, and T. Hazan. Multi-way clustering using super-symmetric nonnegative tensor factorization. *Computer Vision–ECCV 2006*, pages 595–608, 2006.

[186] Dinggang Shen and Christos Davatzikos. Hammer: hierarchical attribute matching mechanism for elastic registration. *IEEE Trans Med Imaging*, 21(11):1421–1439, Nov 2002.

[187] Dinggang Shen and Christos Davatzikos. Very high-resolution morphometry using mass-preserving deformations and hammer elastic registration. *Neuroimage*, 18(1):28–41, Jan 2003.

[188] Vikas Sindhwani and Partha Niyogi. Beyond the point cloud: from transductive to semi-supervised learning. In *In ICML*, pages 824–831, 2005.

[189] Sindhwani, Vikas and Niyogi, Partha. Linear manifold regularization for large scale semi-supervised learning. In *Proc. of the 22nd ICML Workshop on Learning with Partially Classified Training Data*, 2005.

[190] Ajit Singh and Geoffrey Gordon. A unified view of matrix factorization models. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5212 of *Lecture Notes in Computer Science*, chapter 24, pages 358–373. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[191] Karl Sjöstrand, Egill Rostrup, Charlotte Ryberg, Rasmus Larsen, Colin Studholme, Hansjoerg Baezner, Jose Ferro, Franz Fazekas, Leonardo Pantoni, Domenico Inzitari, Gunhild Waldemar, and L. A. D. I. S. Study Group. Sparse decomposition and modeling of anatomical shape variation. *IEEE Trans Med Imaging*, 26(12):1625–1635, Dec 2007.

[192] Aristeidis Sotiras and Nikos Paragios. Deformable Image Registration: A Survey. Rapport de recherche RR-7919, INRIA, March 2012.

[193] N. Srebro, J.D.M. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17:1329–1336, 2005.

[194] J. Sui, T. Adali, G.D. Pearlson, and V.D. Calhoun. An ica-based method for the identification of optimal fmri features and components using combined group-discriminative techniques. *Neuroimage*, 46(1):73–86, 2009.

[195] Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D Greicius. Network analysis of intrinsic functional brain connectivity in alzheimer's disease. *PLoS Comput Biol*, 4(6):e1000100, June 2008.

[196] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[197] Stefan J Teipel, Christine Born, Michael Ewers, Arun L W Bokde, Maximilian F Reiser, Hans-Jürgen Möller, and Harald Hampel. Multivariate deformation-based analysis of brain atrophy to predict alzheimer's disease in mild cognitive impairment. *Neuroimage*, 38(1):13–24, Oct 2007.

[198] Carlos E. Thomaz, James P. Boardman, Derek L.G. Hill, Jo V. Hajnal, David D. Edwards, Mary A. Rutherford, Duncan F. Gillies, and Daniel Rueckert. Using a maximum uncertainty lda-based approach to classify and analyse mr brain images. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, 3216:291–300, 2004.

[199] M.E. Tipping. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

[200] Simon Tong and Daphne Koller. Restricted bayes optimal classifiers. In *In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 658–664, 2000.

[201] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, 2004.

[202] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[203] M.P. van den Heuvel and H.E. Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *Eur Neuropsychopharmacol*, 20:519–534, Aug 2010.

[204] V. Vapnik. Statistical learning theory. 1998, 1998.

[205] Ravi Varadhan and Christophe Roland. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

[206] G. Varoquaux, A. Gramfort, F. Pedregosa, V. Michel, and B. Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, pages 562–573. Springer, 2011.

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

[208] Archana Venkataraman, Yogesh Rathi, Marek Kubicki, Carl-Fredrik Westin, and Polina Golland. Joint modeling of anatomical and functional connectivity for population studies. *IEEE Trans Med Imaging*, 31(2):164–82, February 2012.

[209] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE transactions on pattern analysis and machine intelligence*, 13(6):583–598, 1991.

[210] U. Vovk, F. Pernus, and B. Likar. A robust technique for motion correction in fmri. *Medical Imaging, IEEE Transactions on*, 26(3):405–421, 2007.

[211] K. Wang, M. Liang, L. Wang, L. Tian, X. Zhang, K. Li, and T. Jiang. Altered functional connectivity in early alzheimer's disease: A resting-state fmri study. *Human brain mapping*, 28(10):967–978, 2007.

[212] P. Wang and R. Verma. On classifying disease-induced patterns in the brain using diffusion tensor images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, pages 908–916, 2008.

[213] Y. Wang and S. Ma. Projected barzilai–borwein method for large-scale nonnegative image restoration. *Inverse Problems in Science and Engineering*, 15(6):559–583, 2007.

[214] S. Weisberg. *Applied linear regression*, volume 528. Wiley, 2005.

[215] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation. *SIAM Journal on Scientific Computing*, 32:1832–1857, 2009.

[216] C.F. Westin, SE Maier, H. Mamata, A. Nabavi, FA Jolesz, and R. Kikinis. Processing and visualization for diffusion tensor mri. *Medical Image Analysis*, 6(2):93–108, 2002.

[217] The wiki-based collaborative Radiology resource Radiopaedia.org. T1 relaxation time, 2009. [Online; accessed 2-Jun-2012].

[218] The wiki-based collaborative Radiology resource Radiopaedia.org. T1 weighted image, 2009. [Online; accessed 2-Jun-2012].

[219] Wikipedia. Diffusion mri, 2011. [Online; accessed 2-Jun-2012].

[220] Wikipedia. Fluid attenuated inversion recovery, 2011. [Online; accessed 2-Jan-2012].

[221] Wikipedia. Positron emission tomography, 2011. [Online; accessed 2-Jan-2012].

[222] Wikipedia. Single-photon emission computed tomography, 2011. [Online; accessed 2-Jan-2012].

[223] Wikipedia. Functional magnetic resonance imaging, 2012. [Online; accessed 2-Jun-2012].

[224] D.P. Wipf and B.D. Rao. Sparse bayesian learning for basis selection. *Signal Processing, IEEE Transactions on*, 52(8):2153–2164, 2004.

[225] I. C. Wright, P. K. McGuire, J. B. Poline, J. M. Travere, R. M. Murray, C. D. Frith, R. S. Frackowiak, and K. J. Friston. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2(4):244–252, Dec 1995.

[226] Stephen J. Wright, Robert D. Nowak, and MÁrio A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, July 2009.

[227] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns. *NeuroImage*, 42(4):1414–1429, 2008.

[228] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. Ieee, 2009.

[229] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *Neural Networks, IEEE Transactions on*, 21(5):734–749, 2010.

[230] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. *Advances in neural information processing systems*, pages 689–695, 2001.

[231] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 211–218. ACM, 2009.

[232] Paul Yushkevich, Sarang Joshi, Stephen M Pizer, John G Csernansky, and Lei E Wang. Feature selection for shape-based classification of biological objects. *Inf Process Med Imaging*, 18:114–125, Jul 2003.

[233] Paul A Yushkevich, Hui Zhang, and James C Gee. Continuous medial representation for anatomical structures. *IEEE Trans Med Imaging*, 25(12):1547–1564, Dec 2006.

[234] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *NeuroImage*, 2011.

[235] Yi Zhang, Jie Tian, Kai Yuan, Peng Liu, Lu Zhuo, Wei Qin, Liyan Zhao, Jixin Liu, Karen M von Deneen, Nelson J Klahr, Mark S Gold, and Yijun Liu. Distinct resting-state brain activities in heroin-dependent individuals. *Brain Res*, 1402:46–53, July 2011.

[236] Dajiang Zhu, Kaiming Li, Carlos Cesar Faraco, Fan Deng, Degang Zhang, Lei Guo, L Stephen Miller, and Tianming Liu. Optimization of functional brain rois via maximization of consistency of structural connectivity profiles. *Neuroimage*, 59(2):1382–93, 2012.

[237] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.

[238] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.