Statistical Learning for Sample-Limited High-dimensional Problems with Application to Biomedical Data

by

Tzu-Yu Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2013

Doctoral Committee:

Professor Alfred O. Hero, Chair
Associate Professor Clayton D. Scott, Co-Chair
Professor Jeffrey A. Fessler
Professor Ji Zhu

© Tzu-Yu Liu 2013
All Rights Reserved

This dissertation is dedicated to my loving parents

in honor of their 30th wedding anniversary.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and encouragement of many individuals.

First and foremost, I would like to thank my advisor, Professor Alfred Hero, for his guidance and mentorship. His depth and breadth of knowledge guided me to the world of research. There have been numerous inspiring moments in the meetings when he gave me insights into research projects. These encouraging discussions led to various chapters in this dissertation. An abundance of applications for the developed methodologies would not have been possible without his support. Besides the vast knowledge and support, I have been impressed by his passion for research. I will never forget seeing an elder professor wearing a backpack, meeting young researchers around the world and always willing to provide his insights and encouragement to intellects. Research can take place anytime, anywhere for a devoted intellect.

I would like to extend my thanks to my co-advisor, Professor Clayton Scott, for his expertise in machine learning and Statistics, and to my other committee members, Professor Jeffrey Fessler, Professor Ji Zhu, for their valuable input to this dissertation.

I was fortunate to work with many collaborators. By order of appearance, I am grateful to several postdoctoral research fellows in the lab, Dr. Ami Wiesel, Dr. Koby Todros and Dr. Dennis Wei for their guidance, especially on the PLS project.

It was a memorable experience collaborating with many researchers in different disciplines. I am thankful to Dr. Frank Bogun for the opportunity of working on the Electrocardiology project; Dr. Jean-Christophe Olivo-Marin, Dr. Alexandre Dufour

and Christel Ducroz for the 3D cell microscopy project; Professor Tenenhaus Arthur and Dr. Laura Trinchera for collaborating on the PLS project; Dr. Christopher Woods, Dr. Aimee Zaas and Dr. Geoffrey Ginsburg for their Medicine knowledge.

Finally, I would like to express my deepest thanks to my parents. "I shall be telling this with a *smile*, somewhere ages and ages hence: two roads diverged in a wood, and I - I took the one less traveled by, and that has made all the difference." Thank you for walking with me in the woods no matter if they are dark and scary or lovely with birds singing. If one sees some beautiful scenery along the trails in this dissertation, that is because this dissertation is a proud combination of Mathematics and Biostatistics.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Statistical Learning for Sample-Limited High-dimensional Problems with Application to Biomedical Data

by

Tzu-Yu Liu

Chair: Alfred O. Hero

Co-Chair: Clayton D. Scott

With advancing technology comes the need to extract information from increasingly high-dimensional data, whereas the number of samples is often limited or even acquired from imbalanced populations. This thesis develops strategies for classification and prediction in high-dimensional but poorly sampled problems arising in computational biology and medicine. These strategies are presented in 6 chapters. In Chapter II Support Vector Machine (SVM) classifiers are applied to localizing ventricular tachycardia from electrocardiographical data. In Chapters III, IV, V and VII optimization-driven structured sparsity algorithms are developed. In Chapter VI a class of uneven margin SVMs is proposed for learning binary classifiers with imbalanced training populations.

The major part of this thesis is focused on group structured sparsity constrained statistical learning for sample-limited high-dimensional problems. Variable selection consists of reducing the dimension to a few important variables that contain most

of the information necessary for discriminating between classes or for prediction of continuous responses. This can potentially avoid overfitting problems, improve generalizability of the predictors and provide better interpretation. Novel algorithms based on the augmented Larangian and ADMM methods are developed for various statistical learning problems with group structured sparsity penalty: binary SVMs with application to 3D cell microscopy data to discover important shape information for characterizing highly deformable cells; multi-class SVMs with application to gene expression analysis to improve disease prediction rate and control irrelevant patient variations; PLS regression with application to chemometrics, medicine, and agriculture applications. These applications demonstrate the benefit of sparsity constrained optimization approaches to high dimensional problems with limited data.

CHAPTER I

Introduction

1.1 Motivation

With advancing technology comes the need to extract information from increasingly high dimensional data. Some examples of technologies that produce high dimensional data are microarrays in genomics (dimensions in the tens of thousands of gene expression probes), high-throughput video sequences (dimensions in the millions of pixel intensities per second), electrocardiology (dimensions in the thousands of samples per ECG electrode), and chemometrics (dimensions in the thousands of spectrum analyzer bands per compound), etc. These technologies produce data that have high intrinsic statistical variability, e.g., due to technical noise or biological variation in the sample. To account for statistical variability it is necessary to acquire multiple samples. This gives rise to the problem of information extraction from data of high dimension from relatively few samples. This thesis addresses several specific applications involving extracting information from data of high dimension with small sample size.

We motivate our domain of study by the electrocardiology problem of detecting and localizing ventricular tachycardia (VT), an arrhythmia originates in one of the ventricles of the heart, from ECG and electrograms (EGM) data. These data have the aforementioned properties of:

1. **High dimension**: the ECG and EGM data consist of temporal traces of heart activity, i.e., heart beat waveforms, that are digitized into a data matrix containing thousands of time samples.

2. **Small sample size**: The number of patients that are available for training the classifier is only on the order of tens.

3. **High variability**: Patients within a given heart disease category have ECG and EGM traces that exhibit a high degree of variability in the shape (QRS-complex) of their digitized waveforms.

These properties bring challenges to applying machine learning methods to disease classification, VT detection, and VT source localization. Training classifiers with high dimensional data and few samples can result in overfitted models, reducing the predictability of the classifiers. The high variability of the signals and the limited samples challenge the accuracy of information extraction, and make the interpretation of the data difficult. The accuracy of the model learned from clinical data suffers from small data size and imbalanced proportions among the different disease classes. This motivates the issues treated in this thesis: statistical learning in high dimensional and small sample size data. We address these issues by developing statistical learning methods with sparse feature selection. Regularized Support Vector Machines (SVM) in Chapter IV and Chapter V enable feature selection simultaneously with classification. The formulation of Sparse Partial Least Squares regression in Chapter VII provides a regression tool, combining dimensionality reduction, prediction, and variable selection. Uneven margin SVM of Chapter VI addresses the issue of learning from imbalanced populations.

1.2 Background and Contributions

1.2.1 Background

We motivate the thesis by the electrocardiology problem of detecting and localizing ventricular tachycardia (VT) from ECG and EGM data. VT is a potentially life-threatening arrhythmia because it may lead to ventricular fibrillation and sudden death. A common current therapy is to perform a catheter ablation procedure, which requires finding the location of the VT on the myocardium (cardiac mapping procedure), and eliminating the VT source by high frequency radio waves (catheter ablation procedure). The mapping procedure consists of two stages. First, the electrocardiologist induces VTs in the patient, and records the signals as templates. At the next stage, a catheter at the left ventricle is used to stimulate the wall and the signals are recorded, forming a record containing the so-called pace-maps. The objective is to look for pace-maps that match the morphology of the recorded templates of VT and eliminate the loci of these VT sources by ablation. The accuracy of VT detection and localization algorithms will depend on the spatial resolution and sensitivity properties of the sensing instrument. Two instruments are commonly used in electrocardiology, 12 lead electrocardiograms (ECG), and single lead electrograms (EGM) from implantable cardiac defibrillators (ICDs). EGM's ability to differentiate VTs and to target VT during mapping and ablation have not been thoroughly explored. Furthermore the spatial resolution of pace-mapping within the infarct zone (a localized area of scar tissue due to loss of adequate blood supply) in patients with prior infarction has not been adequately assessed. Developing algorithms that use Electrograms as a surrogate for ECG and automated classification or prediction of the origin of VT based on ECG can potentially result in a reduction of the time duration of the pace-mapping procedure, which usually takes more than 6 hours. In Chapter 2 we introduce machine learning algorithms for classification and spatial localization

designed for pace-mapping. In particular, we developed quantitative measures of the achievable spatial resolution of ECGs and EGMs, examined the potential of using EGMs when ECGs of the VT template are not available, and built classifiers based on ECGs to predict the locations of VT origination sites.

1.2.2 Learning from High-dimensional Features

One of the principal challenges of high dimensional data is that the number of samples is usually much smaller than the dimension. Let rp be the dimension of the feature variables collected, and let n be the number of samples. It is well known that the estimate of the covariance matrix of the high dimensional variables is problematic since when $n < rp$ the empirical estimate of the covariance matrix becomes singular. A similar problem occurs in classification and regression problems. In this thesis we take small sample size and reduce the dimension of the feature space using variable selection techniques. Variable selection consists of reducing the dimension to a few important variables that contain most of the information necessary for discriminating between classes or for prediction of unobserved events. This can potentially avoid overfitting problems, improve generalizability of the predictors, and provide better interpretation through a more parsimonious model [4].

Suppose we have a dataset with n samples,

$$\{\mathbf{x}_i, \mathbf{y}_i, s_i\}_{i=1}^{n}$$

in which $\mathbf{x}_i \in R^{rp}$ are the independent variables, $y_i \in \{1, 2, ..., K\}$ are the dependent variables for categorical responses or $\mathbf{y}_i \in R^q$ for continuous responses, and $s_i \in \{1, 2, ..., m\}$ represents the additional information about the generating sources. All r, p, q, m are positive integers. It is sometimes useful to write the variables in matrix forms, $X \in R^{n \times rp}$ and $Y \in R^{n \times q}$. The data format occurs naturally in

experiments conducted under several conditions. For example, in serially sampled experiments, there could be multiple measurements collected over time, contributing one p dimensional measurement at each time point, then $\mathbf{x}_i$ becomes a multi-block data with r blocks. The data may have been collected from m different individuals, labeled as s_i.

The thesis begins with supervised classification, binary classification with group structured feature selection when $r = 1$, and extends to multi-block multi-class data by adopting a general framework [3]. In multi-block multi-class classification, the task is to correctly predict the label by using serially or spatially diversified samples. As the data dimension increases, variable selection becomes increasingly important in these problems.

This is especially the case for the serially sampled reference-based classification problem, which can be viewed as a special structured case of the multi-block multi-class classification problem [3] [5], as variable dimensions increase linearly in the number of references. For example, a fixed sample for each subject under normal conditions, i.e., before the challenge tasks, can be viewed as a fixed reference. The references and the samples after the challenge tasks form a dataset with $r = 2$ blocks. By using such a fixed reference, irrelevant patient variations can be controlled and enhance our ability to evaluate positive or negative response to drug treatment, or classification of diseases based on gene microarray responses from multiple time points or multiple tissues. Hence it is important to understand which variables are strongly relevant to the classification task, and how they evolve over temporally or spatially different samples.

We treat variable selection for the general multiclass classification problem for which binary classification variable selection has been commonly implemented by using forward/backward selection and parameter estimation with shrinkage. For the high-dimensional multi-block multi-class problems of interest to us, in this thesis we

show that parameter estimation with shrinkage can be cast as a problem of structured variable selection, where the structure is specified by the classes and blocks defining the sampling patterns. The problem can be formulated mathematically as

$$\min_{F} V(F, X, Y) + \lambda R(F) \tag{1.1}$$

in which V is a loss function for classification, R is a sparsity regularization function that induces shrinkage, and λ is the regularization parameter. A convex optimization method is developed to solve for the optimal classifier function F and select the relevant variables simultaneously. This optimization is implemented by variable splitting and augmented Lagrangian methods.

Another dimension reduction technique that has obtained much attention in recent years is the class of constrained eigen-decomposition methods, such as principle component analysis (PCA) [6] and partial least squares regression (PLS) [7]. PLS can be viewed as an extension of PCA since it takes into account the response variables, and is a supervised learning method. PLS combines dimensionality reduction and prediction using a latent variable model. We discuss the PLS model and formulate the sparse PLS in Chapter VII.

Standard PLS performs a sequence of eigen-decompositions to specify the PLS component directions and the latent components of the PLS model. Suppose that the data with n samples consists of predictors $X \in R^{n\times p}$ and responses $Y \in R^{n\times q}$. The general underlying model is $X = TP' + E$ and $Y = TQ' + F$, where T is the latent component matrix, P and Q are the loading matrices, E and F are the residual terms.

As variable dimension increases, variable selection also becomes essential to avoid over-fitting and to provide more accurate PLS predictors. We propose a structured sparsity penalty in which global variable selection is performed such that any variable selected is shared among all PLS components. Analogous to the variable selection

in multi-block multi-class classification, we formulate PLS with structured sparsity as a variational optimization problem, with objective function V in Equation 1.1 equal to the classical PLS criterion with an added mixed norm sparsity constraint on the weight matrix. We propose a novel augmented Lagrangian method to solve the optimization problem. We show that soft thresholding for sparsity occurs naturally as part of the iterative solution.

We summarize the specific formulation of the loss function and the sparsity regularization functions of (1.1) in Table 1.1. Experiments show that both the general multi-block multi-class classification with variable selection and the modified PLS attains better performance with fewer predictor variables and fewer components as compared to previously proposed PLS methods [7] [8] [9].

1.2.3 Learning from Imbalanced Data

Most statistical learning methods are usually designed for data that are well-balanced. However, there are many of real world problems that do not have the same number of training samples in each class. This creates another common small sample size problem, in which the number of samples in a particular class is much smaller than any other. This problem is known as the imbalanced learning problem [10, 11, 12, 13, 14, 15, 16, 17, 18].

There are two main strategies that have been proposed to solve the imbalanced classification problem in the literature. One artificially balances the training data, also known as the external approaches; while the other, adopted here, tries to develop algorithms that can handle imbalanced data in an optimal manner, also known as internal approaches. We compare recent approaches that deal with imbalanced datasets in the context of weighted risks, which is a performance measure related to classification calibration [19]. Let $\eta(x) = P(Y = 1|X = x)$, and $\alpha \in (0, 1)$ be the uneven cost parameter for false positive and false negative. The conditional L-risk

Method	Formulation
Binary SVM Chapter IV	$X \in R^p,\ y_i \in \{-1, 1\}$ $F = \{f\},\ f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ $V(F, X, Y) = \frac{1}{n}\sum_{i=1}^{n} [1 - y_i f(\mathbf{x_i})]_+$ $R(F) = \sum_{g=1}^{G} \|\|\mathbf{w}_{I_g}\|\|_2$
Multi-class SVM Chapter V	$X \in R^{rp}\ ,\ y_i \in \{1, 2, ..., K\}$ $F = \{f_1, f_2, ..., f_K\},\ f_k(\mathbf{x}) = \mathbf{w}_k'\mathbf{x} + b_k$ $W = \begin{bmatrix} \| & \| & & \| \\ \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_K \\ \| & \| & & \| \end{bmatrix} = \begin{bmatrix} - & \mathbf{w}'_{(1)} & - \\ - & \mathbf{w}'_{(2)} & - \\ & \vdots & \\ - & \mathbf{w}'_{(p)} & - \end{bmatrix}$ $V(F, \mathbf{x_i}) = \frac{1}{n}\sum_{i=1}^{n} \left[\max_k (1 - \delta_{y_i,k} + f_k(\mathbf{x_i}) - f_{y_i}(\mathbf{x_i}))\right]_+$ $R(F) = \sum_{j=1}^{p} \left\|\tilde{\mathbf{w}}_{(j)}\right\|_2$ with $\tilde{\mathbf{w}}_{(j)} = \begin{bmatrix} \mathbf{w}_{(j)} \\ \mathbf{w}_{(j+p)} \\ \vdots \\ \mathbf{w}_{j+(r-1)p} \end{bmatrix}$
PLS Chapter VII	$X \in R^p,\ \mathbf{y}_i \in R^q$ $W = \begin{bmatrix} \| & \| & & \| \\ \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_K \\ \| & \| & & \| \end{bmatrix} = \begin{bmatrix} - & \mathbf{w}'_{(1)} & - \\ - & \mathbf{w}'_{(2)} & - \\ & \vdots & \\ - & \mathbf{w}'_{(p)} & - \end{bmatrix}$ $V(W, X, Y) = -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k$ $s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall\ k\ and\ \mathbf{w}_k'X'X\mathbf{w}_i = 0\ \forall\ i \neq k$ $R(W) = \sum_{j=1}^{p} \|\|\mathbf{w}_{(j)}\|\|_2$

Table 1.1: Definitions of V and R in the thesis.

is defined as $C_L(\eta, t) = \eta L_1(t) + (1 - \eta)L_{-1}(t)$. A loss function L is α-classification calibrated if, for all x such that $\eta(x) \neq \alpha$, the value of the prediction function $f(x)$ minimizing the conditional L-risk will have the same sign as the optimal predictor $\eta(x) - \alpha$. The experimental results support what we expect from the calibration

theory: calibrated losses outperform the non-calibrated losses.

1.3 Outline of the Thesis

The thesis is organized as follows. Chapter II presents research on ECGs and EGMs pace mapping analysis. In chapter III, we discuss a general framework for statistical learning with group structured variable section and introduce the general optimization algorithm for solving these problems. Chapter IV and V focus on binary and multi-class classifications with group structured variable selection respectively. We show that the performance becomes better with our proposed method in high dimensional data. Chapter VI addresses the problem of learning optimal classifiers with imbalanced training data, which is motivated by the imbalance applications in chapter II and V. In chapter VII, we present our approach to sparse partial least squares regression with mixed norm penalties. We again impose sparsity constraints as in chapter IV and V, but to a different statistical learning problem. Finally, conclusions and some possible extensions for future work are discussed in Chapter VIII.

1.4 Publications

The publications that have come out of research presented in this thesis are as follows.

Journals

[1] Kentaro Yoshida, Tzu-Yu Liu, Clayton Scott, Alfred O. Hero, Miki Yokokawa, Sanjaya Gupta, Eric Good, Fred Morady, and Frank Bogun. "The value of defibrillator electrograms for recognition of clinical ventricular tachycardias and for pace mapping of post-infarction ventricular tachycardia". *J Am Coll Cardiol*, 56(12), 2010.

[2] Miki Yokokawa, Tzu-Yu Liu, Kentaro Yoshida, Clayton Scott, Alfred O. Hero, Eric

Good, Fred Morady, and Frank Bogun. "Automated analysis of the 12-lead electrocardiogram to identify the exit site of postinfarction ventricular tachycardia". *Heart Rhythm*, 2011.

[3] Tzu-Yu Liu, Clayton Scott, Alfred O. Hero. "Uneven Margin SVM", in preparation.

[4] Alexandre Dufour, Tzu-Yu Liu, Christel Ducroz, Alfred O. Hero, and Jean-Christophe Olivo-Marin. "Spherical Harmonics Based Classification and Analysis of Highly Deforming Cells in 3D Microscopy", in preparation.

[5] Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei, and Alfred O. Hero. "Globally Sparse PLS Regression", in preparation.

[6] Tzu-Yu Liu, Ami Wiesel, Christopher W. Woods, Aimee Zaas, Geoffrey S. Ginsburg and Alfred O. Hero. " Learning Differential Gene Expression Signatures from Personalized High Throughput Screening", in preparation.

Conferences

[1] Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei, and Alfred O. Hero. "Globally Sparse PLS Regression". *7th International Conference on Partial Least Squares and Related Methods (PLS)*, 2012.

[2] Tzu-Yu Liu, Ami Wiesel, Christopher W. Woods, Aimee Zaas, Geoffrey S. Ginsburg and Alfred O. Hero. " Learning Differential Gene Expression Signatures from Personalized High Throughput Screening". *The Great Lakes Bioinformatics Conference*, 2012.

[3] Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei, and Alfred O. Hero. "A new criterion for sparse PLS regression". *Compstat*, 2012.

[4] Tzu-Yu Liu, Laura Trinchera, Arthur Tenenhaus, Dennis Wei, and Alfred O. Hero. "Global Criteria for Sparse Penalized Partial Least Squares". *World Statistics Congress*, 2013, in preparation.

CHAPTER II

SVM classifiers for Electrocardiograph Application

2.1 Introduction

Ventricular tachycardia (VT) is a tachycardia, or fast heart rhythm, that originates in one of the ventricles of the heart. This is a potentially life-threatening arrhythmia because it may lead to ventricular fibrillation and sudden death.

It is usually possible to terminate a VT episode with a direct current shock across the heart. The shock may be delivered to the outside of the chest using an external defibrillator, or internally to the heart by an implantable cardioverter-defibrillator (ICD) if one has previously been inserted. An ICD may also be set to attempt to overdrive the pace of the ventricle. Pacing the ventricle at a rate faster than the underlying tachycardia can sometimes be effective in terminating the rhythm. If this fails after a short trial, the ICD will usually stop pacing, charge up and deliver a defibrillation grade shock.

However, a direct current shock does not eliminate a VT from recurrence. And in [20], it is concluded that among patients with heart failure in whom an ICD is implanted for primary prevention, those who receive shocks for any arrhythmia have a substantially higher risk of death than similar patients who do not receive such shocks.

Catheter ablation has revolutionized the management of patients with tachyarrhyth-

mias [21]. Having evolved from arrhythmia surgery, catheter ablation was initially performed using high voltage direct current (DC); however, since the late 1980s, radiofrequency current has supplanted DC as the energy source of choice and has made catheter ablation a first-line therapy for many tachycardias [22].

Therefore, finding the location of the VT is paramount in radiofrequency ablation, which could depend on the spatial resolution of the 12 lead electrocardiograms (ECG), and electrograms (EGM) from implantable cardiac defibrillators (ICDs). EGM's ability to differentiate VTs and its use to target VT during mapping and ablation procedures have not been described. Furthermore the spatial resolution of pace-mapping within the infarct zone (which results in an macroscopic area of tissue due to loss of adequate blood supply) in patients with prior infarction has not been adequately assessed. We have shown in this study that ICD EGMs can be used to differentiate clinical VTs from other VTs in patients undergoing VT ablation procedures. The spatial resolution of pace-mapping using ICD EGMs is variable but can be used for identification of a VT exit site.

2.2 Background

2.2.1 Electrical Activity of the Heart

The electrical behavior of a single cardiac muscle cell can be investigated by inserting microelectrodes into the interior of a cell from various regions of the heart [23].The various phases of the cardiac action potential are associated with changes in the permeability of the cell membrane, mainly to Na, K, and Ca ions. These changes in permeability produce alterations in the rate of passage of these ions across the membrane. And any process that abruptly changes the resting membrane potential to a critical value (threshold) will result in a propagated action potential.

Electrocardiograms (ECG) are usually recorded from indirect leads, i.e., located

on the skin. Electric activity going through the heart, can be measured by external (skin) electrodes. The ECG registers these activities from these electrodes which have been attached on different places on the body. In total, twelve leads are calculated using ten electrodes. They consist of 6 chest leads (V1,V2,V3,V4,V5 and V6) and 6 extremity leads (I, II, III, aVL, aVR, and aVF).

2.2.2 Ventricular Tachycardia (VT)

The majority of VTs are caused by re-entry involving a region of ventricular scar[24]. Dense fibrotic scar creates areas of anatomic conduction block. Fibrosis between surviving myocyte bundles decreases cell to cell coupling, and distorts the path of propagation causing slow conduction, which promotes re-entry. These re-entry circuits (Fig 2.1) often contain a narrow isthmus of abnormal conduction. Depolarisation of the small mass of tissue in the isthmus is not detectable in the body surface ECG. The QRS complex is caused by propagation of the wavefront from the exit of the circuit to the surrounding myocardium. After leaving the exit of the isthmus, the circulating re-entry wavefront may propagate through a broad path along the border of the scar (loop), back to the entrance of the isthmus.

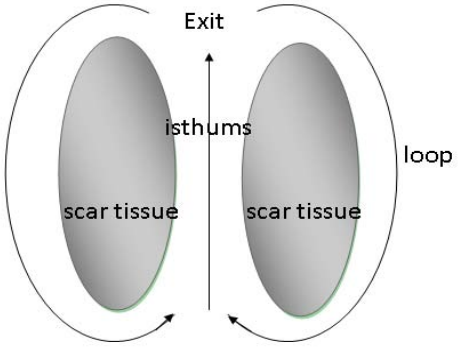

Figure 2.1: the re-entry circuit

A variety of different circuit configurations are possible. Ablation lesions produced with standard RF ablation catheters are usually less than 8 mm in diameter, relatively

small in relation to the entire re-entry circuit, and can be smaller than the width of the re-entry path at different points in the circuit. Successful ablation of a large circuit is achieved either by targeting an isthmus where the circuit can be interrupted with one or a small number of RF lesions, or by creating a line of RF lesions through a region containing the re-entry circuit.

The situation is further complicated by the frequent presence of multiple potential re-entry circuits, giving rise to multiple different VTs in a single patient. Ablation in one area may abolish more than one VT, or leave VT circuits in other locations intact. VTs that have been documented to occur spontaneously are referred to as "clinical VTs". Those that are induced in the electrophysiology lab, but not previously observed, are referred to as "non-clinical VTs".

2.2.3 Pace-map Procedure and Radiofrequency Ablation

An electrophysiology (EP) study of the heart is a nonsurgical analysis of the electrical conduction system (normal or abnormal) of the heart. The test employs cardiac catheters and sophisticated computers to generate electrocardiogram (ECG) tracings and electrical measurements with exquisite precision from within the heart chambers. The EP study can be performed solely for diagnostic purposes. It also is performed to find the exact location of electrical signals (cardiac mapping) with a therapeutic procedure called catheter ablation.

Catheter/Radiofrequency ablation procedure involves the use of a specially designed catheter that is threaded through the leg into the heart. While in the heart, the catheter is used to locate the arrhythmia source, which is then eliminated by high frequency radio waves, i.e., targeting at an isthmus where the re-entry circuit can be interrupted.

This study was approved by the Institutional Review Board at the University of Michigan. After informed consent was obtained, a multipolar electrode catheter

was inserted into a femoral vein and was positioned in the right ventricular apex. A 7 French multipolar catheter was placed at the His bundle position. Programmed ventricular stimulation was performed from 2 right ventricular sites using up to 4 extrastimuli. After ablations, the same stimulation protocol was repeated from 2 ventricular sites. In all patients, either a single or multiple VTs were recorded on either a 12- lead ECG or on a 7-lead telemetry recording before the ablation procedure. The spontaneous VTs were defined as the clinical VTs. The morphology of the clinical VTs was compared to the morphology of the induced VTs.

An electroanatomic mapping system (CARTO, Biosense Webster, Inc, Diamond Bar, California) was used in all patients, with an 8-Fr mapping/ablation catheter that had a 3.5-mm irrigated-tip electrode and a 2-mm ring electrode separated by 1 mm (Thermocool, Biosense Webster, Diamond Bar, CA). Intracardiac electrograms were filtered at 50-500 Hz. The intracardiac electrograms and leads V1, I, II and III were displayed on an oscilloscope and recorded at a speed of 100 mm/sec. The recordings were stored on optical disc (EP Med, Inc). Systemic heparinization was maintained throughout the procedure.

Left ventricular access was obtained using a retrograde aortic approach. A left ventricular endocardial voltage map was constructed during sinus rhythm. Pace-mapping was performed at sites with a voltage < 1.5 mV. Low voltage was defined as a bipolar voltage of < 1.5 mV. Dense scar was defined as < 0.5 mV. The border zone was defined as 0.5-1.5 mV. Bipolar pace-mapping was performed with an amplitude of 10 mA at a pulse width of 2 ms. If no capture occurred, the pacing output was increased progressively up to 20 mA.

The power of radiofrequency energy was titrated to achieve an impedance drop of 10 ohms. The maximal temperature was 45 degree celsius. Radiofrequency energy was delivered at isthmus sites or at VT exit sites.

2.2.4 Implantable Cardioverter Defibrillator (ICD)

ICDs are small devices, about the size of a pager, that are placed below the collarbone. Via wires, or leads, these devices continuously monitor the heart's rhythm. If the heart beats too quickly, the ventricles will not have enough time to fill with blood and will not effectively pump blood to the rest of the body. Left unchecked, the rapid heartbeat could cause death. To intervene, the ICD issues a lifesaving jolt of electricity to restore the heart's normal rhythm and prevent sudden cardiac death. ICDs also can act as pacemakers when a heart beat that is too slow (bradycardia) is detected.

Most ICDs keep a record of the heart's activity when an abnormal heart rhythm occurs. With this information, the electrophysiologist, a specialist in arrhythmias, can study the heart's activity and ask about other symptoms that may have occurred. Sometimes the ICD can be programmed to pace the heart to restore its natural rhythm and avoid the need for a shock from the ICD. Pacing signals from the ICD are not felt by the patient; shock signals are, and have been described as a kick in the chest.

2.2.5 Recorded Signals

In the pace-mapping procedure, a patient could have several VTs induced. We stored the 12 lead ECG and the EGM from the ICD for each VT, named as the templates. Depending on the manufacturers, most ICD stored two waveforms, named the far-field and the near-field, others only have one ICD signal. Therefore each set of template, corresponds to one VT, and has 13-14 waveforms, as shown in figure 2.2.

Then the catheter stimulated the wall of the left ventricle. The ECG and EGM generated, named as the pace-maps (test signals), are compared with those of the template. The figure above also shows a pace-map example.

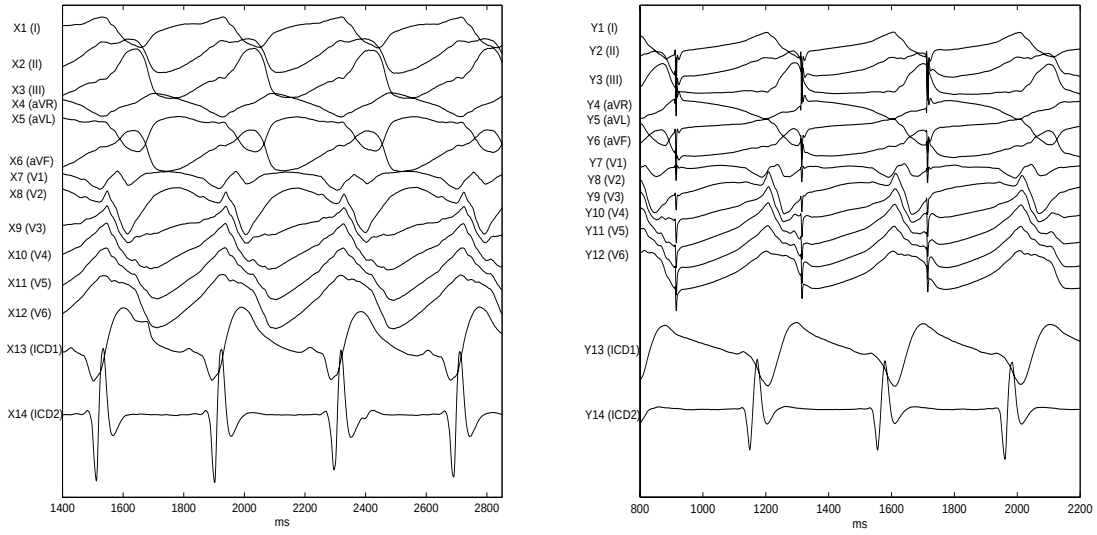

Figure 2.2: ECGs and EGMs for a VT(left) and a pace-map(right)

2.3 Value of Defibrillator Electrograms: Spatial Resolution and Differentiation of Ventricular Tachycardias

If the catheter is at the site of origin of the VT, then theoretically the ECG and EGM should have the same morphology as those of the targeted VT. This explains how pace-mapping works for identifying the site of origin. In clinical practice, pace-mapping is analyzed by visual inspection. A standard way of judging is that a good pace-map should have more than 10 leads among the 12 leads ECG that match with those of the template, then radiofrequency ablation is applied to this pacing site.

We would like to know how close do the 12 leads ECG bring us to the origin of the targeted VT when we observe a match between the targeted VT and the pace-map, defined as the spatial resolution of ECG. Similarly, we would like to know the spatial resolution of EGM. Though the 12 lead ECG is the standard judgment in the pace-mapping procedure, understanding the value of EGM is important. For one thing, the signals recorded in the ICDs might be the only prior information regarding a new patient. For another, this is the first step to evaluate the potential of using

ECG to differentiate VTs and may improve the algorithm in the ICD to reduce the false discharges/shocks.

However, the assessment of the spatial resolution is difficult unless a reproducible, quantitative analysis is performed. The purpose of this study is two-fold: (1) Provide a quantitative measure of the spatial resolution of pace-mapping using the 12 lead ECG and the EGM from the ICD. Compare these two measurements. (2) Provide a quantitative measure of the differentiability among different VTs using the EGMs.

2.3.1 Signal Alignment

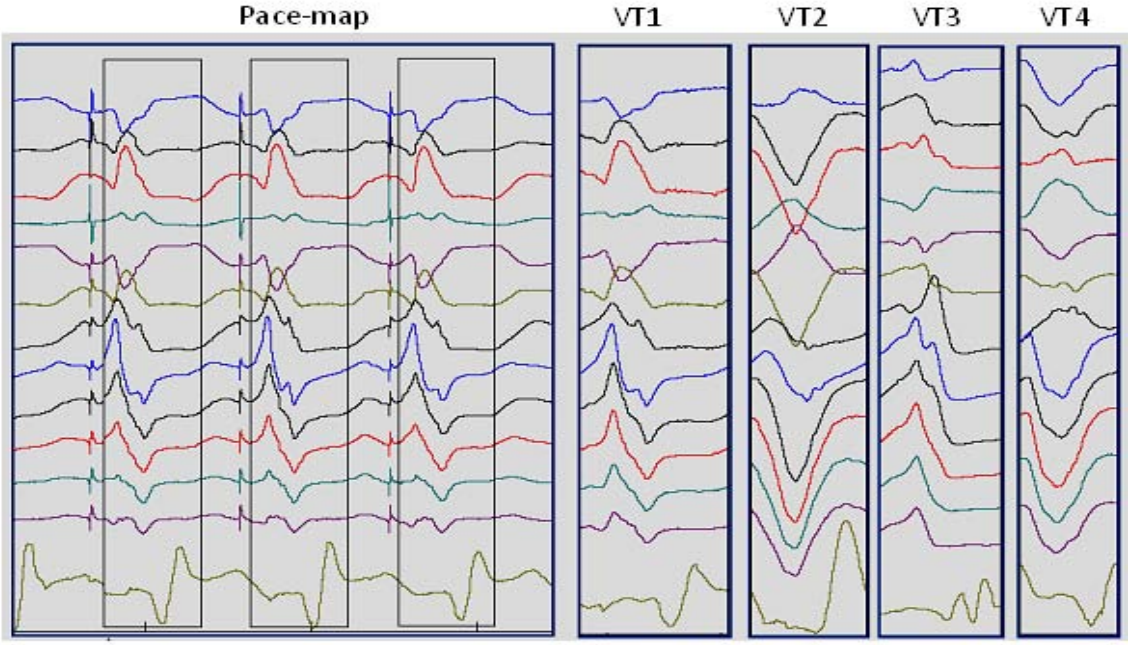

Figure 2.3: signals have different vector length

In Fig 2.3, we have a set of data from a patient, who is diagnosed to have 4 VTs, called VT1, VT2, VT3, VT4. The pace-map chosen is one of the 50 pace-map (testing) sites. Vertically, we see 13 waveforms, in which the first 12 of them are the 12 lead ECG, and the last one is the EGM from the ICD. Horizontally, the number of samples in one period is roughly around 500, with sampling frequency 2 kHz.

In electrophysiology's point of view, the pace-map chosen corresponds to VT1. So we would like to develop a testing method which gives positive results when testing the pace-map with VT1, and negative results with other VT's. However, the vector lengths of the pace-map and VTs are different. To make the vectors comparable, it requires the alignment of signals. Let $l1$, $l2$ be the lengths of the two signals $v1$, $v2$, being compared, in which $l1 > l2$. We set a window length of $l2$ on $v1$, and move the window to get the truncated $v1'$ such that the correlation coefficient between $v1'$ and $v2$ is maximized, or the root-mean-square difference between them is minimized. Fig. 2.4 shows the alignment results for a matching pace-map site and a non-matching pace-map site is shown in Fig. 2.5.

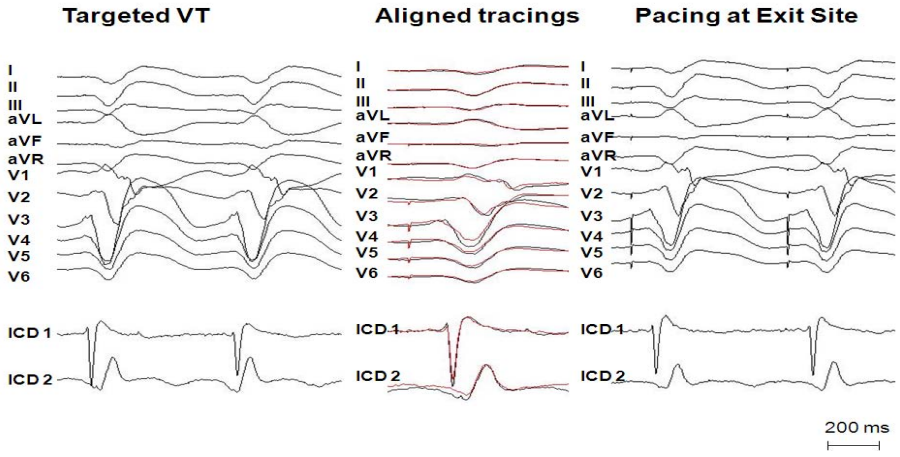

Figure 2.4: the alignment between targeted VT and a matching pace-map

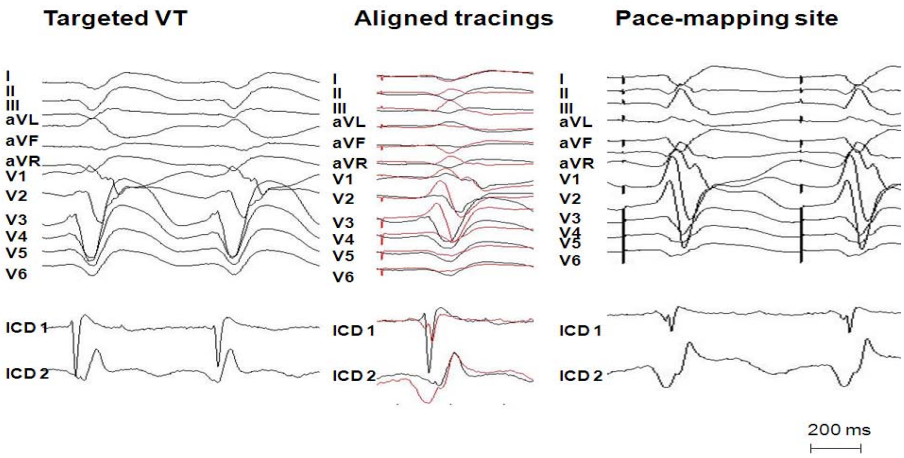

Figure 2.5: the alignment between targeted VT and a non-matching pace-map

2.3.2 Data Analysis

The most widely used technique in pace-mapping procedure is the correlation coefficient and root-mean-square statistics. They are less powerful than parametric methods if the assumptions underlying the latter are met, but are less likely to give distorted results when the assumptions fail.

$$corrcoef(\mathbf{X_i}, \mathbf{Y_i}) = \frac{\left(\mathbf{X_i} - \overline{\mathbf{X_i}}\right)' \left(\mathbf{Y_i} - \overline{\mathbf{Y_i}}\right)}{\sqrt{\left(\mathbf{X_i} - \overline{\mathbf{X_i}}\right)' \left(\mathbf{X_i} - \overline{\mathbf{X_i}}\right)} \sqrt{\left(\mathbf{Y_i} - \overline{\mathbf{Y_i}}\right)' \left(\mathbf{Y_i} - \overline{\mathbf{Y_i}}\right)}}$$

$$RMSd(\mathbf{X_i}, \mathbf{Y_i}) = \sqrt{\frac{1}{l} \left(\mathbf{X_i} - \mathbf{Y_i}\right)' \left(\mathbf{X_i} - \mathbf{Y_i}\right)}$$

The idea came from [25]. Since the matching between VT and pace-map requires the same morphology and the same magnitude, we try to explain the former by the correlation coefficient, and the latter with root-mean-square difference. The analysis was carried out for each of the 12 leads, $\mathbf{X_i}$, and $\mathbf{Y_i}$, where $i = 1, 2, ..., 12$. The correlation coefficients, the correlation coefficients, and RMSd, were averaged over the 12 leads, yielding 2 values that represented the degree of similarity between the test and template siganls based on the 12 lead ECG information. The same analysis was carried out for the ICD signals to represent the similarity based on EGM.

Another statistic analysis tool applied in this project is the receiver operating characteristic (ROC), a graphical plot of the sensitivity (β) versus 1-specificity (α) for a binary classifier system as its discrimination threshold is varied. Define a test function

$$\Phi(x) = \begin{cases} 1, & \text{say } H_1 \\ 0, & \text{say } H_0 \end{cases}$$

The false alarm probability and detection probability are functions of θ.

$$E_\theta[\Phi] = \int_\chi \Phi(x) f(x;\theta) dx = \begin{cases} P_F(\theta), & \theta \in \Theta_0 \\ P_D(\theta), & \theta \in \Theta_1 \end{cases}$$

A test function is said to be of level $\alpha \in [0,1]$ if $\max_{\theta \in \Theta_0} P_F(\theta) \leq \alpha$ and the power function is defined as $\beta(\theta) = P_D(\theta),\ \theta \in \Theta_1$. For the test of simple hypotheses $\theta \in \{\theta_0, \theta_1\}$,

$$H_0 : X \sim f(x;\theta_0)$$
$$H_1 : X \sim f(x;\theta_1)$$

the Neyman-Pearson Strategy is to find the most powerful test Φ^* of level α: $E_{\theta 1}[\Phi^*] \geq E_{\theta 1}[\Phi]$ for any other test satisfying $E_{\theta 0}[\Phi] \leq \alpha$. By Neyman Pearson Lemma, the MP test is a randomized likelihood ratio test of the following form

$$\Phi^*(x) = \begin{cases} 1, & f(x;\theta_1) > \eta f(x;\theta_0) \\ q, & f(x;\theta_1) = \eta f(x;\theta_0) \\ 0, & f(x;\theta_1) < \eta f(x;\theta_0) \end{cases}$$

where η and θ are chosen to satisfy $E_{\theta 0}[\Phi^*] = \alpha$. The threshold test have P_F and P_D indexed by a parameter η, then the receiver operating characteristic is simply the plot of the parametric curve $\{P_F(\eta, q), P_D(\eta, q)\}_{\eta,q}$, a plot of $\beta = P_D$ versus $\alpha = P_F$.

2.3.3 Spatial Resolution

Two independent observers compared the 12-lead ECG pace-maps and determined whether at least 10/12 leads matched with the targeted VT. Discrepancies were resolved by consensus. Labeling a pace-map with match of $\geq 10/12$ leads as positive and negative otherwise, this generated a binary classification. Consider a true positive (the outcome from a prediction is positive and the actual value is also positive), by visual inspection, the pace-maps and the targeted VT should have the same morphology; by quantitative analysis, the corrcoef between them is expected to be high, and the RMSd is expected to be low. Given the corrcoef and RMSd statistics, we varied the discrimination thresholds, and determined a receiver operating characteristic for each kind of statistic per patient per VT. The threshold that maximized the sum of sensitivity(β) and specificity($1-\alpha$) was chosen as the cut-off value for classifying the pace-maps.

To determine the spatial resolution of the pace-maps, the distance of each pace-mapping point from the exit site was measured on the electroanatomic map and correlated with the cut-off values of the compared signals (template signal vs test signal). An exit site was defined as a site where the pace-map matched the targeted VT and where the stimulus-QRS interval was less than 30% of the VT cycle length when pacing was performed during sinus rhythm at the VT cycle length. The area encompassing pacing sites with a corrcoef beyond the cut-off corrcoef value represented the spatial resolution for mapping an exit site (Fig 2.6). The same procedure is performed on the electroanatomic map for the RMSd statistic.

A total of 124 VTs were induced by programmed stimulation. In 13 of 15 patients in whom the clinical VTs were documented on 12-lead ECGs, VTs with matching configurations were inducible by programmed stimulation. In 2 of 15 patients, the documented VTs could not be induced; however, there were frequent premature ventricular contractions during the procedure that matched the clinical VT on the basis

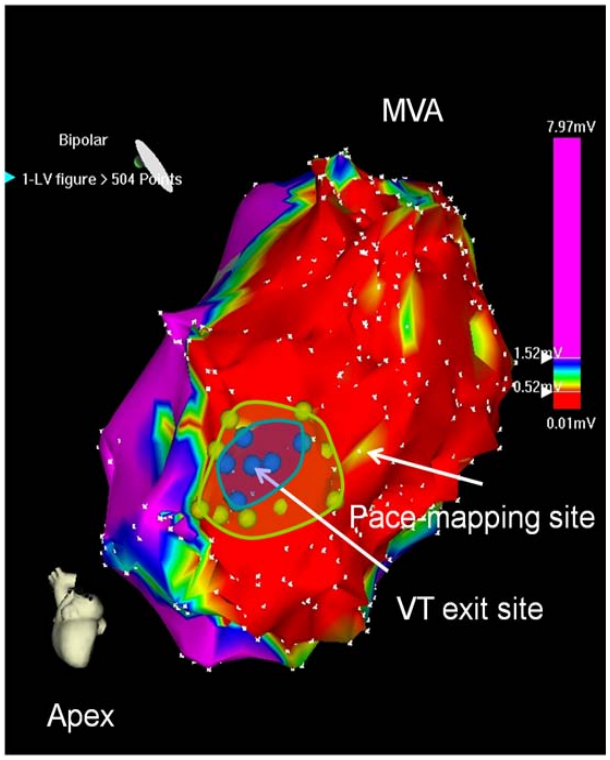

Figure 2.6: voltage map

of 12-lead ECGs. Pace-mapping was performed at 1,296 sites (a mean of 62 sites per patient, resulting in a sampling density of 0.75 $points/cm^2$ within low-voltage tissue) within low-voltage areas, and 62 distinct exit sites were identified. Matching pace maps were identified for an additional 15 VTs, but these were not considered exit sites because they had long stimulus-QRS intervals when pacing was performed during sinus rhythm. VT ablation sites were identified by entrainment mapping in 9 VTs and by pace mapping for 68 VTs. The mean procedure time was 383 ± 97 min. The total amount of radiofrequency energy delivered was 72 ± 41 min. After ablation, 10 VTs (8%) remained inducible. None of the clinical VTs remained inducible. Fourteen patients (67%) had no inducible VTs after ablation. Outcomes were no different if the exit site was identified on the basis of 10, 11, or 12 matching leads in the pace maps. VT exit sites and scar. The VT exit sites were located within low-voltage areas and had a mean bipolar voltage of 0.44 ± 0.43 mV. They were located a mean

of 16 ± 10 mm (range 4-46 mm) from sites with normal voltage (> 1.5 mV). The total mean low-voltage area was 82 ± 35 cm^2 (bipolar voltage < 1.5 mV) and 68 ± 31 cm^2 (bipolar voltage < 1.0 mV).

By the 12-lead ECG, the spatial resolution of pace mapping for identifying the exit site of a VT was 3.8 ± 4.5 cm^2 (range 0 to 17.5 cm^2) for the correlation coefficient and 3.6 ± 4.5 cm^2 (range 0 to 18.3 cm^2) for the RMSd. By the farfield ICD EGMs, the spatial resolution of pace mapping for identifying the exit site of a VT was 10.1 ± 10.0 cm^2 (range 0 to 36.5 cm^2) for the correlation coefficient and 10.2 ± 10.5 cm^2 (range 0 to 35 cm^2) for the RMSd.

2.3.4 Differentiation of the Clinical VT

When ECG documentation of a clinical VT is not available in a patient with an ICD, the ICD electrograms can be used to discriminate the clinical VT from other VTs that are induced in post-infarction patients. Furthermore ICD electrograms may be helpful for pace-mapping when a clinical VT is not inducible in determining the site of origin of targeted VTs during a mapping and ablation procedure. The latter may have relevance in patients in whom VTs are not inducible.

The template signals of the clinical VTs were compared to a total of 64 test signals of VTs that were induced during programmed stimulation. The ICD electrograms were almost as accurate as the 12-lead ECG's in differentiating the clinical VT from non-clinical VTs. All clinical VTs were accurately identified based on the 12-lead ECG from the clinical VT, and 98% of the ICD electrograms had a corrcoef that was below the cut-off value determined by the ROC curve of the clinical VT.

Identification of clinically-relevant VTs in post-infarction patients undergoing VT ablation is difficult unless a multi-lead ECG of the VT is available. Identification of the clinical VT is paramount as this might be the only VT requiring therapy. Ablation of only non-clinical VTs may result in recurrence of VT post-ablation. Fur-

thermore a particular VT may respond to anti-tachycardia pacing, and identification of a particular VT based on ICD EGMs might help to deliver a selected therapy that is effective for a particular VT, and thereby avoid unnecessary ICD discharges. This study demonstrates that the ICD electrograms are capable of identifying a particular VT as clinical or non-clinical.

2.3.5 Discussion

The ability of EGMs to differentiate VT and their use to target VT during mapping and ablation procedures have not been described. The spatial resolution of pace-mapping within the infarct zone in patients with prior infarction has not been adequately assessed. In 21 consecutive patients referred for catheter ablation of post-infarction VT, VTs were induced and ICD EGMs were recorded at the same time. The exit site of a particular arrhythmia was then identified by pace-mapping and the spatial resolution and accuracy of pace-mapping was determined for the 12 lead electrocardiogram (ECG) and the ICD EGMs. This was accomplished by comparing template signals to test signals using a customized Matlab program. Cut-off values were established using ROC curves to separate matching from non-matching pace-maps. The 12 lead ECG morphology of the clinical VTs was compared to 62 distinct VTs that were inducible to assess the discriminatory value of ICD EGMs to differentiate the clinical VT from other induced VTs. We found that ICD EGMs can be used to differentiate clinical VTs from other VTs in patients undergoing VT ablation procedures. The spatial resolution of pace-mapping using ICD EGMs is variable but can be used to confirm the presence of an exit site.

2.4 Automated Analysis of the 12-lead Electrocardiogram to Identify the Exit Site of Postinfarction Ventricular Tachycardia

The value of the 12-lead electrocardiogram (ECG) to identify the exit site of postinfarction ventricular tachycardia (VT) has been questioned. The purpose of this study was to assess the accuracy of a computerized algorithm for identifying a VT exit site on the basis of the 12-lead ECG. In 34 postinfarction patients, pace mapping was performed from within scar tissue. A computerized algorithm that used a supervised learning method (support vector machine) received the digitized pace-map morphologies combined with the pacing sites as training data. No other information (i.e., infarct localization, bundle branch block morphology, axis, or R-wave pattern) was used in the algorithm. The training data were validated in 58 VTs in 33 patients. Only the pace-map and/or VT morphologies were used in this algorithm. The sizes of 10 different anatomic sections within the heart were determined by using the pace maps as the determining factor. Automated identification of a VT exit site based on the 12-lead ECG of postinfarction VT is possible with an accuracy of about 70% for identifying a region of interest with a size of approximately 15 cm^2. Identification of an area of interest up-front will help to facilitate mapping and ablation of complex postinfarction VTs, especially in patients with large scars.

2.4.1 Classification of the 12-lead ECG

Digitized 12-lead ECGs of pace-maps generated within low-voltage tissue in the 34 patients were analyzed. The pacing sites were assigned a particular anatomic location within the heart based on a previously described schema (AJ, Fig. 2.7)[26]. The schema was adapted to this study by using the following guidelines: The distance between the apex and the base was divided into 3 equal segments (basal, mid, and

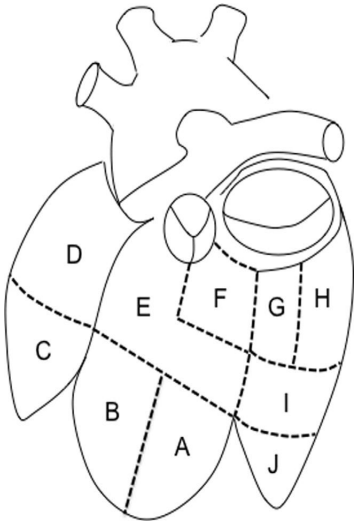

Figure 2.7: Cardiac sections of AJ serving as the regions to which the VT exit sites were assigned

distal). Areas A, B, C, and J were the distal segments; areas I, E, and D were the mid segments; and areas E, F, G, and H were the basal segments. Localization was performed by 2 independent observers. Discrepancies were resolved by consensus. We wanted to assess whether a computerized algorithm is able to distinguish the pacing site based on QRS morphology of the 12-lead ECG. To achieve this, we used a supervised learning method, the Support Vector Machine (SVM).

SVM is a machine learning technique. In binary classification problems, the principle of SVM is to find a hyperplane to separate the signals from the 2 regions, such that the separation between the 2 classes is maximized. In this multi-class problem, we followed a one-against-one approach to break the multi-class classification into several binary problems. We trained the algorithm for each pair of 2 different regions where the 12-lead ECG signals originate from and analyzed how often a particular region was chosen by the algorithm to determine which region (AJ) the ECG signal would be assigned to. Based on the majority vote policy, one can also rank all the possibilities, and use the rank (given the sorted prediction list, from the most probable

	A	B	C	D	E	F	G	H	I	J
A	0.59	0.09	0.02	0	0.08	0	0	0	0.08	0.15
B	0.21	0.50	0.16	0	0.07	0	0.02	0	0	0.02
C	0.04	0.05	0.65	0.12	0.02	0	0	0.01	0	0.09
D	0	0	0.07	0.77	0.11	0	0	0.03	0.01	0.01
E	0.05	0.01	0.01	0.10	0.73	0.05	0.02	0.01	0.03	0
F	0	0	0	0	0.17	0.67	0.09	0.04	0.03	0
G	0	0	0	0	0.02	0.13	0.62	0.16	0.07	0.01
H	0	0	0	0.06	0	0.01	0.06	0.78	0.06	0.07
I	0.05	0	0	0.02	0.03	0.01	0.03	0.10	0.66	0.08
J	0.08	0.01	0.07	0.02	0.01	0	0	0	0.05	0.77

Figure 2.8: Confusion matrix comparing accuracy of each region.

to the least, the position of the correct prediction) as the performance measure.

The algorithm to assign a particular pace-map to a particular region (AJ) was tested with leave one out cross-validation. The results were displayed in a confusion matrix indicating the percentage of correctly classified data (red rings indicate percentages of the correctly identified data in Fig. 2.8). The training data containing pace maps only were then validated by using 58 VTs from 33 patients where exit sites were identified by pace-mapping. The 12-lead ECGs of 58 VTs from 33 postinfarction patients in whom the exit sites were determined by pace mapping were analyzed prospectively. An exit site was defined as a site where the pace map matched the targeted VT and where the stimulusQRS interval was $\leq 30\%$ of the VT cycle length when pacing was performed during sinus rhythm. The VTs of these 33 patients served as testing data, and the pace-maps from the initial 34 patients served as the training data. Data were analyzed for the validation of accuracy of the computerized algorithm. The accuracy was 71% for assigning the testing data into the correct region (AJ). The overall accuracy increased to 88% for identification of a matching region if the 2 top-ranked regions were included. The overall rank to correctly identify a

particular region was an average of 1.7.

2.4.2 Determination of Spatial Resolution of the 12-lead ECG Pattern Based on Anatomic Region

The size of the anatomic area (AJ) that generated a particular ECG morphology during pace mapping within low-voltage tissue was determined. A median of the 12-lead ECG of the pace-maps of a particular region is shown in Fig. 2.9. The median 12-lead ECG morphology was then used as a template signal that was compared with the pace maps assigned to this and other regions and a correlation coefficient was generated. The spatial resolution of such a region was determined on the basis of receiver operator characteristics curves that generated a cutoff value separating the median ECG electrogram of a region (AJ) from pace maps of other regions. This was done for each patient in the low-voltage area where pacing was performed. The gold standard was whether or not a pace map belonged to a particular region or not. Once the cutoff value was determined for each region, the area encompassing sites with a correlation coefficient equal or greater than the cutoff value was measured on the electroanatomic map. The spatial resolution then was averaged for all patients and the areas were reported per region in Table 2.1.

	A	B	C	D	E
Anatomic area (cm^2)	13 (12 – 15)	13 (10 – 17)	22 (20 – 29)	62 (57 – 69)	26 (23 – 32)
Spatial resolution (cm^2)	15 (11 – 20)	20 (15 – 21)	18 (9 – 25)	20 (14 – 27)	16 (9 – 26)
	F	G	H	I	J
Anatomic area (cm^2)	19 (16 – 23)	21 (17 – 23)	22 (19 – 30)	30 (27 – 32)	21 (19 – 25)
Spatial resolution (cm^2)	14 (7 – 23)	10 (5 – 21)	11 (6 – 17)	10 (8 – 21)	18 (12 – 28)

Table 2.1: Spatial resolution of each region. Anatomic area and spatial resolution data are displayed as median values and (interquartile range).

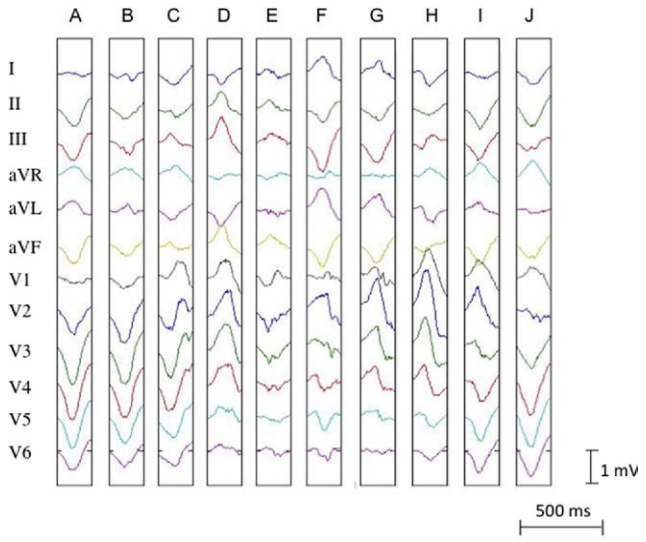

Figure 2.9: Median of ECGs from regions A to J.

2.4.3 Discussion

The 12-lead ECG of postinfarction VT contains localizing information that enables determination of a region of interest in the $10-20\ cm^2$ range for more than 70% of VT exit sites in a given sector. Identification of an area of interest up-front will help to facilitate mapping and ablation of complex postinfarction VTs, especially in patients with large scars. However, the accuracy depends on regions. A previously-used algorithm had an accuracy of 70% for predominantly apical septum regions (A, B)[26]. These were the regions that performed worst on the computerized analysis, suggesting that these algorithms are complementary and that the automated algorithm can be further improved. In the previously published algorithm, 50% of the VTs did not fit any particular pattern and therefore could not be classified, making the algorithm impractical. The main limitation of the previously described algorithm is the lack of applicability for patients with prior anterior wall infarctions and for right bundle branch block VT morphologies. In contrast, the computerized algorithm performed best in regions affected by anterior wall infarcts (region D). The discriminatory value of the computerized algorithm was imperfect in the apical septum area

where the accuracy was around 50%. However, if 2 zones are combined, the accuracy for determining the larger sector improved to approximately 70% in these regions.

Since there are no clear-cut demarcations between left ventricular regions and because infarct scars are not necessarily confined to one region, it seems appropriate to use a ranking classification indicating the best and second-best matching regions for test data. In order to identify a VT exit region (AJ) based on the 12-lead ECG, an average of 1.7 attempts were needed to get to the correct region. Since the mean size of the ECG-determined region is approximately 15 cm^2, combining 2 regions results in an area of 30 cm^2 in which more than 80% of VT exit sites could be assigned to. The mean scar area in the patients in whom the mapping data were obtained is a mean of 85 cm^2; the 12-lead ECG helps to narrow down the area of interest to approximately one-third for 80% of VTs.

2.5 Conclusion

We have shown that an SVM is capable of classifying the 12-lead ECG into 10 anatomic regions with relatively high accuracy, as compared with methods developed in the 80's. There are several questions that deserve further study. The SVM used in the analysis is based on one versus one framework, rather than a unified multi-class classifier. Recent studies have shown that mutli-class classifiers may outperform binary classifiers such as one-versus-all [27, 28]. The samples in each of the 10 anatomic regions are not uniform, and could benefit from a formulation as an imbalanced learning problem, studied in Chapter VI of this thesis. These comments motivate the study of related problems in the following chapters.

CHAPTER III

Review of Optimization for Group Structured Variable Selection

3.1 Introduction

This chapter provides a brief a review of a useful optimization algorithm for solving the general sparse structured statistical learning problems proposed in Chapter I. Augmented Lagrangian Methods and Alternating direction method of multipliers (ADMM) have been successful techniques for solving problems with structured or patterned data matrices, especially when the objective functions can be viewed as the summation of several functions, and each function can be optimized efficiently if they are solved independently. This is exactly the case in the formulation (1.1), involving the summation of a loss function and a regularization function. We start by motivating augmented Lagrangian methods, which frame the constrained optimization problems as penalized unconstrained problems. ADMM is then introduced as an algorithm for optimizing over the augmented objective function.

3.2 Augmented Lagrangian Methods

Consider the constrained problem

$$\begin{aligned} &\min_{\mathbf{w}} f(\mathbf{w}) \\ &s.t.\ \ h(\mathbf{w}) = 0 \end{aligned} \tag{3.1}$$

in which f and h are twice continuously differentiable. We can find the minimum of f over the constraints by computing an unconstrained minimum of the augmented Lagrangian [29], defined as

$$L_\mu(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda' h(\mathbf{w}) + \frac{\mu}{2}||h(\mathbf{w})||_2^2. \tag{3.2}$$

This is the Lagrangian function for

$$\begin{aligned} &\min_{\mathbf{w}} f(\mathbf{w}) + \tfrac{\mu}{2}||h(\mathbf{w})||_2^2 \\ &s.t.\ \ h(\mathbf{w}) = 0. \end{aligned} \tag{3.3}$$

Notice that problem (3.3) has the same local minima as problem (3.1). Assume that $\mathbf{w}^*$ and λ^* satisfy

$$\nabla_{\mathbf{w}} L(\mathbf{w}^*, \lambda^*) = 0,\ \nabla_\lambda L(\mathbf{w}^*, \lambda^*) = 0$$

$$\mathbf{v}' \nabla^2_{\mathbf{ww}} L(\mathbf{w}^*, \lambda^*)\mathbf{v} > 0,\ \forall \mathbf{v} \neq 0, \nabla h(\mathbf{w}^*)'\mathbf{v} = 0$$

where L is Lagrangian $L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda' h(\mathbf{w})$. It can be shown that $\mathbf{w}^*$ is a strict local minimum of f subject to $h(\mathbf{w}) = 0$, and there exist $\gamma > 0$ and $\varepsilon > 0$ such that

$$L_\mu(\mathbf{w}, \lambda^*) \geq L_\mu(\mathbf{w}^*, \lambda^*) + \frac{\gamma}{2}||\mathbf{w} - \mathbf{w}^*||_2^2, \qquad \forall ||\mathbf{w} - \mathbf{w}^*||_2 < \varepsilon$$

for some μ such that $\nabla^2_{ww} L_\mu(w^*, \lambda^*)$ is positive definite [29]. Since $\mathbf{w}^*$ is an unconstrined local minimum of L_μ, we can try to solve problem 3.1 by minimizing L_μ.

3.3 Alternating Direction Method of Multipliers

Alternating direction method of multipliers (ADMM) [30, 31, 32, 33] is applicable to solving problems of the form

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{m}} H(\mathbf{w}) + G(\mathbf{m}) \\ &s.t.\ A\mathbf{w} + B\mathbf{m} = \mathbf{b}. \end{aligned} \tag{3.4}$$

According to the discussion in Section 3.2, we can attach a Lagrange multiplier to the linear constraints, add a quadratic penalty, and try to find the minimum of the augmented Larangian function. The augmented dual function is

$$L_{Dual}(\mathbf{t}) = \min_{\mathbf{w},\mathbf{m}} H(\mathbf{w}) + G(\mathbf{m}) + \mathbf{t}'(\mathbf{b} - A\mathbf{w} - B\mathbf{m}) + \frac{\mu}{2}||\mathbf{b} - A\mathbf{w} - B\mathbf{m}||_2^2.$$

Notice that

$$\nabla L_{Dual}(\mathbf{t}) = \mathbf{b} - A\mathbf{w} - B\mathbf{m},$$

which suggests a dual ascent method [30],

$$\mathbf{t} \leftarrow \mathbf{t} + \alpha \nabla L_{Dual}(\mathbf{t}).$$

The optimization over $\mathbf{w}$ and $\mathbf{m}$ can be implemented by coordinate descent, which results in the ADMM algorithm [30, 31, 32, 33] .

Algorithm 1: ADMM

set $\tau = 0$, choose $\mu > 0$, $\mathbf{m}_0$, $\mathbf{w}_0$, $\mathbf{t}_0$;
while *stopping criterion is not satisfied* **do**
  $\mathbf{w}_{\tau+1} = \arg\min_{\mathbf{w}} H(\mathbf{w}) + \mathbf{t}_\tau'(\mathbf{b} - A\mathbf{w}) + \frac{\mu}{2}||\mathbf{b} - A\mathbf{w} - B\mathbf{m}_\tau||_2^2$
  $\mathbf{m}_{\tau+1} = \arg\min_{\mathbf{m}} G(\mathbf{m}) + \mathbf{t}_\tau'(\mathbf{b} - B\mathbf{m}) + \frac{\mu}{2}||\mathbf{b} - A\mathbf{w}_{\tau+1} - B\mathbf{m}||_2^2$
  $\mathbf{t}_{\tau+1} = \mathbf{t} + \alpha(\mathbf{b} - A\mathbf{w}_{\tau+1} - B\mathbf{m}_{\tau+1})$
  $\tau = \tau + 1$

Convergence analysis of ADMM can be found in [34, 30, 31]. In particular, it has been shown that ADMM converges linearly if H and G are strongly convex [31]. A function f is strongly convex if there exists a constant $\sigma > 0$ such that

$$\eta f(\mathbf{u}) + (1-\eta) f(\mathbf{v}) - f(\eta \mathbf{u} + (1-\eta)\mathbf{v}) \geq \sigma \eta (1-\eta) ||\mathbf{u} - \mathbf{v}||_2^2$$

for $\eta \in [0, 1]$. From the KKT conditions,

$$\mathbf{b} - A\mathbf{w} - B\mathbf{m} = 0$$

$$0 \in \partial H(\mathbf{w}^*) - A'\mathbf{t}^*$$

$$0 \in \partial G(\mathbf{m}^*) - B'\mathbf{t}^*$$

Goldstein, et al. proposed the following residual measurements

$$\mathbf{r}_\tau = \mathbf{b} - A\mathbf{w}_\tau - B\mathbf{m}_\tau$$

$$\mathbf{q}_\tau = \mu A^T B(\mathbf{m}_\tau) - \mathbf{m}_{\tau-1}$$

to examine how closely the iterates satisfy the optimality conditions [31]. They showed that $||\mathbf{r}_\tau||_2^2 \leq O(1/\tau)$ and $||\mathbf{q}_\tau||_2^2 \leq O(1/\tau)$. Luo extended the convergence of 2 block ADMM to K block ADMM ($K > 2$), i.e., where the objective function becomes a summation of K functions [30]. The authors of [30] showed linear convergence for the general multi-block ADMM.

The ADMM algorithm is widely applied to structured optimization [35]. Consider the general regularized statistical learning formulation in (1.1). If minimizing V and R independently is efficient, then ADMM will be an appropriate candidate for developing the algorithm for problem (1.1). Suppose F in (1.1) can be parameterized

by $\mathbf{w}$, we can rewrite the problem in the ADMM formulation as

$$\min_{\mathbf{w},\mathbf{m}} V(\mathbf{w}, X, Y) + \lambda R(\mathbf{m})$$
$$s.t.\ \mathbf{w} - \mathbf{m} = 0.$$

By manipulating the linear and quadratic terms added to the objective in the augmented Lagrangian, and rewriting them in a single quadratic term [35], we have the ADMM algorithm for general optimization problems defined in (1.1). Although the

Algorithm 2: ADMM algorithm for general optimization problems defined by (1.1).

set $\tau = 0$, choose $\mu > 0$, $\mathbf{m}_0$, $\mathbf{w}_0$, $\mathbf{d}_0$;
while *stopping criterion is not satisfied* **do**
  $\mathbf{w}_{\tau+1} = \arg\min_{\mathbf{w}} V(\mathbf{w}) + \frac{\mu}{2}||\mathbf{w} - \mathbf{m}_\tau - \mathbf{d}_\tau||_2^2$
  $\mathbf{m}_{\tau+1} = \arg\min_{\mathbf{m}} \lambda R(\mathbf{m}) + \frac{\mu}{2}||\mathbf{w}_{\tau+1} - \mathbf{m} - \mathbf{d}_\tau||_2^2$
  $\mathbf{d}_{\tau+1} = \mathbf{d} - \mathbf{w}_{\tau+1} + \mathbf{m}_{\tau+1}$
  $\tau = \tau + 1$

convergence is based on strong convex assumptions, ADMM has been widely applied in practice, even to nonconvex problems [33]. The optimization problems for classifications in Chapter IV, Chapter V, and Chapter VI have convex objective functions but not strong convex functions, and the eigen-decomposition objective function for dimension reduction problems in Chapter VII is nonconvex. We empirically demonstrate the effectiveness of solving these statistical learning problems with structured sparsity constraints by the ADMM algorithm.

3.4 Conclusion

We discussed a general optimization algorithm applicable to statistical learning problems with constraints. In the following chapters, we specialize the algorithm to different loss functions V and regularizations R. In Chapter IV, the loss function

is the hinge loss [1, 36] for binary SVMs, and R is a L_1 L_2 mixed norm to impose group structured sparsity. Chapter V also discusses how to learn categorical responses with structured sparsity constraints, and extends sparse binary classification to sparse multi-class classification by specializing the loss function to a multi-class hinge loss. Chapter VII formulates a sparse supervised dimension reduction problem. The function V becomes concave and there are additional equality constraints to satisfy. Although there is currently no convergence guarantee for solving nonconvex problems by ADMM in the literature, we are able to show empirical performance improvements using our ADMM structured sparsity approach.

CHAPTER IV

Binary Classification with Variable Selection: Application to 3D Cell Microscopy

4.1 Introduction

In this chapter we formulate a sparse support vector machine in which the sparsity is imposed on groups of variables by penalizing the loss function with L_1/L_2 mixed norms. The mixed norms penalty avoids overfitting, which is common as variable dimension increases and becomes much larger the number of samples. We propose a novel algorithm to solve the binary SVM with group structured variable selection by ADMM, also known as splitting methods, discussed in Chapter III. The methodology is applied to 3D cell microscopy to learn the most discriminating population-based features to distinguish between different populations of *entamoeba histolytica* parasites. The features are the shape information obtained by spherical harmonics analysis of the cell surface, and the variable group structures are defined according to (1) the degree and order of spherical harmonics, and (2) the sequential time of observance. Experiments show that we can obtain significant improvement in terms of error rate relative to standard SVMs. The selected features provide biological insights and interpretations about the movement of cells.

The chapter is organized as follows. Section 4.2 reviews the binary classification,

and introduces SVM as a nonparametric classification method. Section 4.3 introduces different sparse SVMs that have been proposed in the literature and develops a new formulation to accommodate group structures on the features, followed by the implementation details in Section 4.4. In Section 4.5, we present the application of the SVM with group structured variable selection to 3D cell microscopy. Section 4.6 concludes the chapter.

4.2 Binary Classification

Suppose there is a system that takes an input vector $\mathbf{X} \in R^p$ and generates an output vector $\mathbf{Y} \in R^q$. The goal of statistical decision is to learn a function that describes the relationship between the independent variable $\mathbf{X}$ and dependent $\mathbf{Y}$. The relationship can be described as a function $f(\mathbf{X})$ [4]. When the output $\mathbf{Y}$ is a univariate categorical response, without loss of generality $\mathbf{Y} \in \{1, .., K\}$, it is known as a classification problem. In these problems, the task is to find an accurate prediction of $\mathbf{Y}$ given the input features $\mathbf{X}$, and the accuracy can be measured by the error rate, $E[\mathbf{Y} \neq f(\mathbf{X})]$. The optimal solution to minimize the error rate is the maximum a posteriori estimation, often denoted as the MAP rule [37]. We can view the decision as partitioning the observation space into decision regions, D_k, $k = 1, ..., K$. The error rate can be written as

$$E[\mathbf{Y} \neq f(\mathbf{X})] = 1 - E[\mathbf{Y} = f(\mathbf{X})] = 1 - \sum_k \pi_k \int_{D_k} h(\mathbf{X} = x|\mathbf{Y} = k)dx$$

in which h denotes the density function and π_k denotes the prior probability $P(\mathbf{Y} = k)$. Therefore, the optimal solution to partition the observation space is to assign x to D_k for which $\pi_k h(x|\mathbf{Y} = k)$ is maximized, or equivalently for which $h(\mathbf{Y} = k|\mathbf{X} = x)$ is maximized.

If we know the exact joint distributions of $\mathbf{X}$ and Y under each class, then an

optimal minimum probability of error classifier exists and is equal to a likelihood ratio comparator. The optimal classifier can be found by fitting the parameters in the model to the data. For example, assume that given the class label k, the density function $h(\mathbf{X}|\mathbf{Y}=k)$ is a member of the Gaussian family of models having form,

$$h(\mathbf{X}=\mathbf{x}|\mathbf{Y}=k)=\frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)}$$

in which μ_k and Σ_k are the known mean and known covariance matrix in each class k respectively.The minimium probability of error classifier is given by the MAP decision rule of the form

$$\arg\max_k\{\pi_k h(\mathbf{X}=\mathbf{x}|Y=k)\}=\arg\max_k\{\ln\pi_k-\frac{1}{2}\ln|\Sigma_k|-\frac{1}{2}(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)\}.$$

When the covariance matrices Σ_k are identical the MAP rule classifies an observation $\mathbf{x}$ to the class k having mean μ_k that is at minimum Mahalonobis distance from X, where the Mahalanobis distance is $(\mathbf{x}-\mu_k)'\Sigma_k^{-1}(\mathbf{x}-\mu_k)$. If the covariance matrices Σ_k are identical then the MAP classifier reduces to a linear classifier. In this case the MAP rule simplifies to

$$\arg\max_k\{\pi_k h(\mathbf{X}=\mathbf{x}|Y=k)\}=\arg\max_k\{\ln\pi_k+\mathbf{x}'\Sigma^{-1}\mu_k-\frac{1}{2}\mu_k'\Sigma^{-1}\mu_k\}.$$

When the mean and covariance are unknown but there exists a training data set $(y_i,\mathbf{x}_i), i=1,\ldots,n$, the means and covariance matrices can often be estimated and substituted into the MAP decision rules above. Another parametric model approach to classification is logistic regression. In this approach one models the posterior probability as a logistic function of the independent variables $\mathbf{x}$

$$P(\mathbf{Y}=k|\mathbf{X}=\mathbf{x})=\frac{\exp(\beta_k'\mathbf{x}+\beta_{0,k})}{\sum_l\exp(\beta_l'\mathbf{x}+\beta_{0,l})}. \tag{4.1}$$

where β_k are model parameters that are fitted to the data. In practice, the logistic regression model is fitted by maximum likelihood [4], which is an optimal solution when the prior probabilities π_k are the same for all k.

Support Vector Machines (SVM) [1, 36] have been a popular classification technique recently. Instead of imposing a parametric model on the joint distributions, it was first designed for binary classification, searching for a hyperplane to separate the two classes. Let $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ be a decision function that assigns observation $\mathbf{x}$ to $sign(f(\mathbf{x}))$, then $f(\mathbf{x}) = 0$ is a hyperplane in R^p. In the case where the classes are linearly separable, in other words there exists a hyperplane such that $y_i(\mathbf{w}'\mathbf{x}_i + b) > 0$ for all $y_i \in \{-1, 1\}$, and $i \in \{1, ..., N\}$, there may be multiple solutions of the hyperplane. Vapnik proposed to search for a hyperplane that brings the largest margin between the classes [1]. Consider the following problem

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||_2^2 \\ &s.t. \;\; y_i(\mathbf{x}_i'\mathbf{w} + b) \geq 1 \quad , \quad i = 1, ..., N. \end{aligned} \tag{4.2}$$

The signed distance of a point $\mathbf{x}$ to the hyperplane defined by $f(\mathbf{x}) = 0$ is $\frac{1}{||\mathbf{w}||_2} f(\mathbf{x})$. Hence, when the data points are linearly separable, the optimization is searching for the solution that gives the largest margin, see Fig. 4.1.

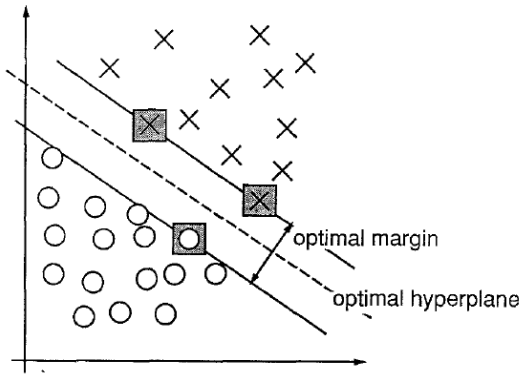

Figure 4.1: Separating Hyperplane in support vector machines [1].

When the data points are not linearly separable, one can introduce slack variables ξ_i to penalize the misclassified points. The generalization to nonseparable/overlap cases is known as the soft margin hyperplane proposed in [1].

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \frac{1}{2}||\mathbf{w}||_2^2 \quad &+ \quad \gamma \sum_{i=1}^{N} \xi_i \\ s.t. \;\; y_i(\mathbf{x}_i'\mathbf{w} + b) \geq 1 - \xi_i \quad &, \quad \xi_i \geq 0, \quad i = 1, ..., N. \end{aligned} \tag{4.3}$$

The solution to the generalized soft margin SVM (4.3) can be found by quadratic programming techniques. It is worthwhile to analyze the solution of the program to obtain insight to SVMs [4]. First we form the Lagrange function

$$L = \frac{1}{2}||\mathbf{w}||_2^2 + \gamma \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{w}'\mathbf{x}_i + b) - (1 - \xi_i)] - \sum_{i=1}^{N} \mu_i \xi_i,$$

in which α_i and μ_i are the Lagrange multipliers. To minimize with respect to the variables $\mathbf{w}$, b and ξ, we set the derivatives with respect to these variables to 0, obtaining

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^{N} \alpha_i y_i &= 0 \\ \alpha_i = \gamma - \mu_i, \quad i &= 1, ..., N \end{aligned}$$

and the Lagrange multipliers α_i , μ_i and the slack variables ξ_i should be nonnegative. The dual objective function can then be expressed as

$$L_{Dual} = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x_j}.$$

From the dual objective function and the invoking KKT conditions

$$\alpha_i[y_i(\mathbf{w}'\mathbf{x}_i + b) - (1 - \xi_i)] = 0$$

$$\mu_i\xi_i = 0$$

$$y_i(\mathbf{w}'\mathbf{x}_i + b) - (1 - \xi_i) \geq 0$$

we obtain the optimal hyperplane weight vector $\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i$. It is obvious from the KKT conditions that the weight vector only depends on the data points whose Lagrange multipliers satisfy $\alpha_i > 0$. These points are the support vectors. In other words, the SVM solution, specified by the hyperplane $\{\mathbf{u} \in R^p : \mathbf{u}'\mathbf{w} + b = 0\}$, only depends on a few samples that are close to the class boundary. It is useful to view the slack variables as part of a loss function, know as the hinge loss function. Since we are minimizing over the summation of ξ_i and ξ_i should satisfy the conditions $\xi_i \geq 0$ and $y_i(\mathbf{x}_i'\mathbf{w} + b) \geq 1 - \xi_i$, one can conclude that $\xi_i = 0$ if $1 - y_i(\mathbf{x}_i'\mathbf{w} + b) \leq 0$, otherwise $\xi_i = 1 - y_i(\mathbf{x}_i'\mathbf{w} + b)$. We can replace ξ_i by the hinge loss, which is represented as $[1 - y_i f(\mathbf{x_i})]_+$.

We rewrite the problem in the general form discussed in Chapter III.

$$\min_f \sum_{i=1}^{n} V(f, \mathbf{x_i}) + R(f) \tag{4.4}$$

where V denotes a convex loss function that upper-bounds the 0-1 loss (misclassification error), and R is a regularization function that enforces smoothness. The loss function should be a good surrogate for the non-convex 0-1 loss, such as the hinge

loss function

$$\begin{aligned} V(f, \mathbf{x_i}) &= [1 - y_i f(\mathbf{x_i})]_+ \\ R(f) &= \frac{\lambda}{2}||\mathbf{w}||_2^2. \end{aligned} \tag{4.5}$$

The fact that the hinge loss function doesn't penalize those data points that satisfy the inequality $1 - y_i f(\mathbf{x}_i) \leq 0$ confirms that the solution only depends on the support vectors. Different modifications of the loss function have been proposed in the literature. For example, the truncated hinge loss is designed to reduce the effect of outliers [38], differentiable approximations or the integral of sigmoid function are proposed to enable second order methods, e.g., Newton method to be applied to SVM [39, 40], and smoothing the hinge loss by a strongly convex function can lead to faster algorithms [41, 42]. In the following discussion, we focus on the standard hinge loss function.

4.3 Sparse Binary Support Vector Machines

In order to encourage a sparse normal vector $\mathbf{w}$ in [43] it was proposed to apply a L_1 norm penalty on the weight vector by replacing the regularization function in (4.3) by

$$R(f) = \frac{\lambda}{2}||\mathbf{w}||_1.$$

A sparse weight vector has advantages over the standard SVM weight vector when there are redundant noise features and the total number of samples is much less than the dimension of the features. The L_1 norm was introduced to induce sparsity in regression problems [44, 45]. Several different approaches have been proposed to solve the sparse SVM optimization, such as newton method [46], unconstrained convex differentiable minimization [47]. A comparison of optimization methods for L_1 regularized SVM can be found in [48]. The parameter for the regularization term

can be found by cross-validation or regularization path techniques [49, 50].

Later more extensions of L_1 penalty were proposed. The weighted L_1 penalty [51] constructs the adaptive weights using the 2-norm SVM, called the hybrid SVM.

$$R(f) = \lambda \sum_{j=1}^{p} \alpha_j |w_j|$$

Fussed lasso was proposed to penalize the L_1 norm of both the coefficients and their successive differences [52], which can expressed as

$$R(f) = \frac{\lambda_1}{2}||\mathbf{w}||_1 + \frac{\lambda_2}{2}\sum_{j=2}^{p}||w_j - w_{j-1}||_1.$$

Elastic-net penalty [53] has also been applied to the SVM [54]. This doubly regularized SVM has a mixture of the L_1 norm penalty and the squared L_2 norm penalty. The former term allows variable selection, whereas the latter one groups the correlated variables together.

$$R(f) = \frac{\lambda_1}{2}||\mathbf{w}||_1 + \frac{\lambda_2}{2}||\mathbf{w}||_2^2.$$

A few more extensions have been proposed recently. For example, the non-convex penalty, smoothly clipped absolute deviation (SCAD) was proposed to overcome the bias of L_1 penalty on large coefficients [55]. The SCAD penalty with $a > 2$ is

$$R(f) = \begin{cases} \lambda||\mathbf{w}||_1 \text{ if } ||\mathbf{w}||_1 \leq \lambda \\ -\frac{(||\mathbf{w}||_2^2 - 2a\lambda||\mathbf{w}|| + \lambda^2)}{2(a-1)} \text{ if } \lambda < ||\mathbf{w}||_1 \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} \text{ if } ||\mathbf{w}||_1 > a\lambda. \end{cases}$$

A more general L_q norm penalty is discussed in [56]

$$R(f) = \lambda \sum_{j=1}^{p} ||w_j||^q$$

and a combination of L_0 and L_1 penalties was proposed in [57]. The former introduces local quadratic approximation algorithm to solve the optimization, and the latter optimization is converted into a mixed integer programming problem.

In this chapter we apply a method to promote group sparsity by applying a mixed L_1/L_2 norm to the weight vector. The technique had been successfully applied to regression problems [58, 59], and can be extended to the SVM by reformulating the regularization function as

$$R(f) = \sum_{g=1}^{G} ||\mathbf{w}_{I_g}||_2 \tag{4.6}$$

where I_g is the set indexing the variables that are in the g_{th} group. This will provide better interpretation when the group structure is given.

4.4 Algorithmic Implementation

To solve the optimization problem (4.4) with the sum of loss function V (4.5) and regularization R specified by (4.6), we use a variable splitting approach that is tailored to optimization of a sum of two functions. Consider an unconstrained optimization problem in which the objective is the sum of two functions [35]:

$$\min_{v} f_1(v) + f_2(v)$$

The variable splitting method introduces a new variable w as the argument of f_2, under the constraint that $v = w$.

$$\min_{v,w} f_1(v) + f_2(w) \ s.t. \ v = w$$

This constrained problem can be solved by unconstrained methods if we add a quadratic penalty

$$\min_{v,w} f_1(v) + f_2(w) + \frac{u}{2} \|v - w\|_2^2$$

which suggests an alternating splitting algorithm, alternating between solving the unconstrained problem with respect to v and w. Applying these ideas to our problem, we can modify the optimization to form an equivalent problem, in which the newly introduced variable $\mathbf{m}$ is constrained such that $\mathbf{m} = \mathbf{w}$:

$$\min_{\mathbf{w},\mathbf{m}} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\sum_{g=1}^{G}\left\|\mathbf{m}_{I_g}\right\|_2, \text{ subject to } \mathbf{m} = \mathbf{w} \text{ and}$$

$$(\mathbf{w}'\,\mathbf{x_i} + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i$$

This alternating splitting method yields the following novel algorithm for group sparse classification:

Algorithm 3: Proposed sparse SVM with group structured variable selection.

1 set $\tau = 0$, choose $\mu > 0$, $\mathbf{m}_0$, $\mathbf{w}_0$, $\mathbf{d}_0$
2 **while** *stopping criterion is not satisfied* **do**
3 $\quad \mathbf{w}_{\tau+1} = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \frac{\mu}{2}\|\mathbf{w} - \mathbf{m}_\tau - \mathbf{d}_\tau\|_2^2$
4 $\quad$ s.t. $\forall i \;\; (\mathbf{w}'\mathbf{x_i} + b) \geq 1 - \xi_i, \xi_i \geq 0$
5 $\quad \mathbf{m}_{\tau+1} = \arg\min_{\mathbf{m}} \lambda\sum_{g=1}^{G}\|\mathbf{m}_{I_g}\|_2 + \frac{\mu}{2}\|\mathbf{w}_{\tau+1} - \mathbf{m} - \mathbf{d}_\tau\|_2^2$
6 $\quad \mathbf{d}_{\tau+1} = \mathbf{d}_\tau - \mathbf{w}_{\tau+1} + \mathbf{m}_{\tau+1}$
7 $\quad \tau = \tau + 1$

Line 3 in Algorithm 3 is a quadratic programming problem. We apply the dual coordinate descent method implemented for large-scale SVM to solve the QP [60, 61]. It is similar to L_2 regularized SVM formulation except the offset term $\mathbf{m}_{it} + \mathbf{d}_{it}$ in the squared L_2 norm. The corresponding dual Lagrange can be expressed as

$$L_{Dual} = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i'\mathbf{x_j} - (\mathbf{m}_{it} + \mathbf{d}_{it})\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i.$$

And given $\boldsymbol{\alpha}$, the corresponding weight vector is $\mathbf{w}_{it} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i + \mathbf{m}_{it} + \mathbf{d}_{it}$. We adopt the algorithm in [60] to iteratively solve for the updates $\mathbf{w}_{it+1}$, shown in (Algorithm

4).

Algorithm 4: The dual coordinate descent method for the subproblem in Algorithm 3

while *stopping criterion is not satisfied* **do**
  for *i=1* **to** N **do**
    $G = y_i \mathbf{w}' \mathbf{x}_i - 1$
    $\mathrm{PG} = \begin{cases} \min(\mathrm{G}, 0) & \text{if } \alpha_i = 0 \\ \max(G, 0) & \text{if } \alpha_i = \gamma \\ G & \text{if } 0 < \alpha_i < \gamma \end{cases}$
    If $PG \neq 0$
      $\Delta\alpha_i = \min(\max(\alpha_i - G/\mathbf{x}_i' \mathbf{x}_i, 0), \gamma) - \alpha_i$
      $\mathbf{w} \leftarrow \mathbf{w} + \Delta\alpha_i y_i \mathbf{x}_i$

Line 5 in Algorithm 3 has a closed form solution. Let $\mathbf{c} = \mathbf{w}_{\tau+1} - \mathbf{d}_\tau$, then the solution of the subvector $\mathbf{m}_{I_g}$ is given as $\mathbf{m}_{I_g} = [||\mathbf{c}_{I_g}||_2 - \frac{\lambda}{\mu}]_+ \frac{\mathbf{c}_{I_g}}{||\mathbf{c}_{I_g}||_2}$. This is a multidimensional shrinkage-thresholding operator [62].

4.5 Application: Spherical Harmonics Based Classification and Analysis of Highly Deforming Cells in 3D Microscopy

Characterizing cell morphology is of crucial importance to study many fundamental biological processes involving cellular dynamics. For instance, cells and unicellular organisms characterized by amoeboid motion exhibit an ordered cycle of complex shape changes in order to generate movement [63]. Understanding cellular shape and movement requires efficient shape quantification tools to describe and classify the wide variety of shape configurations, with the aim of deciphering the biological mechanisms underlying cell motion.

In the context of highly deformable cells, such as cells exhibiting amoeboid motion, robust shape description and analysis is particularly challenging, due to the high degree of variability that can be observed within a so-called homogeneous population, while different populations may exhibit visually similar deformation patterns.

Therefore, traditional shape descriptors based on a voxelized representation of the cell volume are either too sensitive to small shape variations, and thus cannot be used to discriminate different cell populations. To address this issue, the community has recently shifted toward more advanced mathematical descriptors based on frequency analysis, such as the SPherical HARMonics (hereafter SPHARM) transform [64, 65, 66]. The SPHARM transform considers any closed surface as a function of the unit sphere, and decomposes this function into a unique set of coefficients in a basis of polynomial functions, facilitating subsequent shape characterization and classification. This technique offers interesting properties such as position and orientation invariance (when properly handled), and is thus well suited for shape sets with high variability such as living cells [2].

We present a novel approach to classify populations of living cells based on shape information obtained via spherical harmonics analysis of the cell surface. Classification is achieved using a Support Vector Machine classifier with group structured variable selection, using the extracted spherical harmonics coefficients and their group-based correlation as feature vector. The variable selection lets the classifier isolate the most representative features, which can be further analyzed in a qualitative manner.

Supervised classification algorithms such as support vector machines (SVM), linear discriminant analysis (LDA), and nearest neighbor algorithms [67] are most often associated with machine automation of pattern recognition. When the feature dimension is high, sparsity-constrained classifiers are especially desirable since they reduce otherwise severe over-fitting errors. Such classifiers perform variable selection by eliminating those feature dimensions that are the least powerful discriminants, retaining only the most important ones. These important variables alone are of interest as they can provide an explanatory model, similarly to the principal component in LDA that best differentiate between classes. By applying this approach to SPHARM features, we wish to build an optimal classifier which can be used as a tool for exploring the

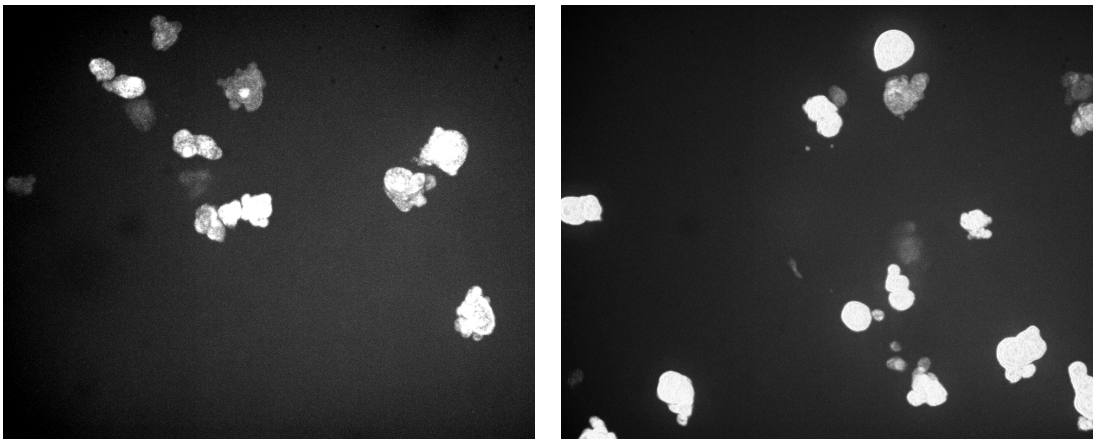

Figure 4.2: Image samples of two cell populations observed in fluorescence microscopy. Left: WT population. Right: ΔCP5 population. Images are Maximum Intensity Projection of the original 3D imaging data (cf. section 4.7 for more details). These samples illustrate the difficulty in distinguishing the two populations using simple visual assessment.

major shape differences between cell populations.

The proposed approach is used to classify populations of *entamoeba histolytica* parasites, which are unicellular organisms responsible for the amoebiasis disease. A recent study has shown that parasites with chemically altered collagen degradation ability were still able to migrate through collagen gels at the same speed as unaltered parasites [68]. This finding suggested that cell deformation in the chemically modified condition was also altered, however only limited shape information was available to precisely describe this difference. A raw image sample of each population is shown in Fig. 4.2, illustrating the difficulty of the problem regarding simple visual distinction.

4.5.1 Approach

The goal is to characterize the shape of highly deforming cells evolving in a 3D environment, and to use this shape information to classify different populations with visually similar deformation patterns. We describe below the successive steps of the

approach, and defer technical aspects on how the cell shape information is extracted in the Methods section.

4.5.1.1 Spherical Harmonics Analysis

On the unit sphere, an orthonormal basis for the Hilbert space of square-integrable function is given by the spherical harmonics (cf. Fig. 4.3):

$$Y_l^m(\theta, \varphi) = k_l^m P_l^m(\cos\theta) e^{im\varphi},$$

where l and m are respectively the degree and order of the harmonic, k_l^m is the expansion coefficient and P_l^m is the associated Legendre polynomial.

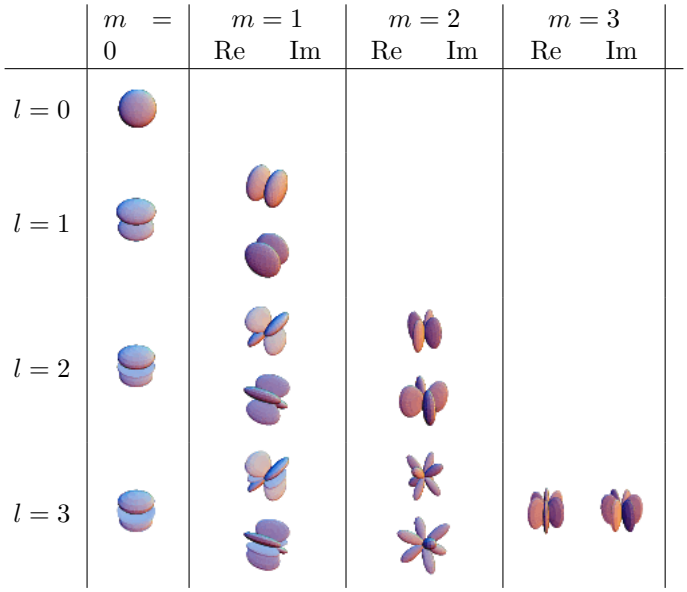

Figure 4.3: First Few Spherical Harmonics (l: degree, $m \leq l$: order). Adapted from *Weisstein, E.W. "Spherical Harmonic" from MathWorld – A Wolfram Web Resource, http://mathworld.wolfram.com/SphericalHarmonic.html*

Using this basis, any spherical scalar function $f(\theta, \varphi)$ can be expanded as follows:

$$f(\theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \hat{f}_l^m Y_l^m(\theta, \varphi),$$

where $\hat{f}(l,m)$ is the (l,m) harmonic coefficient, given by:

$$\hat{f}_l^m = k_l^m \int_0^{\pi} \int_0^{2\pi} e^{-im\varphi} f(\theta,\varphi) P_l^m(\cos\theta) \sin\theta \; d\varphi \; d\theta.$$

The coefficients $\hat{f}_l^m$ are unique and can thus describe any arbitrary shape. The spectral decomposition of the input signal is then straightforward: lower degrees (i.e., l) correspond to low frequencies and hence describe the global shape of the object, while higher degrees describe the details of the surface, as illustrated in Fig. 4.4.

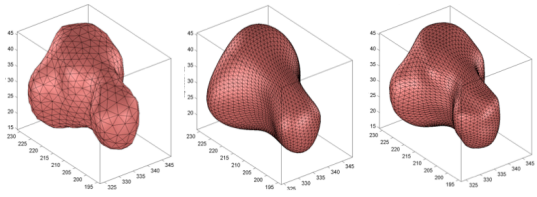

Figure 4.4: SPHARM reconstruction of an arbitrary mesh (left) with a maximum level $l_{\max}$ of 5 (middle) and 11 (right). Higher levels reconstruct the finer details of the surface. Adapted from [2].

Higher dimensional (non-scalar) spherical functions can also be described by expanding each component of the function independently. Here we perform the transform on a surface defined in Cartesian space (x,y,z) and mapped to a spherical signal defined in the polar system (θ,φ) as $\mathbf{v}(\theta,\varphi) = (x(\theta,\varphi)\ y(\theta,\varphi)\ z(\theta,\varphi))'$. As (θ,φ) runs over the sphere, $\mathbf{v}(\theta,\varphi)$ runs over the object surface. By applying the SPHARM transform to each component of $\mathbf{v}(\theta,\varphi)$ independently, we obtain coefficients with three components.

A classical problem in shape comparison and classification resides in the degrees of freedom of the shape space. In order to enforce translational invariance, all surfaces are translated to an arbitrary origin before the mapping step. For rotational invariance, one may rely on the intrinsic property of the SPHARM transform that any arbitrary rotation in the parameter space corresponds to a rotation in the ob-

ject space. Following [64], we expand each component of the spherical signal $\mathbf{v}(\theta, \varphi)$ independently, yielding three set of coefficients C_x, C_y, C_z, and compute rotationally invariant descriptors as $\hat{C} = \sqrt{(C_x^2 + C_y^2 + C_z^2)}$. These translationally and rotationally invariant descriptors will be used in the remainder of this work.

4.5.2 Population Shape Discrimination and Group Structured Variable Selection

Assume that we have a dataset of 2 populations, comprised of n samples. Each sample is observed over K cells from the same type of population and each cell is recorded in a a video of Q frames. The tracking of each cell is described in Section 4.7. Motivated from the idea of identifying the cell types in a Petri dish, in which often multiple cells are present in the microbiological culture, we propose to use a SPHARM representation, and extract from each sample a set of population feature vector $\mathbf{x_i} \in R^p$ and a population label $y_i \in \{1, -1\}$, where 1 represents the ΔCP5 population and -1 represents the WT population. A standard binary classifier can be trained over this data to give minimum classification error probability. However, when the feature dimension p is large such a classifier will suffer from severe over-fitting error. To overcome this deficiency, sparsity-penalized classifiers have been developed [43, 69], that "sparsify" the feature vector, i.e., finding a reduced number of features that attain the minimum cross-validated error. This is tantamount to selection of the most discriminating features. We adopt a similar sparse penalty approach to include group structure unique to the problem of estimating shape parameters in a cell population.

4.5.2.1 Population features

We obtain the rotationally invariant descriptors for each cell, in each frame at a given time point t, expressed over the SPHARM from $l = 0$ to $l = 5$. These features

can be represented as $\hat{C}_l^m(i, cell_k, t)$, where $i \in \{1, 2, ..., n\}$ is the sample indicator, $k \in \{1, ..., K\}$ indicates the cell and $t \in \{t_1, ..., t_Q\}$ indicates the time at which the frame is acquired. The feature space of SPHARM representation does not include the information shared among the cells in the same sample, which can not be observed by looking at a single cell image. We propose a set of population features to capture the inter-cell and intra-cell information, defined as follows.

Inter-cell features: Mean, auto-correlation and cross-correlation averaged over cells provide features of dimensions $21Q$, $C_2^{21}Q$, and $21C_2^Q$ respectively.

$$\mu_l^m(i,t) = \frac{1}{K}\sum_{k=1}^{K} \hat{C}_l^m(i, cell_k, t)$$

$$\rho_{(l_1,m_1),(l_2,m_2)}(i,t) = \frac{\sum\limits_k \hat{C}_{l_1}^{m_1}(i, cell_k, t)\hat{C}_{l_2}^{m_2}(i, cell_k, t)}{\sigma_{\hat{C}_{l_1}^{m_1}}(i,t)\sigma_{\hat{C}_{l_2}^{m_2}}(i,t)}$$

$$R_l^m(i,t_1,t_2) = \frac{\sum\limits_k \hat{C}_l^m(i, cell_k, t_1)\hat{C}_l^m(i, cell_k, t_2)}{\sigma_{\hat{C}_l^m}(i,t_1)\sigma_{\hat{C}_l^m}(i,t_2)}$$

Intra-cell features: Mean and cross-correlation averaged over time provide features of dimensions $21K$ and $C_2^{21}K$ respectively.

$$\tilde{\mu}_l^m(i, cell_k) = \frac{1}{Q}\sum_{q=1}^{Q} \hat{C}_l^m(i, cell_k, t_q)$$

$$\tilde{\rho}_{(l_1,m_1),(l_2,m_2)}(i, cell_k) = \frac{\sum\limits_{q=1}^{Q} \hat{C}_{l_1}^{m_1}(i, cell_k, t_q)\hat{C}_{l_2}^{m_2}(i, cell_k, t_q)}{\sigma_{\hat{C}_{l_1}^{m_1}}(i, cell_k)\sigma_{\hat{C}_{l_2}^{m_2}}(i, cell_k)}$$

4.5.2.2 Loss Function and Regularization

We follow the framework of sparse SVM discussed in section 4.3.

$$\min_f \sum_{i=1}^{n} V(f, \mathbf{x_i}) + \lambda R(f)$$

In order to obtain better interpretation of the features with group structures, we select the loss function as the hinge loss function and the regularization function as the mixed L_1, L_2 norm penalty.

$$V(f, \mathbf{x_i}) = [1 - y_i f(\mathbf{x_i})]_+, \quad R(f) = \sum_{g=1}^{G} ||\mathbf{w}_{I_g}||_2.$$

I_g is the set indexing the variables that are in the g_{th} group, which is useful as we define the groups in section 4.5.2.3. Details about the algorithm for solving the optimization problem is discussed in section 4.4.

4.5.2.3 Group Sparsity Constraints for Spherical Harmonic Features

We propose a novel set of group structures based on the population features, taking into account that the SPHARM representation has several orders m corresponding to the same degree level l and the auto-correlation provides several features with the same time difference parameter $\Delta t = |t_1 - t_2|$.

$$I_{\mu_{l^*}} = \{\mu_l^m(i,t) : l = l^*, \forall i, t\}$$

$$I_{\rho_{(l_1^*, l_2^*)(t)}} = \{\rho_{(l_1,m_1),(l_2,m_2)}(i,t) : l_1 = l_1^*, l_2 = l_2^*, \forall i\}$$

$$I_{R_{l^*}(\Delta t)} = \{R_l^m(i,t_1,t_2) : l = l^*, |t_1 - t_2| = \Delta t, \ \forall i\}$$

$$I_{\tilde{\mu}_{l^*}} = \{\tilde{\mu}_l^m(i, cell_k) : l = l^*, \forall i, k\}$$

$$I_{\tilde{\rho}_{(l_1^*, l_2^*)}} = \{\tilde{\rho}_{(l_1,m_1),(l_2,m_2)}(i, cell_k) : l_1 = l_1^*, l_2 = l_2^*, \forall i, k\}$$

4.5.3 Results

We apply the presented approach on *entamoeba histolytica* parasites, with the goal to distinguish a wild-type population (WT) from a ΔCP5 population. In order to assess the performance of the proposed approach, and hence how well the two populations can be distinguished, we compare classification results obtained by classical SVM and by the proposed sparse SVM with group structured variable selection. In each class, WT and Δ CP5, K cells are randomly selected to form a sample, and each cell provides a video of length Q, in which the starting point of the video is random if Q is less than the entire length of the video. To avoid imbalance in the training samples, we randomly subsample the larger class, so that the sizes of each class are exactly the same. The regularization parameter λ is chosen with 2-fold cross-validation, 5 random permutations, and the performance is evaluated by leaving one sample from each class out for the test set, repeated until every sample has been tested. The entire experiment is conducted with 10 trials, and the number of cells in each sample $K \in \{1, 2, ..., 14\}$, the number of frames in each video $Q \in \{10, 15, 20, ..., 40\}$ are varied to find the best combination. The best performance is found with the combination of sparse SVM and all the population features. The averaged performances can be found in Table 4.1. The best result by sparse SVM with all the population features occur when $K = 12$ and $Q = 30$, whereas standard SVM without sparsity attains the best when $K = 12$ and $Q = 30, 35$. The combinations of K and Q are discussed in Section 4.7.

method	SVM	sparse SVM
error rate	0.1250	0.0875
standard error	(0.0347)	(0.0353)

Table 4.1: Comparison between the classical SVM and sparse SVM with group structured variable selection.

The sparse SVM with group structured variable selection provides not only better

performance but useful interpretation of the features. To understand the roles of each set of population features, we plot the selection frequency of them, examples of inter-cell features and intra-cell features are illustrated in Fig. 4.5 and Fig. 4.6 respectively.

It is interesting to notice that the mean of $l \geq 1$ SPHARM averaged over cells are important for discriminating the WT and ΔCP5 populations. We plot the mean of the features, i.e., $\frac{1}{|\{i:y_i=1\}|}\sum_{y_i=1}\mu_l^m(i,t)$ and $\frac{1}{|\{i:y_i=-1\}|}\sum_{y_i=-1}\mu_l^m(i,t)$, within the ΔCP5 and WT populations independently in Fig. 4.7 and Fig. 4.8 for SPHARM $l=1$ and $l=5$ respectively to illustrate the differences. Notice that in lower degree SPHARM, the WT population has lower coefficients than the ΔCP5 population does, whereas in higher degree SPHARM, the ΔCP5 shows larger coefficients than the WT population does.

The auto-correlation of higher order SPHARM also plays an important role in the classification. By showing the mean of these features averaged over each population $\frac{1}{|\{i:y_i=1\}|}\sum_{y_i=1}R_l^m(i,t_1,t_2)$ and $\frac{1}{|\{i:y_i=-1\}|}\sum_{y_i=-1}R_l^m(i,t_1,t_2)$ in Fig. 4.9, we notice that the WT population de-correlates faster than the ΔCP5 population. We can quantify the decay of the correlation in each population by measuring

$$v_l^m(\Delta t,1)=\frac{\sum_{|t_1-t_2|=\Delta t,\ y_i=1}R_l^m(i,t_1,t_2)}{|\{i,t_1,t_2:|t_1-t_2|=\Delta t,\ y_i=1\}|}$$

$$v_l^m(\Delta t,-1)=\frac{\sum_{|t_1-t_2|=\Delta t,\ y_i=-1}R_l^m(i,t_1,t_2)}{|\{i,t_1,t_2:|t_1-t_2|=\Delta t,\ y_i=-1\}|}$$

for ΔCP5 and WT respectively. Fig. 4.10 shows one of the major differences between ΔCP5 and WT populations, which suggests that the cell shape of the ΔCP5 population at a given time t is greatly related to the shape at $t-\Delta t$, in which $\Delta t>0$, whereas the WT population can change the cell shape with less restriction.

Furthermore, the variable selection frequency is relatively high on the cross corre-

lation SPHARM features averaged both over the cells and over time between degree $l = 4$ and $l = 5$. This may suggest that the higher degree SPHARM features describe subtle shape information that is important in explaining the movement of *entamoeba histolytica* parasites.

4.6 Conclusion

Cell morphology is a key factor implicated in numerous biological processes, from organ development to disease models. Yet, characterizing highly deforming cells based on their morphology is particularly challenging due to the great variability of shapes within a given population. Our goal is to develop quantitative tools to classify highly deforming cells based on 3D shape information, and to extract key features describing the differences between populations in a qualitative manner.

We have presented a supervised classification approach to characterize and classify populations of highly deforming cells observed in 3D microscopy. By performing classification with group structured variable selection using spherical harmonics analysis of the cell shape, we obtained a classification accuracy far beyond what can be achieved via visual inspection. Moreover the variable selection process allowed to isolate several features of interest, allowing to pinpoint the differences between the observed populations in a qualitative manner. We believe the proposed set of tools is sufficiently generic to be applied on numerous types of cells. An immediate application of this tool can be thought in the context of rapid disease diagnosis, where cell populations can be rapidly analyzed to establish a diagnosis.

In the current study, we have exploited static shape information, although this information was extracted from temporal sequences. We are currently extending the current study spatiotemporal cellular morphodynamics, in order to embed trajectory analysis into the process, with the goal to build spatiotemporal models of cellular deformation, and complement current studies in the field of biology or biophysics

[70, 71].

4.7 Appendix

Imaging protocol The cells are embedded into a 3D collagen matrix, letting them deform and move freely by carving their way through the matrix. Images are acquired on a laser scanning confocal microscope, producing 3D stacks of about 32 consecutive 2D slices of size 1024×1024 pixels each. Finally, 3D time-lapse sequences are obtained by acquiring 90 of these 3D stacks with an interval of 10 seconds between stacks (i.e., the total observation time per condition is 15 minutes), yielding a typical data size of 5.625 GigaBytes per 4D dataset. The total number of cells is 173 and the total number of frames, 12863.

Cell surface extraction Cell surfaces were automatically segmented and tracked throughout the sequence the Icy software [72]. Cells were first pre-detected on the first frame of each sequence using the H-KMeans plugin (V4N7Q1), which creates a binary volume for each object in the image. Using the 3D Active Meshes plugin (P6W8D6), each binary volume is triangulated into a 3D triangular mesh, which is then deformed automatically toward the cell boundary by minimizing an energy functional comprising image-based and geometry-based terms. Segmentation is achieved when the mesh reaches a steady-state, corresponding to a minimizer of the energy. Tracking over time is done by initializing each subsequent frame by the resulting mesh obtained on the previous frame. More details on this method can be found in [73].

Spherical mapping and harmonics transform Spherical mapping consists in mapping the extracted cell surface onto the unit sphere, thus parameterizing the original surface into a spherical signal that can be analyzed using spherical harmonics. We refer the reader to [74] for a review of such parameterization methods. Surface

mapping and spherical harmonics features were performed using the SPHARM-MAT toolbox (http://www.nitrc.org/projects/spharm-mat). In order to minimize shape distortions inherent to the mapping process, we adopted the CALD approach [75], which provides a bijective mapping on any arbitrary closed triangular mesh onto the unit sphere while minimizing area and length distortions of the mapped mesh. An illustration of the surface mapping process is given in Fig. 4.11.

Optimal feature vector construction The optimal set of parameters for the length of the video (Q), and the number of cells in each sample (K) are determined by grid search. Fig. 4.12 shows how the performance in terms of error rate varies as these parameters change. As the number of cells in each sample increases, the performance improves. However, the error rate also increases when the number of cells in each sample reaches about 13. This is because the total number of cells is fixed, the number of training samples decreases as we increase the number of cells in each sample.

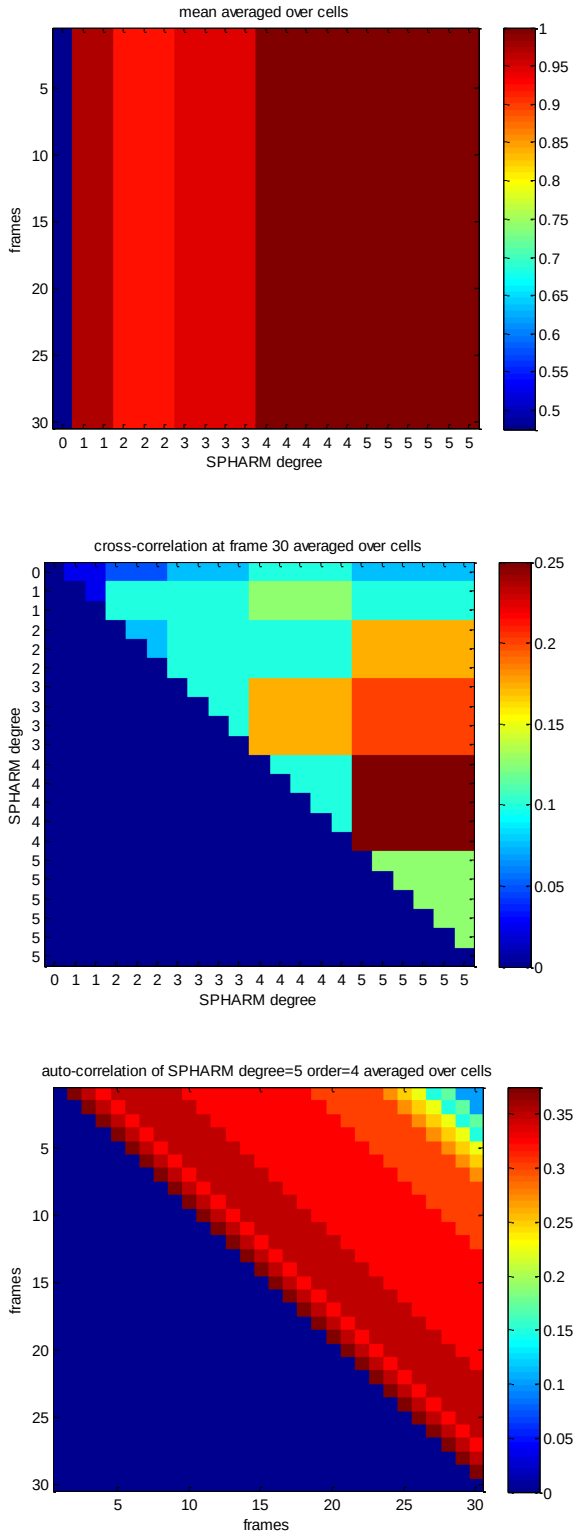

Figure 4.5: The selection frequency of inter-cell features by sparse SVM with group structured variable selection. For a given SPHARM degree l, the SPHARM order $m = 0, 1, ..., l$ are in ascending order from the left to the right and from the top to the bottom in the top and middle figures.

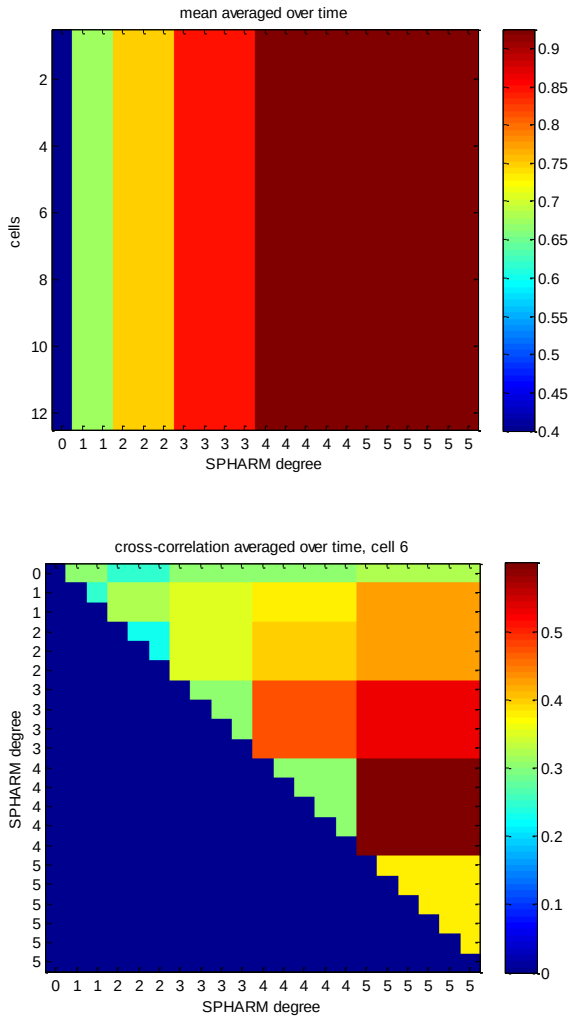

Figure 4.6: The selection frequency of intra-cell features by sparse SVM with group structured variable selection.

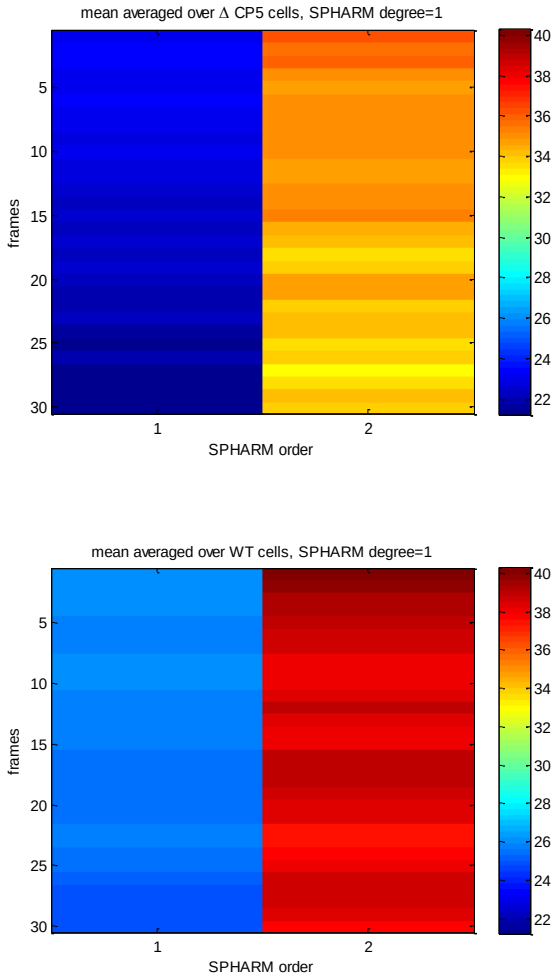

Figure 4.7: Heatmaps of the population features averaged over the ΔCP5 and WT populations independently. Notice that the ΔCP5 population has smaller $l = 1$ SPHARM coefficients than the WT population does.

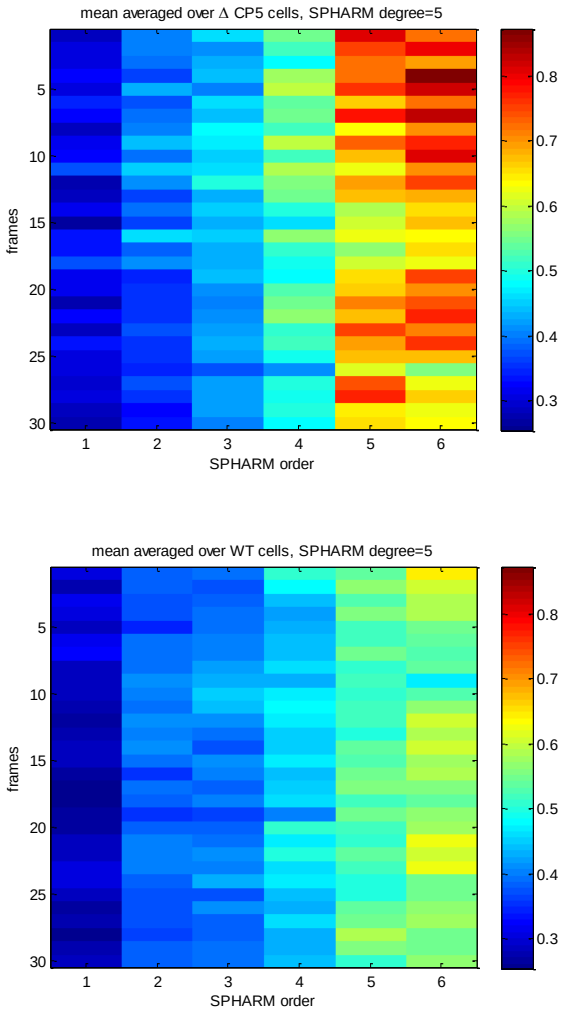

Figure 4.8: Heatmaps of the population features averaged over the ΔCP5 and WT populations independently. Notice that the ΔCP5 population has larger $l = 5$ SPHARM coefficients than the WT population does.

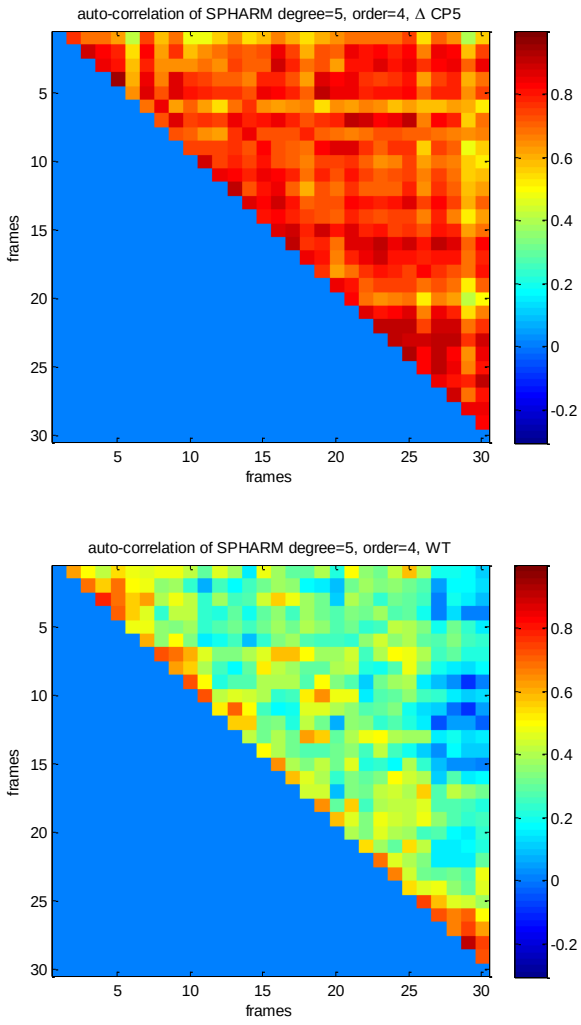

Figure 4.9: Heatmaps of the population features averaged over the ΔCP5 and WT populations independently. Notice that the ΔCP5 population presents larger correlation than the WT population, especially in the upper right corner in the heatmaps. This suggests that the de-correlation speed is an important feature for discriminate these two populations, which can be further understood in Fig. 4.10.

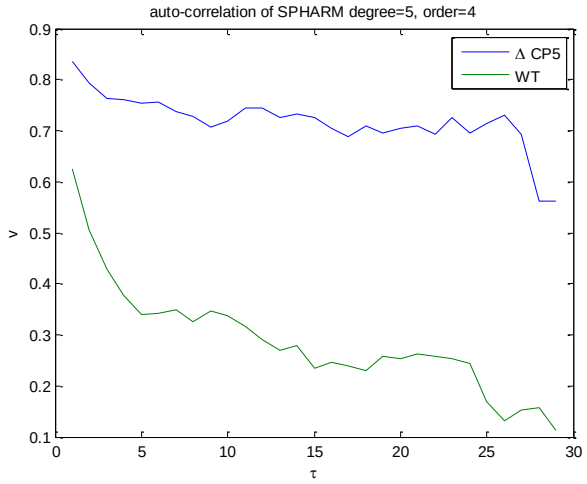

Figure 4.10: The WT population de-correlates faster than the ΔCP5 population.

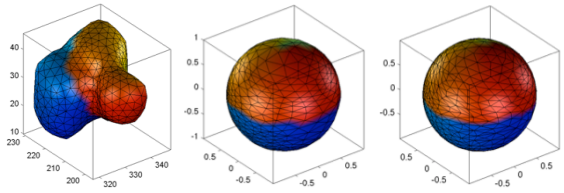

Figure 4.11: Spherical parametrization process. The original mesh (left) is first mapped to the sphere (middle), then the mapping is re-adjusted to minimize local and global distortions (right). The two colors represent the arbitrary North and South hemispheres of the mapped surface, and their correspondence on the original surface.

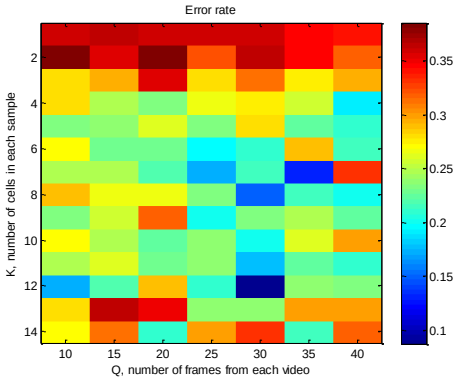

Figure 4.12: The heat map of error rate under the sparse SVM classifier, with the number of frames varying from 10 to 40, and the number of cells in each sample varying from 1 to 14. The best combination occurs at $Q = 30$ and $K = 12$.

CHAPTER V

Serially Sampled Multi-class Classification with Variable Selection: Application to Gene Expression Analysis

5.1 Introduction

In serially sampled multi-block multi-class classification, the task is to correctly predict the label of a subject based on a set of time samples. Examples include evaluation of a subject's positive or negative response to drug treatment, or classification of diseases based on gene microarray responses from multiple time points. In these applications, the common clinical test is to use a single test sample taken immediately prior to patient diagnosis. However, as gene expression genotyping becomes more prevalent in the era of personalized medicine, use of each patient's baseline reference sample can be expected to improve prediction performance. This chapter develops a method for including such a baseline reference sample in the predictor.

High-dimensional applications, such as genomics expression analysis and ECG classification, require parsimonious modeling. By pruning the total number of independent variables or features, variable selection is a first step in building parsimonious models. Accurate variable selection avoids the over fitting problem, and provides interpretations of the most relevant variables for a predictive model. It is important to

understand which variables are strongly relevant to the classification task, and how their importance depends on time or subject in the population. For example, sparsity penalized lasso techniques [4, 44] provide a computationally tractable way to perform variable selection driven by objective function minimization. Here we introduce an objective function minimization approach for variable selection that applies to more general problems than simple prediction and classification.

First to generalize the current methods in binary classification [43] to multi-class problems, we define a unified multi-class classifier. We adopt the Support Vector Machine (SVM) approach. The idea of maximizing the margin between two classes can be extended to multi-class problems. There are two common strategies that have been proposed: (1) solving the multi-class problem by a series of binary SVM classifiers [76, 77, 27]; (2) formulating a single unified multi-class SVM [78, 79, 80, 3, 81, 82]. The former approach has the advantage of building on the binary SVM framework; the latter is more direct. We propose a unified multiclass classifier with variable selection following the latter approach.

This chapter is organized as follows. In section 2, we formulate our problem by discussing loss function used as surrogates in classification, and the proper regularization that selects variables relevant simultaneously to all classes, and all references. In section 3, we propose a general algorithm to train the classifier. Two real data applications are presented in section 4 and 5, followed with discussion in section 6.

5.2 Multi-block Multi-class Classification

We have discussed the SVM in Chapter IV, which is designed for binary classification problems. Different methods have been proposed in the literature on the extension from binary to multi-class classification. A discussion can be found in [27, 28]. Two main strategies are (1) solving a series of binary SVM classifiers and (2) a unified multi-class SVM, which is referred to as the "single machine" approach

in [28].

Examples of the former strategy includes one-against-all, one-against-one, and averaged one-against-one. The one-against-all method trains a binary classifier for each class $k \in \{1, 2, ..., K\}$ versus all the other classes. The classification is made by selecting the class with the largest margin from the decision boundary. The one-against-one strategy trains one binary classifier for each pair of classes, and uses the C_2^K binary classifiers to vote for the class. The decision is made by majority vote, also known as "max wins" strategy [76, 77, 27]. The unified multi-class SVM depends on designing a generalized hinge loss function [78, 79, 80, 3, 81, 82]. It is not clear from the comparisons in [27, 28] which strategy is the best, since there is no single strategy that shows consistent improvement in terms of error rate performance relative to all the other strategies. The unified SVM involves a solution of more complex optimization problem, hence one may argue that training a series of binary classifiers is much more practical. However, the unified strategy has the advantage that interpretation of the features is simplified when using a sparse multi-class classifier. A unified sparse SVM solution can provide the features that are important for distinguishing between all the K classes.

Assume that we have a dataset of n samples from K populations observed under r conditions. A multi-block multi-class classifier, for example [3], can be trained over this data to give minimum classification error probability. However, when the feature dimension rp is large such a classifier will suffer from severe over-fitting error. To overcome this deficiency, sparsity-penalized classifiers have been developed [5], that "sparsify" the feature vector, i.e., finding a reduced number of features that attain the minimum cross-validated error. This is tantamount to selection of the most discriminating features. In this section we introduce a new sparsity-constrained multi-class classification algorithm for finding these features.

Let $\{(\mathbf{x_i}, y_i)\}_{i=1}^n$ be the set of data for classification, in which $\mathbf{x_i}$ is a rp dimensional

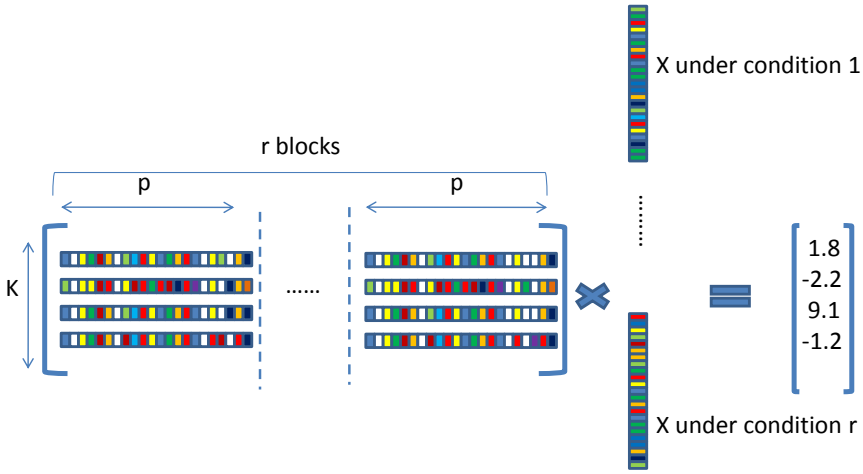

Figure 5.1: Classifying K different classes based on their temporal/spatial evolution is a multi-block classification problem. The colored column vector corresponds to a vector of features of different samples acquired over r blocks of time. The matrix containing four colored rows generates a different score (at right of equality) for each of the possible classes. The special multi-block structured sparsity (white colored entries) of this matrix minimizes overfitting.

vector, and $y_i \in \{1, 2, \ldots, K\}$ is the corresponding label. When $r = 1$, this reduces to the standard multi-class classification problem, whereas when $r > 1$, this is the multi-block multi-class classification problem. We introduce a general and unified way of solving such sparsity-penalized multi-class problems. The problem is to find rp-dimensional hyperplanes to partition the feature space,

$$F = \{f_1, f_2, \cdots, f_K\}, \text{ where } f_k(\mathbf{x}) = \mathbf{w_k}'\mathbf{x} + b_k$$

and the decision rule is to assign the label that gives the largest confidence, $\arg\max_k \{f_k(\mathbf{x})\}$, where f_k are smooth functions to be determined. This problem can be formulated as the optimization

$$\min_F \frac{1}{n} \sum_{i=1}^{n} V(F, \mathbf{x_i}) + \lambda R(F)$$

where V denotes some convex loss function that upper-bound the 0-1 loss, and R is a regularization function. The loss function should be a good surrogate for the non-convex 0-1loss, and the regularization function should provide capability of variable selection.

The first unified multi-class Support Vector Machine (SVM) was introduced in [3]. This work introduced a generalized notion of the margin for multi-class problems. They suggested solving the above optimization by using the representations

$$\begin{aligned} V(F, \mathbf{x_i}) &= \Big[\max_r (1 - \delta_{y_i,r} + f_r(\mathbf{x_i}) - f_{y_i}(\mathbf{x_i}))\Big]_+ \\ R(F) &= \frac{1}{2}\sum_{i=1}^{K} ||\mathbf{w_i}||_2^2. \end{aligned} \tag{5.1}$$

We adopt the same loss function for our classifier, but introduce another regularization function that accounts for group sparsity. The idea of introducing variable selection penalty originates from [83], in which Frank and Friedman introduced a generalization of ridge regression and subset selection through the addition of a penalty of the form $\lambda \sum_j ||\beta_j||_1{}^q$ to the residual sum of squares. By imagining the unit ball for different values of q, one can view the parameter q as the degree to which the variables is concentrated along the favored directions. For example, $q = \infty$ places maximum concentration along the diagonals, and $q = 0$ place the entire mass on the coordinate axes, which corresponds to L_0 subset selection. Tibshirani [84] further proposed the least absolute shrinkage and selection operator (*lasso*), corresponding to the case $q = 1$. Tibshirani's proposal has the advantage of being the penalty such that it is the closest to $q = 0$, the L_0 penalty, yet remaining convex. To generalize the method of variable selection for binary classification to multi-class, multi-block

classification ($r > 1$) , define W as follows:

$$W = \begin{bmatrix} \mathbf{w_1'} \\ \vdots \\ \mathbf{w_K'} \end{bmatrix} = \begin{bmatrix} \mathbf{w_{(1)}} \cdots \mathbf{w_{(rp)}} \end{bmatrix}$$

in which $\mathbf{w}_{(j)}$ is the j^{th} column in the matrix W, and $\mathbf{w}_k$ represents the k^{th} row. Notice that given $j \in \{1, 2, \ldots, p\}$, any elements in the i^{th} column of W such that mod $(i, p) = j$, are related to the same variable j, and all elements in the k^{th} row are related to scoring the confidence of the sample belonging to class k.

In multi-class classification problems ($r = 0$), one can ensure that the predictor variables are shared over all classes by forcing the columns of W to satisfy a coupled sparsity condition: the number of non-zero terms should be small. By imposing the coupled sparsity constraints, the complexity of the model is controlled by a few variables that are important for discriminating all the classes. Variable selection under this framework becomes more complicated than in binary classification, because one would expect that an unrelated variable corresponds to a zero column in W rather than a zero scalar. Wang and Shen extended the L_1 SVM to the L_1 MSVM by imposing a penalty with $q = 1$ on the coefficients. They solved a problem of the form [85]:

$$\min_{b,W} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i \neq k)[b_k + \mathbf{w_k'x_i} + 1]_+ + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} ||w_{kj}||_1.$$

Although the L_1 norm has the advantage of being directly related to *lasso*, it treats all the w_{kj}'s equally, which does not guarantee the variable sharing condition. Zhang et al accounted for variable sharing by treating the coefficients in groups by imposing a L_∞ penalty as follows[5].

$$\min_{b,W} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(y_i \neq k)[b_k + \mathbf{w_k'x_i} + 1]_+ + \lambda \sum_{j=1}^{p} ||\mathbf{w_{(j)}}||_\infty$$

To extend the standard multi-class classification to multi-block multi-class classification, we propose solving the optimization with

$$R(F) = \sum_{j=1}^{p} \left\| \tilde{\mathbf{w}}_{(j)} \right\|_2 \text{ with } \tilde{\mathbf{w}}_{(j)} = \begin{bmatrix} \mathbf{w}_{(j)} \\ \mathbf{w}_{(j+p)} \\ \vdots \\ \mathbf{w}_{j+(r-1)p} \end{bmatrix}. \tag{5.2}$$

This ensures that coefficients corresponding to a shared variable over the measurements under r conditions and over all classes are grouped appropriately. The L_2 norm instead of the L_∞ norm is chosen in our formulation, because the L_∞ ball tends to favor solution with the scaled version of hadamard matrix, whereas L_2 norm penalizes any direction uniformly.

We can also extend the adaptive lasso [86, 87] to our multi-block multi-class formulation to reduce the over-estimation issue. Suppose the initial estimation is F_{init}, we can form a new optimization problem that depends on F_{init}.

$$\min_F \frac{1}{n} \sum_{i=1}^{n} V(F, \mathbf{x}_i) + \lambda_{adapt} R_{adapt}(F, F_{init}) \tag{5.3}$$

in which V is defined as the same generalized hinge loss function and

$$R_{adapt}(F, F_{init}) = \sum_{j=1}^{p} \frac{||\tilde{\mathbf{w}}_{(j)}||_2}{||\tilde{\mathbf{w}}_{init,(j)}||_2}.$$

An interesting property about adaptive lasso is that if the coefficients in the initial estimation is 0, then the new estimation will also be 0. In other words, adaptive lasso is a stage-wise screening process. The process can be repeat for multiple stages.

5.3 Algorithmic Implementation

To solve our optimization problem with the combination of loss function (5.1) and regularization (5.2), we apply the augmented Lagrangian method discussed in Chapter III to the optimization.

We modify the optimization to form an equivalent problem, in which the newly introduced variable M is constrained such that $M = W$, and the row vectors $\mathbf{m_k}$ of M obey the same structural pattern as the rows of W:

$$\min_{W,M} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} ||\tilde{\mathbf{m}}_{(j)}||_2$$

$$subject\ to: \forall i,\ k \quad (\mathbf{w'_{y_i}} \mathbf{x_i} + b_{y_i}) + \delta_{y_i,k} - (\mathbf{w'_k} \mathbf{x_i} + b_k) \geq 1 - \xi_i,\ M = W$$

The slack variables ξ_i from the generalized hinge loss function depend on W by the constraints, and the regularization function penalizes the mixed L_1/L_2 norm of M. By alternating splitting method we have an algorithm for the multi-block multi-class classification problem (Algorithm 5).

Algorithm 5: Multi-block Multi-class Classification

set $\tau = 0$, choose $\mu > 0$, M_0, W_0, D_0
while *stopping criterion is not satisfied* **do**
  $W_{\tau+1} = \arg\min_{W} \frac{1}{n} \sum_{i=1}^{n} \xi_i + \frac{\mu}{2} ||W - M_\tau - D_\tau||_F^2$
  s.t. $\forall i, k \ \ (\mathbf{w'_{y_i}} \mathbf{x_i} + b_{y_i}) + \delta_{y_i,k} - (\mathbf{w'_k} \mathbf{x_i} + b_k) \geq 1 - \xi_i$
  $M_{\tau+1} = \arg\min_{M} \lambda \sum_{j=1}^{p} ||\tilde{\mathbf{m}}_{(j)}||_2 + \frac{\mu}{2} ||W_{\tau+1} - M - D_\tau||_F^2$
  $D_{\tau+1} = D_\tau - W_{\tau+1} + M_{\tau+1}$
  $\tau = \tau + 1$

In each iteration, W can be solved by subgradient method. This enables us to solve any convex loss function besides the generalized hinge loss we adopted, if its subgradient exists. The subgradient method [88] is a simple iterative algorithm for minimizing non-differentiable convex function. Suppose f is convex. To minimize f,

at each iteration, we take a step in the direction of a negative subgradient: $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$, where $x^{(k+1)}$ is the kth iterate, $g^{(k)}$ is any subgradient of f at $x^{(k)}$. Similar approaches based on subgradient methods can be found in [89]. Since the objective function is exactly a quadratic programming problem and there exits fast algorithm tailored to SVM classification problems, we adopt the sequential dual method [90, 61].

The dual of line 3 in Algorithm 5 can be written as

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \tfrac{1}{2} \sum_{k=1}^{K} ||\mathbf{w}_k||_2^2 + \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_i^k e_i^k \\ &\text{s.t.} \sum_{k=1}^{K} \alpha_i^k = 0,\ \alpha_i^k \leq \tfrac{1}{n\mu} \delta_{y_i,k}\ \forall i, k \end{aligned}$$

in which $\mathbf{w}_k = \sum_{i=1}^{n} \alpha_i^k \mathbf{x}_i + \mathbf{m}_{\tau,k} + \mathbf{d}_{\tau,k}$ and $e_i^k = 1 - \delta_{y_i,k}$. Coordinate descent method can be extended to decompose the dual problem into n subproblems, and each problem corresponds to one of the n samples.

$$\begin{aligned} &\min_{\alpha_i^1,\ldots,\alpha_i^K} \sum_{k=1}^{K} \tfrac{1}{2} A(\alpha_i^k)^2 + B_k \alpha_i^k \\ &\text{s.t.} \sum_{k=1}^{K} \alpha_i^k = 0,\ \alpha_i^k \leq \tfrac{1}{n\mu} \delta_{y_i,k} \forall k \end{aligned}$$

where $A = \mathbf{x}_i \mathbf{x}_i'$ and $B = \mathbf{w}_k' \mathbf{x}_i + e_i^k - A\alpha_i^k$. This is the same optimization problem discussed in [90, 61] except the representation of $\mathbf{w}_k$. We can adopt the subproblem solver based on coordinate descent method.

In the second step, M has a close form solution. Let $C = W_{\tau+1} - D_\tau$, then the solution of each concatenated column of M is given as $\tilde{\mathbf{m}}_{(j)} = [||\tilde{\mathbf{c}}_{(j)}||_2 - \frac{\lambda}{\mu}]_+ \frac{\tilde{\mathbf{c}}_{(j)}}{||\tilde{\mathbf{c}}_{(j)}||_2}$, [62].

Given that we can solve the multi-block multi-class classification with group structured sparsity, we can also find the solution for adaptive lasso formulation. We can

reformulate problem (5.3) as the initial estimation problem [87]. Define

$$x_{new,l,i} = x_l ||\tilde{\mathbf{w}}_{init,(\bmod (l,p))}||_2,\ l = 1, .., rp,\ i = 1, ..., n$$
$$\tilde{w}_{new,(j)} = \frac{\tilde{\mathbf{w}}_{(j)}}{||\tilde{\mathbf{w}}_{init,(j)}||_2},\ j = 1, ..., p$$

then the adaptive lasso for multi-block multi-class classification can be formulated as

$$\min_{F_{new}} \frac{1}{n} \sum_{i=1}^{n} V(F_{new}, \mathbf{x}_{new,i}) + \lambda_{adapt} R(F_{new}).$$

The algorithm for multi-block multi-class classification can be applied to this problem.

5.4 Simulation Experiments

We implement two simulation models in [5], a five-class example and a four-class example. The five-class example has independent variables with dimension p, and the first two variables are generated according to $N(\boldsymbol{\mu}_k, \sigma_1^2 I_2)$, where

$$\boldsymbol{\mu}_k = 2(\cos([2k-1]\pi/5), \sin([2k-1]\pi/5)),\ k = 1, 2, 3, 4, 5.$$

The remaining $p-2$ variables are generated independently from $N(0, \sigma_2^2)$, and $\sigma_1 = \sqrt{2}$, $\sigma_2 = 1$. 250 samples are generated evenly from the model for training, another 250 samples for tunning the regularization parameter, and 50,000 samples for the test set. We compare the proposed multi-class classification with L_1/L_2 mixed norm penalty for group structured variable selection with the unified multi-class classifier in [3]. We also test the proposed algorithm with prescreening each pairwise binary classifications, i.e., the input variables to the multi-class classifier are the union of the variables selected by any pairwise binary classification. The entire experiment is repeated for 20 trials. We find that when $p < n$, the unified multi-class classifier without variable selection performs the best. However, when $p > n$, the multi-class

method	error rate	number of var.	CZ	IZ
the ideal classifier	0	2	998	0
1. unified linear SVM	0.61	1000	0	0
2. sparse unified SVM	0.57	47	952.5	0.5
3. sparse unified SVM, prescreen	0.50	13.95	985.8	0.25

p-values of one sided paired t-test:

method	ER	number of var.	CZ	IZ
(1,2)	$1 > 2$: 0.1438	$1 > 2$: 7.46×10^{-19}	$1 < 2$: 7.31×10^{-19}	$1 < 2$: 0.0105
(2,3)	$2 > 3$: 0.0399	$2 > 3$: 0.0783	$2 < 3$: 0.0771	$2 > 3$: 0.096
(1,3)	$1 > 3$: 2.42×10^{-4}	$1 > 3$: 2.02×10^{-31}	$1 < 3$: 1.87×10^{-31}	$1 < 3$: 0.0481

Table 5.1: Simulation model 1, five-class example, with $p = 1000$. We compare the performance of the unified multi-class SVM [3], the proposed multi-class classification with group structured variable selection, and the proposed method with prescreening. CZ: number of correct zeros in the multi-class classifier, IZ: number of incorrect zeros in the classifier.

classifier with structured variable selection outperforms the non-sparse classifiers. The results for $p = 1000$ are listed in Table 5.1. We also change the signal to noise ratio by modifying

$$\boldsymbol{\mu}_k = 4(\cos([2k-1]\pi/5), \sin([2k-1]\pi/5)),\ k = 1, 2, 3, 4, 5.$$

The performance for all the methods improve, and the sparse unified SVM works better than the nonsparse SVM. However, the sparse SVM with prescreening technique performs worse than the one without prescreening. The results are shown in Table 5.2.

The second simulation model was designed for generating variables that are important for some of the classes but not all the classes [5]. This four-class example has independent variables with dimension p, in which the first four variables are generated from $Unif[-1, 1]$, and the rest $p - 4$ of them from $N(0, 8^2)$. 4 linear functions are

method	error rate	number of var.	CZ	IZ
the ideal classifier	0	2	998	0
1. unified linear SVM	0.3303	1000	0	0
2. sparse unified SVM	0.0988	2.05	997.95	0
3. sparse unified SVM, prescreen	0.1217	9.60	990.40	0

p-values of one sided paired t-test:

method	ER	number of var.	CZ	IZ
(1,2)	$1 > 2$: 7.26×10^{-31}	$1 > 2$: 2.52×10^{-71}	$1 < 2$: 2.52×10^{-71}	$1 \neq 2$: 1
(2,3)	$2 < 3$: 3.39×10^{-6}	$2 < 3$: 0.0475	$2 > 3$: 0.0475	$2 \neq 3$: 1
(1,3)	$1 > 3$: 6.88×10^{-23}	$1 > 3$: 2.21×10^{-33}	$1 < 3$: 2.21×10^{-33}	$1 \neq 3$: 1

Table 5.2: Simulation model 1, five-class example, with $p = 1000$. We compare the performance of the unified multi-class SVM [3], the proposed multi-class classification with group structured variable selection, and the proposed method with prescreening. CZ: number of correct zeros in the multi-class classifier, IZ: number of incorrect zeros in the classifier.

generated for defining the class labels.

$$f_1 = -5x_1 + 5x_4$$
$$f_2 = 5x_1 + 5x_2$$
$$f_3 = -5x_2 + 5x_3$$
$$f_4 = -5x_3 - 5x_4$$

The class labels are generated according to multinomial distribution

$$p_k(\mathbf{x}) = P(Y = k|X = \mathbf{x}) \propto \exp(f_k(\mathbf{x})),\ k = 1, 2, 3, 4.$$

200 samples are generated evenly from the model for training, another 200 samples for tunning the regularization parameter, and 40,000 samples for the test set. Similar to the simulations in model 1, we compare 3 different methods with $p = 1000$. Results are listed in Table 5.3. For larger signal to noise ratio experiment, we generate the first four variables from $Unif[-20, 20]$. The performance is summarized in Table 5.4.

In simulation model 1, the informative variables are important for all the five

method	error rate	number of var.	CZ	IZ
the ideal classifier	0	4	996	0
1. unified linear SVM	0.7489	1000	0	0
2. sparse unified SVM	0.7495	200.5	795.5	4
3. sparse unified SVM, prescreen	0.7495	242.8	753.2	4

p-values of one sided paired t-test:

method	ER	number of var.	CZ	IZ
(1,2)	$1 < 2$: 0.2333	$1 > 2$: 6.60×10^{-12}	$1 < 2$: 7.2×10^{-12}	$1 < 2$: 0
(2,3)	$2 > 3$: 0.3249	$2 < 3$: 0.1747	$2 > 3$: 0.1747	$2 \neq 3$: 1
(1,3)	$1 < 3$: 0.1636	$1 > 3$: 2.34×10^{-17}	$1 < 3$: 2.58×10^{-17}	$1 < 3$: 0

Table 5.3: Simulation model 2, four-class example, with $p = 1000$. We compare the performance of the unified multi-class SVM [3], the proposed multi-class classification with group structured variable selection, and the proposed method with prescreening. CZ: number of correct zeros in the multi-class classifier, IZ: number of incorrect zeros in the classifier.

method	error rate	number of var.	CZ	IZ
the ideal classifier	0	4	996	0
1. unified linear SVM	0.4609	1000	0	0
2. sparse unified SVM	0.0756	28.4	971.6	0
3. sparse unified SVM, prescreen	0.0733	27.65	972.35	0

p-values of one sided paired t-test:

method	ER	number of var.	CZ	IZ
(1,2)	$1 > 2$: 9.42×10^{-27}	$1 > 2$: 4.76×10^{-25}	$1 < 2$: 4.76×10^{-25}	$1 \neq 2$: 1
(2,3)	$2 > 3$: 0.2081	$2 > 3$: 0.3198	$2 < 3$: 0.3198	$2 \neq 3$: 1
(1,3)	$1 > 3$: 1.17×10^{-26}	$1 > 3$: 8.64×10^{-32}	$1 < 3$: 8.64×10^{-32}	$1 \neq 3$: 1

Table 5.4: Simulation model 2, four-class example, with $p = 1000$. We compare the performance of the unified multi-class SVM [3], the proposed multi-class classification with group structured variable selection, and the proposed method with prescreening. CZ: number of correct zeros in the multi-class classifier, IZ: number of incorrect zeros in the classifier.

classes, and as dimension increases, the structured sparsity becomes important. When p is large, the unified SVM without sparsity constraints suffers from overfitting problems. Notice that prescreening the variables by pairwise binary classifications helps to reduce the error rate, when the noise level is relatively large. The second simulation model is designed so that the informative variables are important to some of the classes but not all. Therefore, the structured sparsity constraints that penalizes the coefficients across classes may not perform as good as the standard unified SVM without sparsity. However, the number of variables is greatly reduced with the sparse classifiers. When the relative noise level decreases, the prescreening technique for sparse SVM marginally outperforms the one without prescreening in this simulation model. In summary, the sparse SVM outperforms the nonsparse SVM when the dimension is large, and the informative variables are important for all the classes. The prescreening helps when the noise level is high. When the model does not satisfy the group sparsity structure, it depends on the noise level to decide whether to apply the group sparsity structure and prescreening.

5.5 Application: Learning Differential Gene Expression Signatures from Personalized High Throughput Screening

Personalized medicine has become a prominent research direction as longitudinal genomic information becomes available. For example, longitudinal gene expression were collected and analyzed in [91, 92], in which the authors discussed the systems biology of vaccination for yellow fever and seasonal influenza. Temporal dynamics of host molecular responses were studied to differentiate symptomatic and asymptomatic health status [93]. In a recently published paper by Chen et al. [94], the authors have demonstrated the ability of personal omics profile to reavel dynamic molecular and medical phynotypes by monitoring a single individual over 14 months. As the di-

mension of the information available increases, for example the large number of genes present, and the integration of genomic, trasncriptomic, proteomic, metabolomic, and autoantibody profiles, the interpretation and accuracy of the model become new challenges to classic statistical methods. Moreover, intersubject variations can play an important role in the fitted model when the number of subjects is limited while each provides an information-rich profile. The former suggests the need of variable selection, and the latter points to the necessity of new techniques to eliminate irrelevant patient variations.

This section treats an important problem of differential analysis of high dimensional data generated from a serially sampled population undergoing two or more treatments. Specifically, for a population of m subjects let y_i denote the label of the i-th subject, e.g., treatment (viral inoculation type) or outcome (symptom severity), and let $\mathbf{x}_i$ denote a set of molecular samples, r control samples taken at baseline time points, and a test sample taken at some posterior time. We propose an algorithm (section 5.3) for learning the best classifier of the label y_i given the data $\mathbf{x}_i$ in the high-dimensional case where the number m of subjects is much less than the number p of variables, e.g., gene probes on the microarray. This is a non-standard learning problem due to the fact that the classifier is a function of both a baseline reference sample and a test sample. Our results show significant gains in classifier accuracy as compared to standard multi-class classification methods that do not use a personalized reference sample.

5.5.1 Background

We conduct analysis on two H3N2 challenge studies. The H3N2 D2 challenge study in 2008 consisted of 17 pre-screened volunteers without recent influenza-like illness in the preceding 45 days. These subjects had samples taken 12 hours prior to inoculation of A/Wisconsin/67/2005 (H3N2) and immediately prior to inoculation.

Peripheral blood was taken at baseline, then at 8 hour intervals for the initial 120 hours and then 24 hours for the remaining 2 days of the study, as shown in Fig 5.2. The H3N2 D5 study in 2010 had 22 pre-screened volunteers. Subjects had a single reference (30 hours prior to inoculation of A/Wisconsin/67/2005 (H3N2)) and peripheral blood was taken at baseline, then at 8 hour intervals for the initial 170 hours and again at 680 hours.

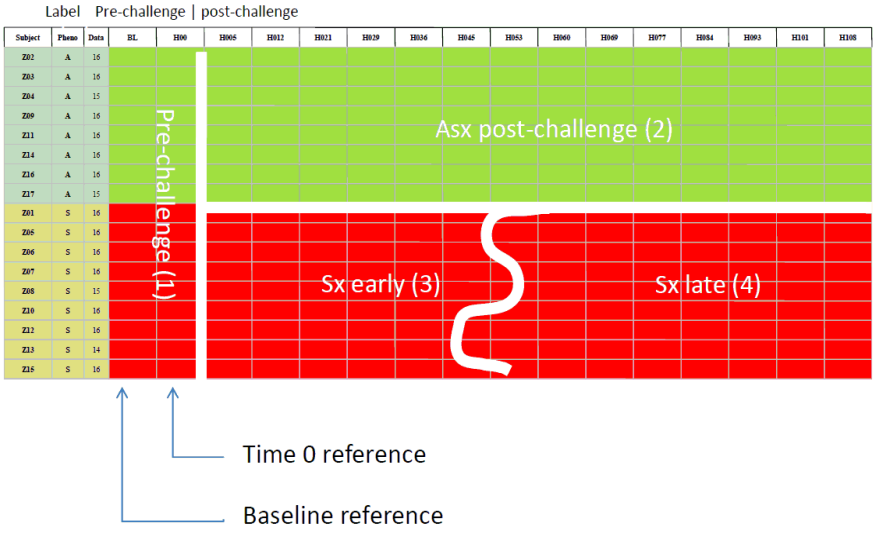

Figure 5.2: H3N2 D2 challenge study

After the inoculation, subjects who never become infected are labeled uninfected post-challenge. Chips obtained from subjects who eventually became infected are further divided into pre-infection class, acute-infection class and post-infection class. Viruses can not reproduce themselves outside a host cell, instead they assemble within the infected cells. The viral replication in respiratory infection is known to occur soon after exposure at the site of infection. Hence, viral titer measurement is a key indicator for infection. In the challenge studies, viral titers from daily nasopharyngeal washes were used as corroborative evidence of successful infection [93]. The onset time and offset time are determined by the virus titer measurement for defining the boundaries between pre-infection and acute-infection, acute-infection and post-infection classes. Figure 5.3 shows the titration measurement for both challenge studies. The D2 study

had measurement every 24 hours, whereas the D5 study measured twice a day. Given the clinical labels of infected and uninfected subjects, we set the onset time for uninfected subjects as the time point at which there was detectable virus titration, and the offset time as the the time point at which no more virus titration is above the detection level. Peripheral blood samples acquired before virus inoculation are labeled as pre-challenge (Class1) and after inoculation, measurements from the uninfected subjects are labeled as uninfected (Class 2). Any peripheral blood sample acquired before the onset time from infected subjects as labeled as pre-infection (Class 3), between onset and offset time points are the acute-infection samples (Class 4), and after the offset time are the post-infection samples (Class 5). Additional exploration by symptom scores and BLU [93] analysis to define the classes are shown in Figure 5.4 and 5.5 respectively. We focus on the classification problems with classes defined by either viral titer measurement or symptom scores, because the BLU analysis was performed on the gene expression which may lead to overfitting. Besides, the analysis procedure of viral titer takes a longer time than the gene expression microarray analysis does, which motivates the use of titer measurement as the class labels.

It is of great interest if we can classify the different regions after inoculation, especially with feature selection to find related biomarkers. The immune system consists of innate immune system and adaptive immune system. It will be useful if we can understand the uninfected and infected subjects in terms of gene expression to discover gene function associated with immune properties. For example, innate immune system is non-specific response, and there is no immunological memory, whereas adaptive immune system usually has a time lag between exposure and response. These properties should be related the virus titration measurements and the symptoms. The reference-based classification problem we are interested in this study is to classify a subject's sample by comparing it with his/her baseline samples or time 0 samples. The hypothesis is that significant gains in infection prediction performance are

methods	reference	error rate	number of selected genes
logistic SVM, one v.s. one	w/o	0.39	12023
logistic SVM, one v.s. one	w/	0.45	12023
linear SVM, one v.s. one	w/o	0.43	12023
linear SVM, one v.s. one	w/	0.47	12023
unified linear SVM	w/o	0.34	12023
unified linear SVM	w/	0.38	12023

Table 5.5: Performance of H3N2 challenge study by classic methods: classes defined by titration. The inclusion of reference increases the data dimension, and these classic methods without sparsity regularization suffers more from overfitting problems as dimension increases.

possible using serial samples.

5.5.2 Approach

The model is to classify a sample into one of the three post-inoculation regions based on virus titration measurements: the uninfected, the pre-infection and the post-infections. We first explore the task by classic methods in Table 5.5 by analyzing data from H3N2 D2 and D5 studies. The samples right before inoculation in H3N2 D2 and the ones obtained 30 hours prior to inoculation in H3N2 D5 are treated as the reference samples. One subject is left out as the test set, and the rest as the training set. The parameters for all the methods are selected by 2-fold cross validation, and the samples are grouped by subjects, i.e., samples from the same subject should exist in the same sets for cross validation. The experiment is conducted over each subject as a test set, and repeated for several times to report the average performance. Given the high-dimensional data, $p = 12023$, and limited number of samples or subjects, it is easy to overfit the model. For example, we compare each method under two cases, the one without the baseline or time 0 references and the one with the references. Since all these methods are without sparsity constraints, the ones with reference suffer more than the ones without references from overfitting problems as expected.

Multi-block Group Structures To solve the overfitting problem,we apply the multi-block multi-class classifier discussed in section 5.2. If no reference is compared in the model, it is the case when $r = 1$ and when the model takes into account the reference chips, it corresponds to $r = 2$. We adopt the multi-block sparsity structure because it is of interest to identify a small number of biomarkers that can be implemented in a device that measures the temporal evolution. Adaptive lasso regularization is applied to this application.

Stratified Sampling The classification may suffer from the fact that the data are imbalanced. The number of infected and uninfected subjects are about the same. However, class 2 (uninfected region) has a lot more samples than the sub-region in the infected subjects, i.e., the pre-infection, acute-infection and post-infection classes. We down sample the majority classes by limiting the number of samples per class per subject to be around 7, which is the averaged number of chips per subject in class 4.

Prescreening by Pairwise Classifications We further test the multi-block multi-class classifier by restricting the classifiers to use the biomarkers selected in pairwise problems. This is represented as the prescreening method in Figure 5.6. The overall flow for each subproblem remains the same, leaving one subject out for the test set, and performing 2-fold cross-validation to decide the regularization parameters. The difference is that we input the selected variables found in 2 v.s. 3, 3 v.s. 4 and 2 v.s. 4 binary classifications to the multi-class classifier and restrict the classifier to optimize over these subset of variables.

5.5.3 Results

The performance of the sparse multi-class classifier is presented in Table 5.6. Different combinations of classes are considered in the experiments. For example, pairwise comparison of classes 2 v.s. 3, 2 v.s. 4, 3 v.s. 4, and 4 v.s. 5, as well

as the 2,3,4 multi-class classification. We notice that the performance greatly improved when the reference is taken into account in pairwise 2 v.s. 3 classification, which is classifying the uninfected subjects from the pre-infection subjects. This is generally a hard problem, since the inflammatory responses are not yet developed in region 3. Our results on the multi-class problem of class 2,3,4 has better performance than the classic approaches. This demonstrates the importance of the sparsity constraints. The results of the prescreened multi-class multi-block classification show a 30% performance gain from the one without reference, and a 36% improvement from the multi-class with reference but without prescreening. We also present the performance of classification when the reference is included, but treated as concatenated data without corresponding group structure between the reference and the target, i.e., a single block data with $r = 1$. In most of the classification tasks, the single block classifiers with reference perform worse than the multi-block classifiers, or include more biomarkers than the multi-block classifiers. To understand the difficulty and improvement of the multi-block multi-class classification, we plot the error rate of each chip as a heatmap in Figure 5.7. The heatmaps show that the improvements mostly arise in regrion 3, which is a consistent results if we compare it to the 2 v.s. 3 binary classification improvement.

The reason that the reference helps to improve the accuracy may be better illustrated by examining the solution matrix of the multi-block classifier. Take the 2,3,4 classification problem with prescreening as an example. We take the average of the multi-block multi-class classifier solution matrix W and plot bar charts for the genes that have been selected with 100% frequency in Figure 5.8. The bars labeled as R represent the average of the coefficients $w_R = w_{j,k}$, in which R stands for the reference chip; the ones labeled as T represent the average of the coefficients $w_T = w_{j+rp,k}$, in which T stands for the target chip, and $k \in \{1, 2, 3\}$. It is interesting to notice that some of the genes serve as normalizing biomarkers, those are the ones that have

the same sign in w_R and w_T, whereas some of them serve as contrasting biomarkers, in which the signs in w_R and w_T are opposite to each other. And the contrasting biomarkers are mostly the inflammatory genes.

Another interesting observation is that the number of selected genes increases in binary 2 v.s. 3, 2 v.s. 4 classifications and 2,3,4 multi-class problems when the reference is considered in the model, while the number decreases in binary 3 v.s. 4 and 4 v.s. 5 classifications when the reference is included. The classification tasks in the former category are classifying across different subjects, i.e., no subjects has samples in both classes simultaneously. The classifier with reference recruits a few more genes to reduce the control the irrelevant patient variations. The tasks in the latter category are to distinguish sub-classes in the subjects. The classifiers with references can do the discrimination at a similar level of accuracy but with significantly fewer biomarkers.

We list the genes with high selection frequency in Table 5.7, 5.8 and 5.9. Notice that a lot of interferon genes related to the inflammatory responses are selected, in addition to genes associated with T cell apoptosis. In order to find early biomarkers we arbitrarily shifted the boundaries between pre-infection and acute-infection by 2 chips, corresponding to 16 hours, earlier and reran the analysis of 3 v.s. 4 and multi-class 2,3,4 classifications. To emphasize the boundary, larger weights are applied to the boundary and the weights linearly decreases in time. The focus is more on feature selection, and we list the biomarkers in the same tables. We notice a few genes with sharper and less homogeneous expression in the acute-infection region are selected as we apply larger weights on the boundary, such as STAT1, IRF9, TAP1 and OAS3, see Figure 5.10.

We also apply the same comparisons on the classes defined by symptom scores. The results are in Table 5.10 and figure 5.9. The performance is not as good as the results based on virus titers. One of the reason may be that the symptom scores are self reported measurements, which can be noisy. However, the inclusion of reference

classification task	reference	error rate	number of genes
1. class 2,3	w/o	0.26	51.79
2. class 2,3	w/ , r=2	0.17	52.69
3. class 2,3	w/ , r=1	0.22	199.76
4. class 2,4	w/o	0.11	20.39
5. class 2,4	w/ , r=2	0.12	28.04
6. class 2,4	w/ , r=1	0.13	314.53
7. class 3,4	w/o	0.18	39.67
8. class 3,4	w/ , r=2	0.16	28.25
9. class 3,4	w/ , r=1	0.16	248.92
10. class 4,5	w/o	0.22	46.89
11. class 4,5	w/ , r=2	0.23	33.72
12. class 4,5	w/ , r=1	0.21	255.19
13. class 2,3,4	w/o	0.28	52.53
14. class 2,3,4	w/ , r=2	0.29	71.16
15. class 2,3,4	w/ , r=1	0.28	394.08
16. class 2,3,4 prescreen	w/o	0.26	118.76
17. class 2,3,4 prescreen	w/ , r=2	0.19	119.84
18. class 2,3,4 prescreen	w/ , r=1	0.25	331.22

p-values of one sided paired t-test:

method	error rate	number of genes
(17,16)	$17 < 16$: 8.40×10^{-4}	$17 > 16$: 0.4634
(17,18)	$17 < 18$: 9.98×10^{-8}	$17 < 18$: 8.30×10^{-18}

Table 5.6: Performance of H3N2 challenge study: classes defined by titration. The classifications with reference and r=1: the reference is taken into account, but not treated as multi-block data, i.e., no corresponding group structure between the corresponding variables in the reference chip and the target chip. The classifications with reference and r=2: the reference is included, and data is treated as a two-block classification problem. These results show that: (1) imposing multi-block sparse structure on the classifier indeed gives superior performance with respect to single block structure; and (2) the difference in performance is statistically significant as determined by the paired t-test.

2,3	2,3	2,4	2,4	4,5	4,5
w/o ref	w/ ref	w/o ref	w/ ref	w/o ref	w/ ref
TFG	IKZF1	IFI27	LY6E	SLC31A1	XAF1
IGLV3-25	ZNF43	'IFI44	IFI27	XAF1	C11orf75
DSP	HSPBAP1	IFI44L	PRMT2	PLAC8	IFI6
C4BPA	STAG3	XAF1	HERC6	C11orf75	PSMB9
115648_at	HRASLS3	HERC6	IFIT1	DPEP2	MYD88
HLA-DQA1	HMG2L1	IFIT1	HLA-DQA1	IFI35	SLC31A1
IKZF1	MSLN	SIGLEC1	SCO2	SCO2	IFI35
TLK1	TNFRSF14	IRF7	IFI44	RRM2	EIF2AK2
MYOM2	LILRA3	'MS4A4A	RRM1	PARP12	RRM2
TUBB6	GFOD2	BLVRA	NFATC3	EIF2AK2	DPEP2
ABCA7	PSTPIP1	ISG15	MS4A4A	M97935_3_at	IRF7
GM2A	LYST	SCO2	IKZF1	'IRF7	PARP12
LILRB2	VPS13D		IFI44L	STAT1	UBE2L6
PRSS21	RRM1		IRF7	UBE2L6	M97935_3_at
PF4V1	ZNF135		OAS1		STAT1
	LILRB2		ISG15		
	BTN3A2				
	HLA-DQA1				
	PCDHB11				
	SETD1A				
	DCTD				
	RHCE				
	GM2A				
	TUBB6				
	ZNF646				
	FKBP11				
	ARSA				
	NFATC3				
	PF4V1				
	C4BPA				
	RSRC1				
	UBE2W				
	NUBP2				
	PSMB1				
	MYOM2				
	PRSS21				
	ANK3				
	APOL3				

Table 5.7: List of genes with selection frequency $\geq 60\%$

3,4	3,4	3,4 W	3,4 W	3,4 W,E	3,4 W,E
w/o ref	w/ ref	w/o ref	w/ ref	w/o ref	w/ ref
IRF7	XAF1	KYNU	PMEPA1	ATP9A	SCO2
XAF1	EMR3	BTN2A2	GRPEL1	DHRS1	IFI44
GBP1	SERPING1	GBP1	KYNU	HOOK2	TUBB2A
OAS1	SCO2	LY6E	SYNGR3	DAAM2	OAS3
ICAM3	ICAM3	IFI44L	XAF1	XAF1	LOC442257
HLA-DPA1	OAS1	HLA-DPA1	SCO2	ALDH9A1	DAAM2
EMR3	IFI44	IFNAR1	IFNAR1	LOC285412	REC8
SERPING1	IFI44L	ISG15	GBP1	TRAF3	HAL
IFI44	IFI27	SCO2	ICAM3	C4orf27	STAT1
IFI44L	LY6E	IFI27	ZNF516	CYP21A2	COIL
IFI27	ISG15	KIAA0157	HLA-DPA1	SCO2	ISG15
LY6E			OAS1	NEK1	ADAMTS5
ISG15			IFI44	COIL	PCK2
SCO2			ISG15	TUBB2A	TGM3
			IFI44L	LOC442257	FXYD1
			IFI27	REC8	ZNF343
			LY6E	TAP1	IRF9
				HAL	PLA2G10
				ADAMTS5	IFI44L
				FXYD1	
				IRF9	
				ZNF343	
				ISG15	
				TGM3	
				STAT1	
				IFI44L	
				PCK2	
				PLA2G10	

Table 5.8: List of genes with selection frequency $\geq 60\%$: W refers to the larger weights being applied to the boundary between pre-infection (class 3) and acute-infection (class 4); E refers to the shifting of that boundary 16 hours earlier.

2,3,4	2,3,4	2,3,4 W	2,3,4 W	2,3,4 W,E	2,3,4 W,E	2,3,4 prescreen	2,3,4 prescreen
w/o ref	w/ ref	w/o ref	w/ ref	w/o ref	w/ ref	w/o ref	w/ ref
LILRB2	IKZF1	LILRB2	NUBP2	LILRB2	NUBP2	IRF7	NUBP2
IFI27	PLAC8	IKZF1	MSLN	IKZF1	IKZF1	MS4A4A	IKZF1
SCO2	MS4A4A	ABCA7	IKZF1	ABCA7	IRF9	PF4V1	PRSS21
IRF7	VPS13D	CLEC10A	PRSS21	CLEC10A	PRSS21	SCO2	IFI44L
MS4A4A	IRF7	PRSS21	IFI44L	UTS2	IFI44L	BLVRA	ANK3
PF4V1	APOL3	CKB	BTN3A2	PRSS21	BTN3A2	SIGLEC1	IRF7
IKZF1	NUBP2	115648_at	IGLV3-25	ADAMTS5	SCAP	ISG15	NFATC3
RRAS	ANK3	LPAR1	ANK3	CKB	C20orf91	IFIT1	OAS1
TUBA4A	IFI27	GPM6A	HLA-DQA1	115648_at	ANK3	LILRB2	RRM1
PRSS21	ZNF43	IGLV3-25	APBA2	DSP	IFI27	PRSS21	APOL3
SIGLEC1	HLA-DQA1	HLA-DQA1	IFI27	ALDH9A1	IFI35	XAF1	ISG15
APOBEC3B	NFATC3	HSPA1B	LY6E	IGLV3-25	NFATC3	HERC6	IFI44
CKB	SETD1A	IFI27	NFATC3	HLA-DQA1	OAS1	GM2A	MS4A4A
HLA-DQA1	MYOM2	IFNAR1	OAS1	IFI27	LAP3	ABCA7	PSMB1
	SCO2	HCG26	MS4A4A	MYL5	FKBP11	IKZF1	MYOM2
	ISG15	LY6E	FZR1	PF4V1	RSRC1	IFI44	RSRC1
	HMG2L1	MICA	PDZK1	PMM2	MS4A4A	TUBB6	UBE2W
	PRSS21	CD200	PF4V1	XAF1	PF4V1	MYOM2	C4BPA
	LAMP3	PLAC8	CEP27	BLVRA	FXYD1	IFI44L	PF4V1
	LY6E	PF4V1	VPS13D	C4BPA	XAF1	IFI27	SCO2
		LIN37	C4BPA	TUBB2A	C4orf27		IFIT1
		ERO1LB	ZNF43	TUBB6	VPS13D		HLA-DQA1
		C4BPA	HSPBAP1	IL18RAP	RRM1		ARSA
		TUBA4A	APOL3	MYOM2	C4BPA		FKBP11
		HSPBAP1	PSTPIP1	FADS2	TUBB2A		ZNF646
		PAQR6	MYOM2	APOBEC3B	ZNF43		LY6E
		TUBB6	SH3BP5	ISG15	ZNF204		IFI27
		IL18RAP	APOBEC3B	TLK1	HSPBAP1		
		MYOM2	ISG15	SCO2	APOL3		
		APOBEC3B	SETD1A		IFT88		
		ISG15			MYOM2		
		ARHGEF10			APOBEC3B		
		TLK1			ISG15		
		SCO2			SETD1A		
					SCO2		

Table 5.9: List of genes with selection frequency $\geq 80\%$: W refers to the larger weights being applied to the boundary between pre-infection (class 3) and acute-infection (class 4); E refers to the shifting of that boundary 16 hours earlier

classification task	reference	error rate	number of genes
class 2,3	w/o	0.36	39.78
class 2,3	w/ , r=2	0.33	43.93
class 2,4	w/o	0.47	31.19
class 2,4	w/ , r=2	0.54	34.55
class 3,4	w/o	0.24	60.29
class 3,4	w/ , r=2	0.24	41.39
class 4,5	w/o	0.25	49.56
class 4,5	w/ , r=2	0.24	33.98
class 2,3,4	w/o	0.50	58.39
class 2,3,4	w/ , r=2	0.43	72.95
class 2,3,4 prescreen	w/o	0.47	116.80
class 2,3,4 prescreen	w/ , r=2	0.44	145.40

Table 5.10: Performance of H3N2 challenge study: classes defined by symptom scores. The performance is not as good as the one with classes defined by virus titer measurements, but the inclusion of reference improves the performance.

leads to improvement in most of the classification tasks.

5.5.4 Discussion

We apply the reference based classification to the gene microarray data of a challenge study where serial samples were acquired from a population of subjects inoculated with live (H3N2) flu virus. While the methodology is generally applicable to many computational bioinformatics problems, we focus on a specific problem in personalized medicine: pre-infection detection and prediction of disease outcome from a serial assay over time of the person's gene expression profile. The reference-based classification problem in this application is a special case of our multi-block multi-class classification. The references of each subject have been concatenated with the testing samples to provide the history of the subject. The experimental results showed significant improvement of the additional references. By quantitative comparision of a person's current expression profile to that observed at previous times a more accurate health assessment can be made. The group sparsity penalty greatly reduces the

number of variables and selects the important ones for the classification task. This is particularly useful in the large p (dimension) small m (number of samples) problems, and becomes more important when multiple references are taken into account, which greatly increases the dimension.

The proposed method may have a lot of other applications, when the decision should be made based on the current status of subject but also the baseline status. The selected features may provide interesting interpretation, for example, in this experiment, the selected immune genes that are closely related to the infection and inflammatory process. This application develops a new framework for learning a classifier from a population of personalized serial samples. The power of this framework is that the incorporation of the reference sample into the classifier significantly improves classifier accuracy. The framework specifically accounts for small population size, high dimensional classifier variables, and increased complexity of the classifier that operates on personalized serial samples including a reference sample.

5.6 Conclusion

In reference based classification we discussed, we predict the label by using serially or spatially diversified samples. By using such a fixed reference, irrelevant patient variations can be controlled to enhance our ability to evaluate positive or negative response to drug treatment, or to classify a disease based on clinical-molecular from multiple time points or multiple tissues. As the data dimension increases, variable selection becomes increasingly important in these problems. This is especially the case for the serially sampled reference-based classification problem, as variable dimensions increase linearly in the number of references. Hence it is important to understand which variables are strongly relevant to the classification task, and how they evolve over temporally or spatially different samples.

Variable selection in multi-block multi-class problems is more challenging than

in binary classification. We show that parameter estimation with shrinkage can be cast as a problem of structured variable selection, where the structure is specified by the classes and blocks defining the sampling patterns. We have formulated the optimization to solve the multi-class support vector machine with a mixed L_1 and L_2 norm penalty. The sparsity penalty enables group variable selection over multiple data blocks and classes. A convex optimization method is developed to solve for the optimal classifier function and select the relevant variables simultaneously. This optimization is implemented by variable splitting and augmented Lagrangian methods. We apply our algorithm to predictive health problems. The results show that the addition of sample reference of gene microarray under normal conditions greatly improves the classification by gene expression to predict the health status. Our method is able to greatly reduce the number of features and pick out immune genes that mediate the response to viral pathogens and are predictive of severe symptomatic illness. We presented the necessity of structured variable selection in personalized high throughput screening problems. The health status can be better predicted with interpretable features by our proposed method.

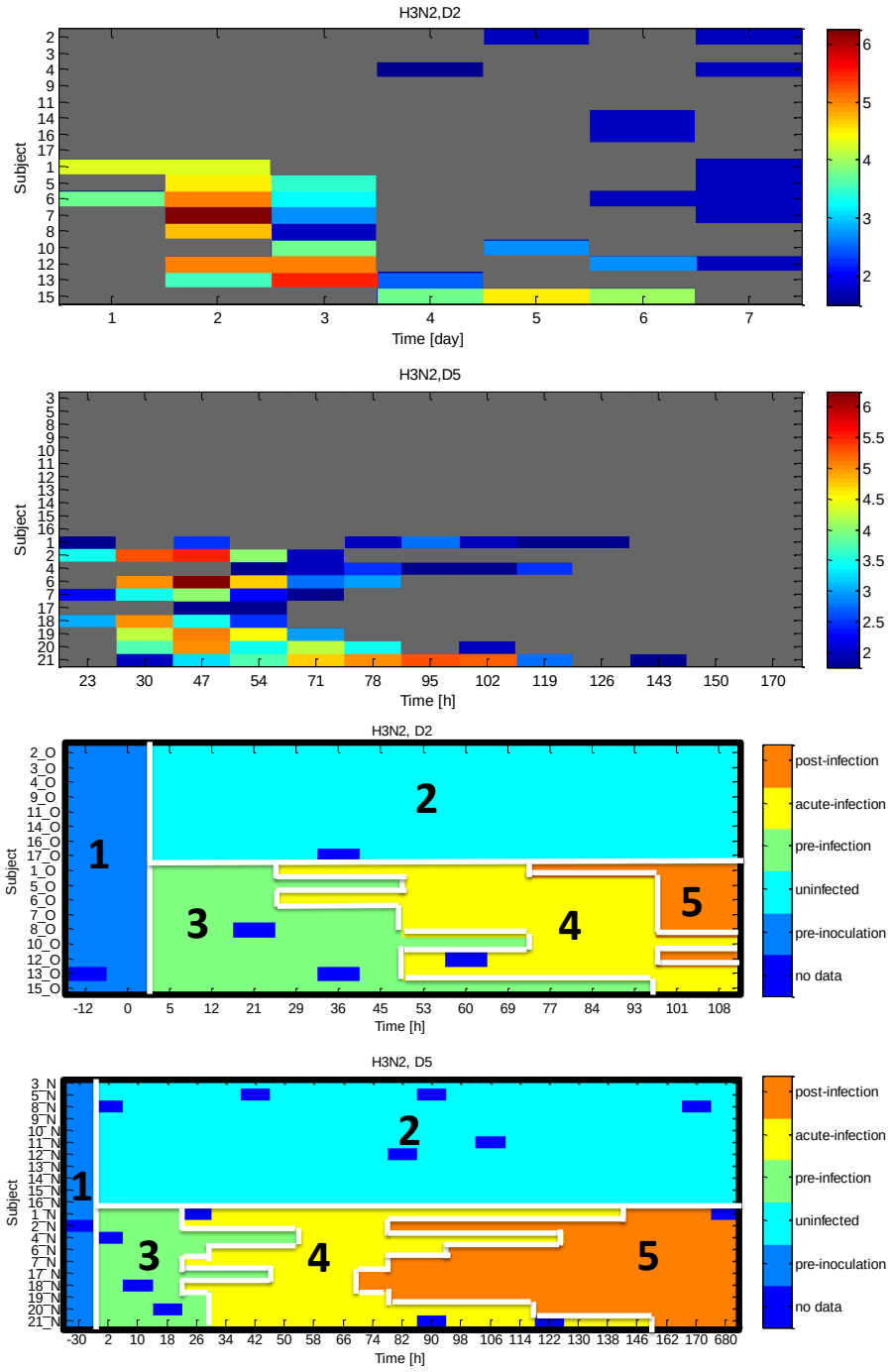

Figure 5.3: H3N2 study titration measurements and the classes. Top two figures show the titration measurements. Bottom two figures show the classes defined by these measurements. The onset and offset time of detectable titration are used to set the boundaries between class 3 and 4 and class 4 and 5 respectively.

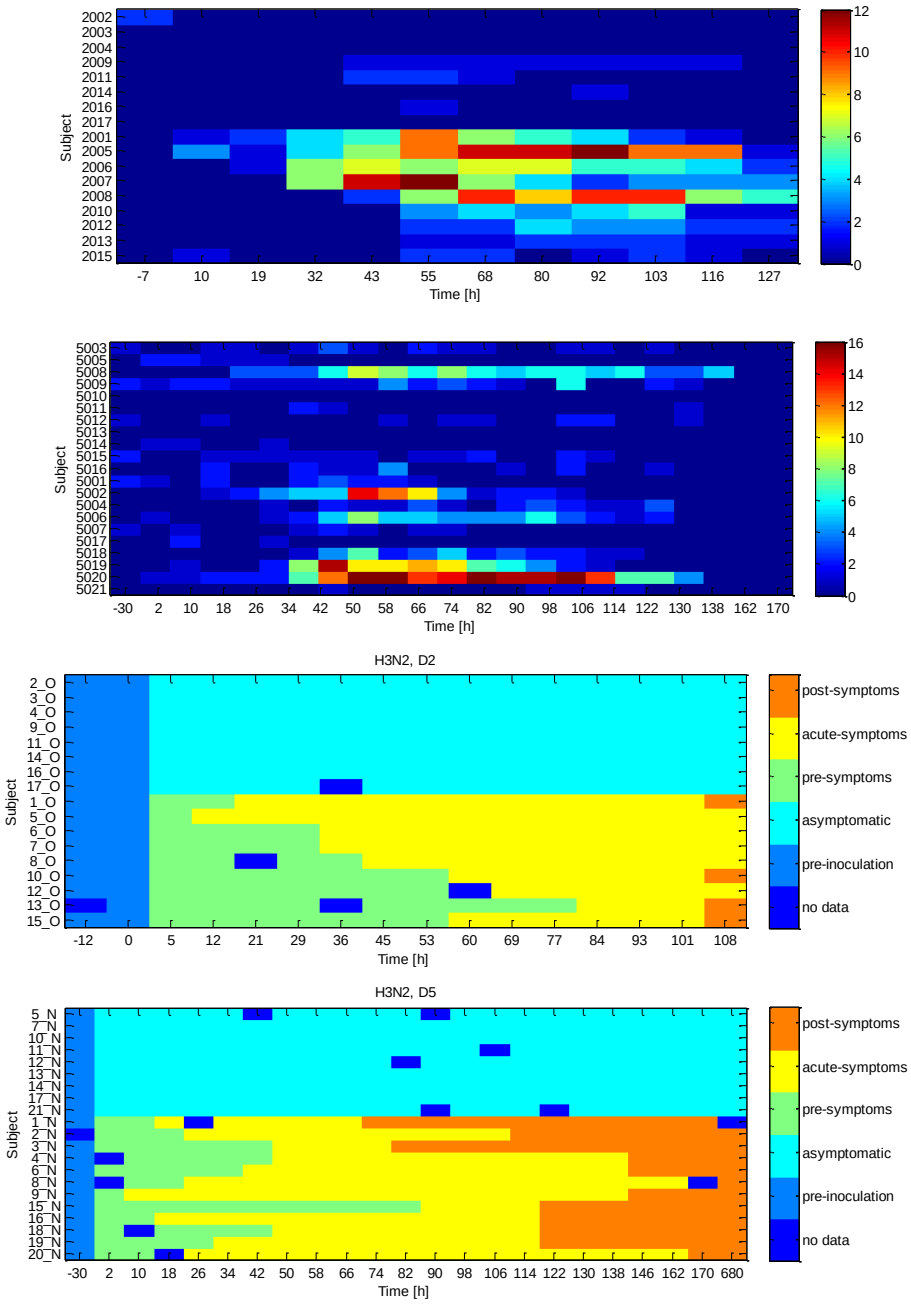

Figure 5.4: H3N2 study symptom scores and the classes. Top two figures show the sum of symptom score measurements. Bottom two figures show the classes defined by these measurements. The onset and offset time of detectable titration are used to set the boundaries between class 3 and 4 and class 4 and 5 respectively.

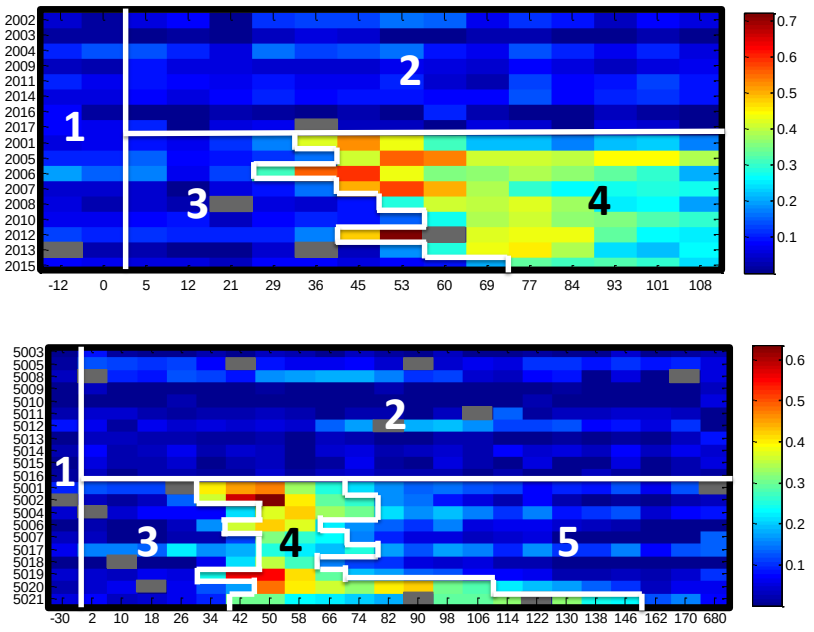

Figure 5.5: BLU analysis on the H3N2 challenge study.

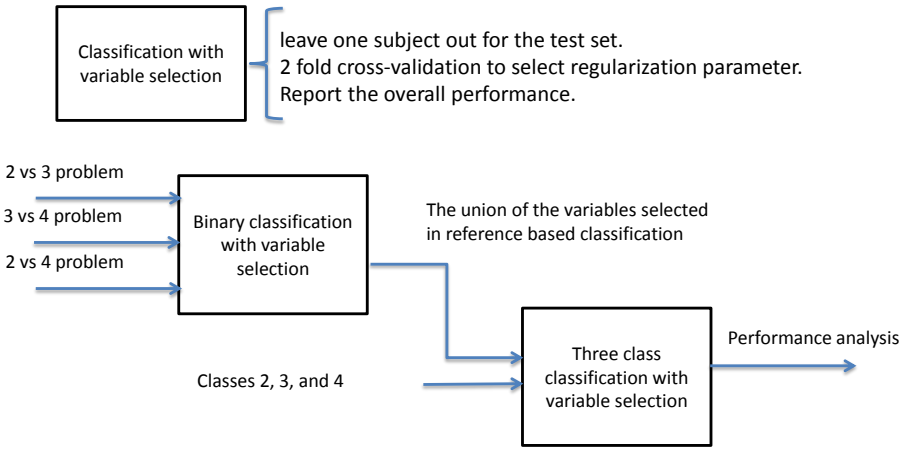

Figure 5.6: Presreening Flow chart for classifying class 2,3, and 4.

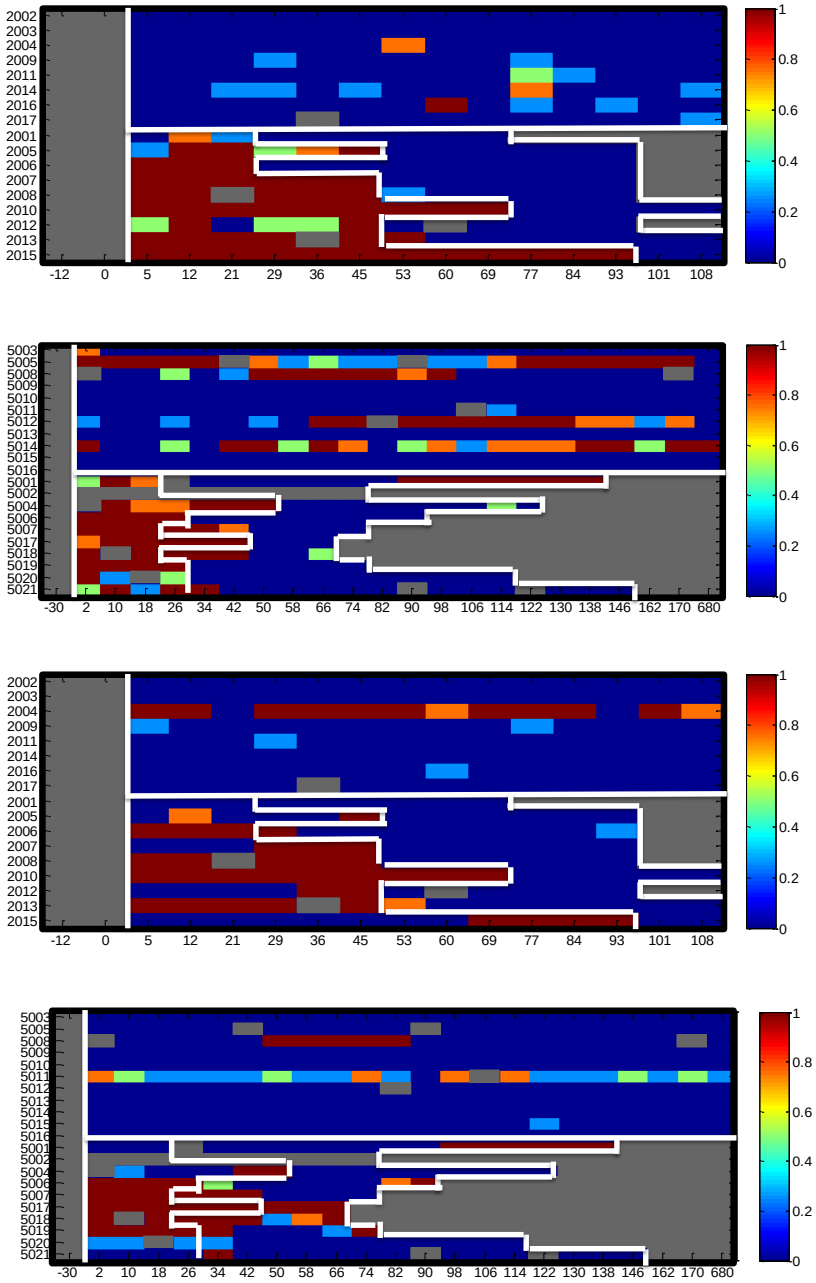

Figure 5.7: The heatmaps of the error rate classifying class 2,3, and 4. The top 2 figures show the error rate of classifying without the reference, whereas the bottom 2 figures show the results when the reference chips are included. The performance in class 3 (pre-infection) improves when the reference is provided. The classes are defined by virus titers.

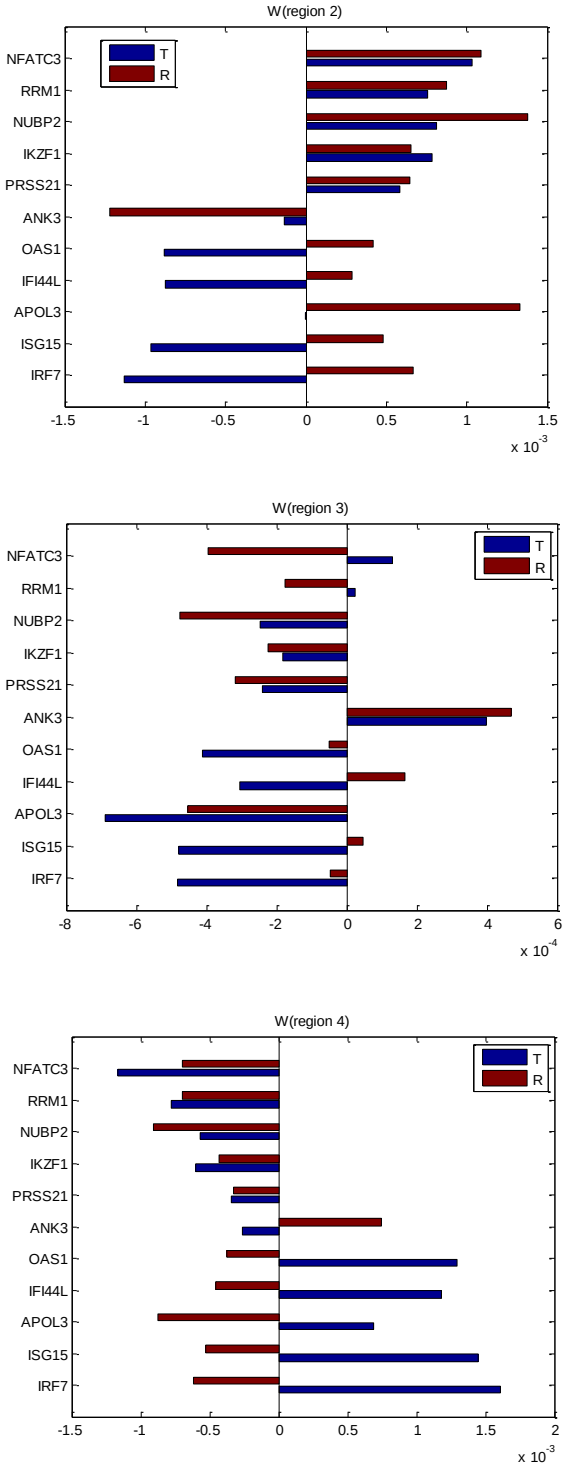

Figure 5.8: The solution matrix of the multi-class classifier.

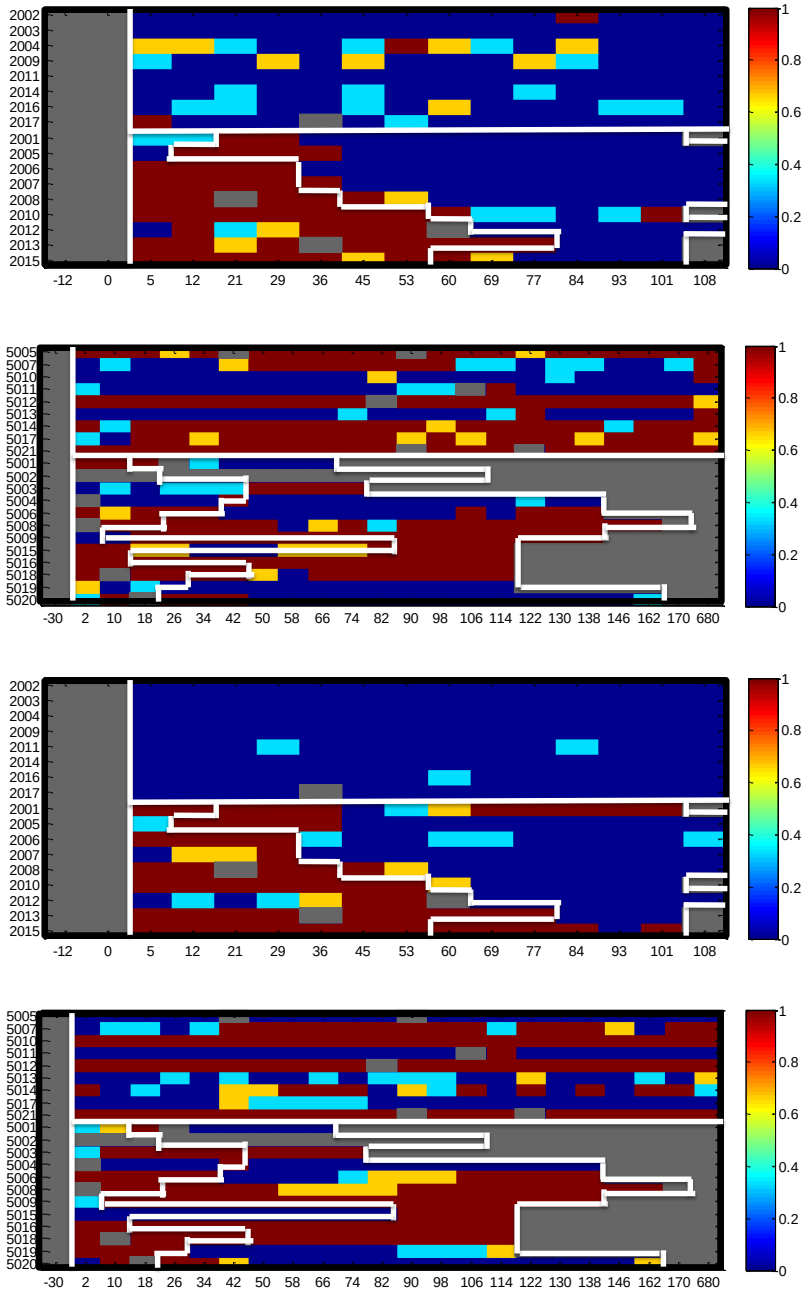

Figure 5.9: The heatmaps of the error rate classifying class 2,3, and 4. The top 2 figures show the error rate of classifying without the reference, whereas the bottom 2 figures show the results when the reference chips are included. The classes are defined by symptom scores.

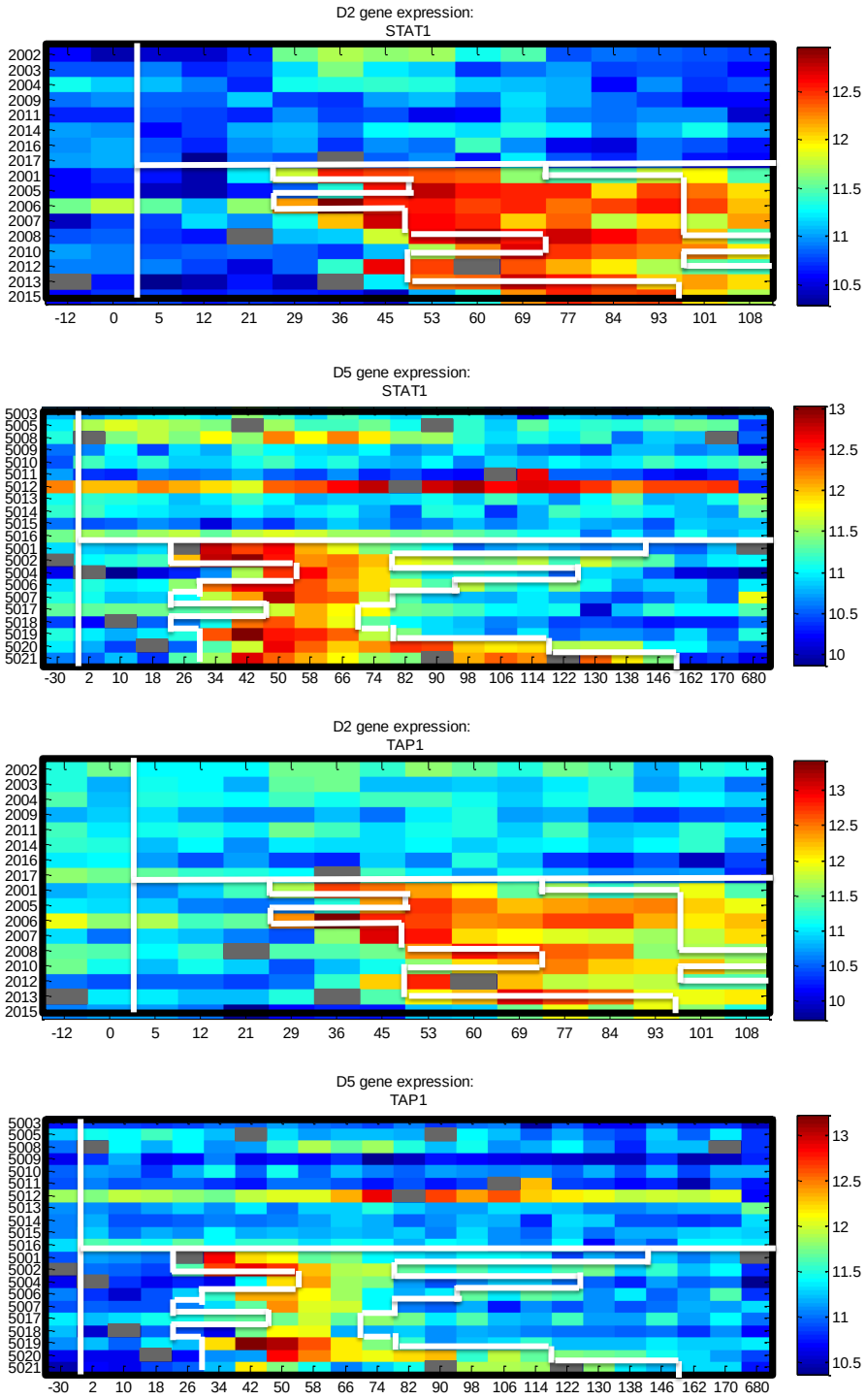

Figure 5.10: The heatmaps of the gene expression: The top 2 figures are the heatmaps of STAT1 and the bottom 2 figures are heatmaps of TAP1. These are examples of the genes selected when larger weights are applied to the boundary between the pre-infection and acute-infection classes.

CHAPTER VI

Uneven Margin SVMs for Imbalanced Training Samples

6.1 Introduction

Classification with imbalanced datasets is an important problem in which one or more of the classes has many fewer training samples than the other classes. For example, this arises when the samples from the class of interest are outnumbered by those of the majority class. At the data preprocessing level, over-sampling the minority class, under-sampling the majority class, or the combination of these two methods have been proposed to solve the imbalance problem [95]. At the algorithmic level, uneven cost and uneven margin classifiers have been studied and shown to improve the performance [10, 11]. In this chapter we focus on the algorithmic level approaches. Specifically we evaluate a category of algorithms, called α-calibrated classifiers, for training binary classifiers from imbalanced samples [19]. Our experiments show that the α-calibrated classifiers with an additional uneven margin parameter performs the best.

6.2 Classification with Imbalanced Data

One of the challenges in machine learning is learning to classify from from data that have few samples from one of the classes: the so-called imbalanced training problem. The traditional statistical learning methods are usually designed for data that are well-balanced between the classes. However, there are many of real world problems that do not fit into the balanced framework. For example, in data mining for direct marketing, banks and insurance companies have large database of customers, and they would like to discover patterns of buyers. The number of customers in the database who have already bought the product is usually small compared to the entire database [12]. In clinical applications, diagnosis may require detection of anomalies, which are by definition rare events. An example is automated detection of intestinal contractions in endoscopic video images, in which the prevalence of contractions is very low, yielding to highly skewed training sets [13]. Text classification is another well-known problem that requires learning from imbalanced data [10], in which the task is assigning documents to appropriate categories. .

There are two common strategies to solve the imbalanced classification problem that have been proposed in the current literature. One is a data level strategy that randomly resamples the data in each class to equalize the number of samples followed by training a standard classifier on this balanced sample. Another is an algorithm-level strategy that uses the imbalanced sample with an adjusted classifier loss function.

6.2.1 Data-level Resampling Strategy

Data level methods are based on resampling one of the classes to produce a balanced sample which can be used to train a classifier using a standard loss function. It can be viewed as the preprocessing step. Resampling methods can be divided into two categories: (1) of over-sampling the minority class, or (2) under-sampling the majority class [14]. One can also consider the combination of over-sampling and

under-sampling, performing over-sampling the minority and under-sampling the majority simultaneously [15].

6.2.2 Algorithmic-level Loss Calibration

Algorithmic level methods use the entire sample but adjust the decision threshold or the loss function to account for imbalance. Examples are: adjusting the margin of the Support Vector Machine(SVM)[16][10][11], applying different weights to the two classes in logistic regression [17], developing cost-sensitive SVM [18]. An advantage of these methods is that they do not require extra manipulation or preprocessing of the data. These methods can usually be implemented using off-the-shelf machine learning tools by adjusting the algorithms.

The impact of these adjustments on classifier performance is not well-studied. In the following sections, we examine the consistency of imbalance-compensated SVM. As the decision boundary of SVM only depends on the support vectors, one might expect it to suffer less from imbalanced data than other classification techniques. This is one of the reasons that SVM has been a popular technique for imbalanced learning.

6.3 Calibrated Surrogate Losses

Suppose we have a binary classification problem with feature data $X \in R^p$ and label data $Y \in \{-1, 1\}$. The task is to learn from the training data $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ a prediction function $f : X \to R$, where the decision on the value of Y is given by $sign(f)$. The prediction error rates are the average number of false positives (FP) $\frac{1}{2n}\sum_{y_i=1}(y_i - sign(f(\mathbf{x}_i))$ and the average number of false negatives (FN) $\frac{1}{2n}\sum_{y_i=-1}(sign(f(\mathbf{x}_i) - y_i)$. In the classical learning problem, the prediction function is chosen to minimize the average prediction error rate. However, in many applications, the cost of FPs and FNs are not necessarily the same. We follow the approach of cost-sensitive classification risk defined in [19]. Let $\alpha \in (0, 1)$ be a cost parameter.

The α-risk is defined as

$$R_\alpha(f) = E_{X,Y}[(1-\alpha)1_{\{y=1\}}1_{\{f(x)\leq 0\}} + \alpha 1_{\{y=-1\}}1_{\{f(x)>0\}}].$$

Let $\eta(x) = P(Y=1|X=x)$. The conditional risk is

$$C(\eta(x), f) = (1-\alpha)\eta(x)1_{\{f(x)\leq 0\}} + \alpha(1-\eta(x))1_{\{f(x)>0\}} \tag{6.1}$$

for which the optimal decision rule becomes [19]

$$sign\left(\frac{\eta(x)(1-\alpha)}{(1-\eta(x))\alpha} - 1\right) = sign(\eta(x) - \alpha)$$

which is also known as the maximum a posteriori probability (MAP) estimate. In even-cost classification, $\alpha = 0.5$, and the optimal decision rule reduces to the standard MAP classifier: $\arg\max_y\{\mathrm{P}(y|X=x)\}$.

In practice, as it is discontinuous, the indicator function on $f(x)$ (6.1) is replaced with a a smoother surrogate loss function L. Then the average loss, or risk, can be which can be further decomposed into average partials losses L_1 and L_{-1}.

$$R_L(f) = E_{X,Y}[1_{\{y=1\}}L_1(f(x)) + 1_{\{y=-1\}}L_{-1}(f(x))]$$

Given properties of L, such as convexity, and or smoothness, minimizing R_L is usually easier than minimizing the average loss (6.1) [19].

The conditional L-risk is defined as

$$C_L(\eta, t) = \eta L_1(t) + (1-\eta)L_{-1}(t),$$

the optimal L-risk is $C_L^*(\eta) = \inf_{t\in R} C_L(\eta, t)$ and the smallest L-risk with $\eta < \alpha$ is $C_{L,\alpha}^-(\eta) = \inf_{t\in R: t(\eta-\alpha)\leq 0} C_L(\eta, t)$. These quantities will be important when we examine

the consistent property of the algorithms.

Definition VI.1. The loss L is α-classification calibrated if $C^-_{L,\alpha}(\eta) - C^*_L(\eta) > 0$ for all $\eta \in [0,1], \eta \neq \alpha$.

In other words, for all x such that $\eta(x) \neq \alpha$, the value $f(x)$ minimizing the conditional L-risk will have the same sign as the optimal predictor $\eta(x) - \alpha$. For an α-classification calibrated L-risk, there exists an invertible, nondecreasing function $\psi_{L,\alpha}$, satisfying $\psi_{L,\alpha}(0) = 0$, such that

$$\psi_{L,\alpha}(R_\alpha(f) - R^*_\alpha) \leq R_L(f) - R^*_L.$$

This surrogate regret bound guarantees that if an algorithm is consistent for the L-risk, it will also be consistent for the α cost-sensitive classification risk.

As mentioned in section 6.2.2, adjusting parameters in each class is a popular solution to imbalanced learning. Uneven cost and uneven margin losses are widely used. They can be represented as in general form by

$$L_\alpha(y,t) = (1-\alpha)1_{\{y=1\}}\phi(t) + \alpha 1_{\{y=-1\}}\beta\phi(-\gamma t) \tag{6.2}$$

where $\phi : R \to [0,\infty)$ is convex and $\beta, \gamma > 0$. By[19](Theorem 7), the condition that a loss function be α-classification calibrated can be simplified for losses that are convex and differentiable at 0. The conditions become $\beta = \frac{1}{\gamma}$ and $\phi'(0) < 0$ for the loss (6.2). In Section 6.4, we compare several uneven cost, uneven margin losses that are popular in imbalanced learning, and show that the ones satisfying α-classification calibrated conditions result in better classification performance.

6.4 Simulation Experiments

We compare different popular loss functions in imbalanced learning. The standard SVM hinge loss, the hinge loss with the cost parameter tuned, the hinge loss with the cost parameter set as the pre-defined α, and lastly the hinge loss with the cost parameter set as α but with the margin parameter tuned. The losses are listed below in the same order, in which $\tilde{\alpha}, \rho \in (0, 1)$.

$$
\begin{aligned}
L_1(y,t) &= 1_{\{y=1\}}(1-t)_+ + 1_{\{y=-1\}}(1+t)_+ \\
L_2(y,t) &= (1-\tilde{\alpha})1_{\{y=1\}}(1-t)_+ + \tilde{\alpha}1_{\{y=-1\}}(1+t)_+ \\
L_3(y,t) &= (1-\alpha)1_{\{y=1\}}(1-t)_+ + \alpha 1_{\{y=-1\}}(1+t)_+ \\
L_4(y,t) &= \tfrac{(1-\alpha)}{2(1-\rho)}1_{\{y=1\}}(1-2(1-\rho)t)_+ + \tfrac{\alpha}{2\rho}1_{\{y=-1\}}(1+2\rho t)_+
\end{aligned}
$$

By the discussion in Section 6.3 calibrated surrogate losses, we know that L_1 is not α-classification calibrated (α-CC) if $\alpha \neq \frac{1}{2}$; L_2 is α-CC only if the tuning parameter is $\tilde{\alpha}$ set to α; L_3 and L_4 are both α-CC, and L_4 has an additional tuning parameter. The current algorithms for solving standard SVM or the kernelized SVM can be modified easily to solve these problems. The details on the modification can be found in the Appendix.

To understand the effect of tuning the parameters, we first set up an experiment, testing each loss function on two Gaussian distributed toy examples in Fig. 6.1 and Fig. 6.2 . The predefined cost parameter (listed as ALPHA in the figures) is set as 0.95, $Pr(Y = 1) = 0.9$. The regularization parameter λ is chosen over a grid $logspace(-5, 5, 50)$ and the tuning $\tilde{\alpha}$ and ρ are searched over $[0.01 : 0.01 : 0.99]$.

In the first experiment shown in Fig. 6.1, the data is generated according to $\mu_+ = [\sqrt{2}, -\sqrt{2}]$, $\mu_- = [-\sqrt{2}, +\sqrt{2}]$, $\Sigma_+ = \Sigma_- = I$. The optimal decision boundary is denoted as MAP in the figure, which is the solid black line. The Classifier trained by the loss function L_2 is the green line, whereas the one trained by loss function L_3 is

the dotted black line, and the one by loss function L_4 is represented as the dotted blue line. Similarly, in the second experiment, Fig. 6.2, the data is synthesized according to $\mu_+ = [\sqrt{2}, -\sqrt{2}], \mu_- = [-\sqrt{2}, +\sqrt{2}], \Sigma_+ = I, \Sigma_- = 2I$, with the same legend as in Fig. 6.1. Since we are using linear classifiers in these examples, the boundaries are straight lines, except for the optimal decision boundary in Fig. 6.2. Notice that in both experiments, the standard SVM, using L_3 as the loss function, suffers from the imbalanced data. In particular, the decision boundary of the standard SVM is pushed away from the optimal MAP boundary. By tuning the margin parameter ρ correctly, one can move the decision boundary closer to the desired solid black line/curve specifying the MAP boundary. Tuning the uneven cost parameter α may make the performance worse, as depicted in Fig. 6.2.

6.5 Application: H3N2 Challenge Study

In this section, we illustrate the uneven margin SVM for imbalanced data in the H3N2 predictive health problem introduced in Chapter V. Recall that in the 2 v.s. 3 problem (uninfected class v.s. pre-infection class), the majority samples are from class 2, 347 samples from the uninfected class and 90 samples from the pre-infection class. We downsample the majority class so that the number of chips is limited to 7 per class per subject in Chapter V to address the imbalance issue. In this section, we compare different common approaches in the literature as in section 6.4. L_1 and L_3 losses are the same in this experiment, since we pick $\alpha = 0.5$.

The data dimension is first reduced by ANOVA down to the order of 800, in which the treatment is time.

$$x_{i,j,t} = x_{i,j} + \mu_t \textit{ where } i = 1, ..., n \ , \ j = 1, ..., p \textit{ and } t \textit{ represents time.}$$

In other words, the null hypothesis is that mean is constant over time. Only those

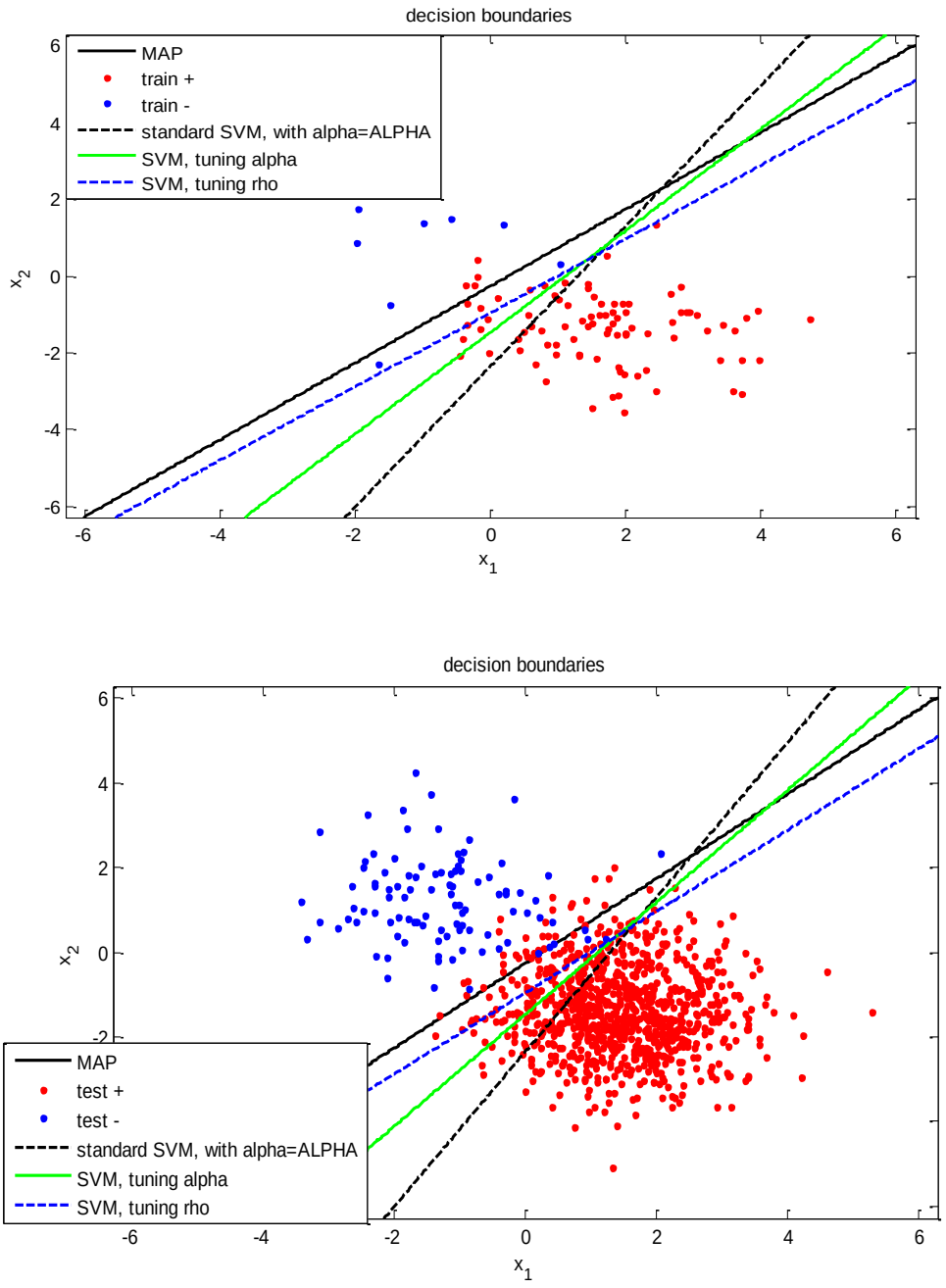

Figure 6.1: The decision boundaries by training the classifiers using different loss functions. The decision boundary trained by L_4 is the one closest to the optimal decision rule. Standard SVM with alpha=ALPHA refers to the loss function L_3; SVM with alpha tuned refers to loss function L_2; SVM with rho tuned refers to the loss function L_4.

Loss function	error rate	number of selected genes
1. L_1	0.1945	21.54
2. L_2	0.1625	57.84
3. L_4	0.1556	48.89
4. downsampling	0.1831	42.00

p-values of one sided paired t-test:

method	error rate	number of genes
(3,1)	$3 < 1$: 0.0011	$3 > 1$: 0.0063
(3,2)	$3 < 2$: 0.2460	$3 < 2$: 0.2038
(3,4)	$3 < 4$: 0.0729	$3 > 4$: 0.3184

Table 6.1: Performance of H3N2 challenge study. The classifier trained with uneven margin α-CC loss function L_4 performs the best in term of error rate.

variables whose mean changes over time are retained for the analysis. Since we perform variable selection and classification together, there is a regularization parameter to be tunned. Due to the uneven margin parameter ρ and the uneven weight parameter $\tilde{\alpha}$, the parameter tunning problem becomes a cross-validation on two-dimensional grids. We speed up the cross validation by line search: cross validation on one of the parameters, while other parameters are fixed to the cross-validated selection in the previous stage. Alternatively search over each parameter, until the selected parameters become stable. All the classifiers are trained with reference. Results are listed in Table 6.1.

From Table 6.1 we conclude that the classifier that minimized the average loss function L_4, which is an α-CC loss function with uneven margin performs the best. Down sampling the majority class to balance the class sample sizes improves the performance as compared to the standard SVM, i.e., having loss L_1, but not as good as the algorithmic approaches using loss functions L_2 and L_4. However, the difficulty of algorithmic approaches is that they involve an additional tuning parameter, and the number of parameters increases linearly as the number of classes increases. In the H3N2 predictive health study, we have a three-dimensional parameter space to search,

the UMSVM tunning parameter ρ or the wights $\tilde{\alpha}$, the initial regularization parameter λ_{init}, and the adaptive regularization parameter λ_{adapt}. In theory, we can extend the uneven margin binary SVM to multi-class SVM, but as the number of parameters increases, it becomes impractical to search for the parameters by cross-validation.

6.6 Conclusion

We have performed several experiments with simulated and real data to explore the benefits of loss functions that are α-classification calibrated for problems where the number of training samples is imbalanced over the classes. We conclude that classifiers trained by loss functions that are α-classification calibrated with uneven margin perform better than those that are not. α-classification calibration is an important property a classifier has to satisfy in order to be consistent for α weighted risk. Our results suggest that tuning the margin (ρ in our notation) is better than tuning the uneven cost ($\tilde{\alpha}$ in our notation) for imbalanced problems.

A direction for future work is to more systematically explore the relation between the degree of imbalance and the improvement from tuning the margin parameter. This can be done by fixing the number of total samples and varying the proportion between the classes, or by fixing the number of the minority class and varying the number of the majority class. Another direction is to examine the consistency of these uneven margin SVM.

6.7 Appendix 1: Kernelized Uneven Margin SVM:

The SVM with uneven margin can be formulated as the following optimization problem

$$\arg\min_{w,b,\xi} \frac{\lambda}{2}||w||^2 + \frac{1}{n}\sum\left(\frac{1-\alpha}{2(1-\rho)}1_{\{y_i=1\}}\xi_i + \frac{\alpha}{2\rho}1_{\{y_i=-1\}}\xi_i\right)$$

$$such\ that\ 2y_i(1-\rho)w^T x_i \geq 1-\xi_i,\ if\ y_i = 1$$

$$2y_i\rho w^T x_i \geq 1-\xi_i, if y_i = -1$$

$$\xi_i \geq 0$$

The dual form can be written as

$$\min_a \frac{1}{2}a^T Q a - e^T a$$

$$such\ that\ y^T a = 0$$

$$0 \leq a_i \leq \frac{1-\alpha}{2(1-\rho)n\lambda},\ if\ y_i = 1$$

$$0 \leq a_i \leq \frac{\alpha}{2\rho n\lambda},\ if\ y_i = -1$$

where $Q_{ij} = 4y_i y_j k(x_i, x_j)(1-\rho)^{1+\frac{y_i+y_j}{2}}\rho^{1-\frac{y_i+y_j}{2}}$, and e represents a column of ones.

Given the solution of the dual form, the decision function becomes

$$f(x) = \sum_i 2a_i y_i k(x_i, x)(1-\rho)^{\frac{1+y_i}{2}}\rho^{\frac{1-y_i}{2}}$$

To solve the kernelized SVM classifier, one can use existing algorithms for uneven-cost SVM, and simply substitute Q for the kernel matrix. If the kernel in linear, another option is to scale each input data x_i by the quantity$2(1-\rho)$ or 2ρ, depending ot the label y_i.

6.8 Appendix 2: Consistency of Support Vector Machines

From the surrogate regret bound, we know that minimizing L-risk is sufficient for the task of minimizing α-risk if L is α-classification calibrated. Hence, we would like to examine the consistency of algorithms, to ensure that the L-risk will be minimized. In [96], Steinwart provided the theory for consistency of kernelized SVM. The argument was based on cost-insensitive classification, but it can be generalized to cost-sensitive

learning.

Consider the optimization problem

$$f_{T,\lambda_n} = \arg\min_{f \in H} \{\Omega(\lambda_n, ||f||_H) + \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))\} \tag{6.3}$$

$$\tilde{f}_{T,\lambda_n} = \arg\min_{f \in H} \{\Omega(\lambda_n, ||f||_H) + \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + b\} \tag{6.4}$$

where H is a reproducing kernel Hilbert space, Ω is a regularization function and L is the loss. Equation (6.3) and (6.4) are very similar, except for the additional offset term in Equation (6.4). In the standard SVM, the regularization is $\Omega(\lambda, ||f||_H) = \lambda||f||_H^2$ and the loss function is the hinge loss $L(y,t) = \max\{0, 1-yt\}$. Let $R_{L,T}(f_{T,\lambda_n})$ be the empirical risk of the decision function, and $R_{L,P}(f_{T,\lambda_n})$ be the true risk. Steinwart showed that for the standard SVM, $|R_{L,T}(f_{T,\lambda_n}) - R_{L,P}(f_{T,\lambda_n})| \to 0$ in probability as $n\lambda_n \to \infty$. In addition, for $\delta > 0$, there exists an integer n_0, such that for all $n > n_0$, $R_{L,P}(f_{T,\lambda_n}) \leq R_{L,P} + \delta$. Steinwart assumed that the loss function L is admissible, defined as

Definition VI.2. A continuous function L is called an admissible loss function if for every $\eta \in [0,1]$ and t_α with $C_L(\eta, t_\eta) = \min_t C_L(\eta, t)$ we have $t_\eta < 0$ if $\eta < 0.5$ and $t_\eta > 0$ if $\eta > 0.5$.

The definition of admissibility VI.2 is a special case of the α-classification calibrated definition VI.1 with $\alpha = 0.5$. In our problem with the loss function (1), and $\phi(t) = \max\{0, 1-t\}$, we can apply the concentration theory [96], since the theory assumed the loss to be admissible, but did not use the admissible property in the argument until proving that it suffices to minimize L-risk to minimize the Bayes risk [96](Proposition 3.3). Indeed, the loss function (1) with $\alpha \neq \frac{1}{2}$ is not admissible, meaning it is not $\frac{1}{2}$-classification calibrated in our terminology, thus consistency of the algorithm does not imply consistency of the error rate. However, it is α-classification

calibrated. If an algorithm is consistent in the error rate then the algorithm is guaranteed minimize the α-risk, implying consistency of the α-risk. The theorems that support the argument are presented as follows. The statements are readjusted to extend Steinwart's work of proving consistency for Bayes risk to cost-sensitive α risk. The main difference occurs when proving that approximating the minimal L-risk is sufficient to approximately achieve the Bayes risk in Steinwart's paper. This should be replaced by the surrogate regret bound for α-classification calibrated loss function to minimize the α risk, presented in section 6.3.

Let $k : X \times X \to R$ be a positive semidefinite function, called a kernel, and an associated Hilbert space (RKHS) $H := \left\{ \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in N, \alpha_i \in R, x_i \in X, i = 1, ..., n \right\}$. Define $K := \sup \left\{ \sqrt{k(x,x)} : x \in X \right\}$, $\delta_\lambda := \sup \{ t : \Omega(\lambda, t) \leq L_1(0) + L_{-1}(0) \}$, and $L_\lambda := L_{|Y \times [-\delta_\lambda K, \delta_\lambda K]}$. The concentration inequality stated below is based on covering numbers. The covering number of a metric space (M, d) is difined by

$$\mathcal{N}((M,d),\varepsilon) := \min\{n \in N | x_1, ..., x_n : M \subset \bigcup_{i=1}^{n} B(x_i, \varepsilon)\}$$

in which $B(x_i, \varepsilon)$ represents a unit ball with center x and radius $\epsilon > 0$. For convenience, the logarithmic covering numbers $\mathcal{H}((M,d),\varepsilon) := \ln \mathcal{N}((M,d),\varepsilon)$ is used in the following statements.

Theorem VI.3. *Let L be an α-classification calibrated (α-CC) loss function, Ω a regularization function , and k a continuous kernel on X. Then for all Borel probability measures P on $X \times Y$ and all $\lambda > 0$, there exists an $f_{P,\lambda} \in H$ with*

$$R^{reg}_{L,P,\lambda}(f_{P,\lambda}) = \inf_{f \in H} R^{reg}_{L,P,\lambda}(f)$$

and $||f_{P,\lambda}|| \leq \delta_\lambda$.

Theorem VI.4. *Let k be a universal kernel on X, L be an α-CC loss function, and*

Ω be a regularization function. Then for every Borel probability measure P on $X \times Y$ we have

$$\lim_{\lambda \to 0} R^{reg}_{L,P,\lambda}(f_{P,\lambda}) = R_{L,P}.$$

Theorem VI.5. *Let L be an α-classification calibrated (α-CC) loss function and P a Borel probability measure on $X \times Y$. The for all $\epsilon > 0$, there exists a $\delta > 0$ such that for all measurable f with $R_{L,P}(f) \leq R_{L,P} + \delta$ we have $R_\alpha(f) \leq R_\alpha + \epsilon$.*

Theorem VI.6. *Let L be an α-classification calibrated (α-CC) loss function, Ω a regularization function , and k a continuous kernel on X. Then for every Borel probability measure P on $X \times Y$, $\epsilon > 0$, $\lambda > 0$, and all $n \geq 1$ we have the outer measure of P^n*

$$\begin{aligned} &\Pr^*(T \in (X \times Y)^n : |R_{L,T}(f_{T,\lambda}) - R_{L,P}(f_{T,\lambda})| \geq \varepsilon) \\ &\leq 2e^{\mathcal{H}(\delta_\lambda I, \varpi^{-1}(L_\lambda, \varepsilon/3)) - \frac{2\varepsilon^2 n}{9||L_\lambda||^2_\infty}} \end{aligned}$$

in which $I : H \to C(X)$ represents the canonical embedding.

The proofs of Theorem VI.3, VI.4, and VI.6 in [96] do not use the admissible property, and Theorem VI.5 is an extension from admissible function to α-CC function, which is proved in [19]. By these results, the consistency theorem in [96] holds for cost-sensitive α-risk.

Theorem VI.7. *Let k be a universal kernel on X, L be an α-CC loss function, and Ω be a regularization function. Suppose we have a positive sequence $\lambda_n \to 0$ and*

$$\frac{||L_{\lambda_\infty}||^2_\infty}{n} \mathcal{H}(\delta_{\lambda_n} I, \varpi^{-1}(L_{\lambda_n}, \varepsilon)) \to 0$$

for all $\epsilon > 0$. Then the classifier based on equation 6.3 is universally consistent.

Similarly, the classifier with offset, based on equation 6.3, is universally consistent if $\frac{||L_{\lambda_\infty}||^2_\infty}{n} H(I, \frac{\varepsilon}{\delta_{\lambda_n} |L_{\lambda_n}|_1}) \to 0$, proved in [96](Theorem 3.12).

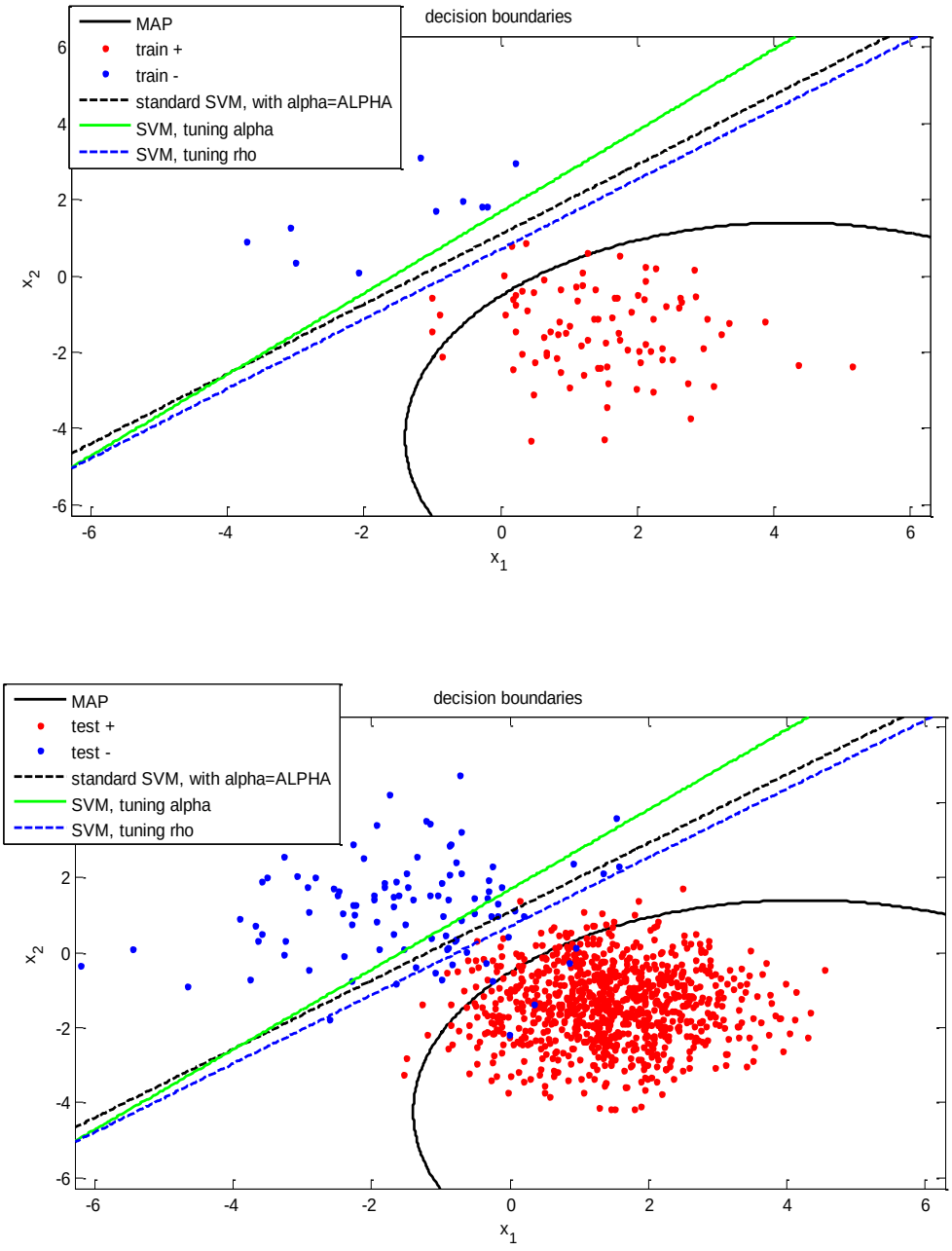

Figure 6.2: The decision boundaries by training the classifiers using different loss functions. In this example, the decision boundary trained by L_4 is the one closest to the optimal decision rule, and the one trained by L_2 is worse than the standard SVM. Standard SVM with alpha=ALPHA refers to the loss function L_3; SVM with alpha tuned refers to loss function L_2; SVM with rho tuned refers to the loss function L_4.

CHAPTER VII

Jointly Sparse Global SIMPLS

7.1 Introduction

Partial least squares (PLS) regression combines dimensionality reduction and prediction using a latent variable model. It was first developed for regression analysis in chemometrics [8], and has been successfully applied to many different areas, including sensory science and more recently genetics [97, 98, 99, 100]. Since Partial least squares regression (PLS-R) does not require matrix inversion or diagonalization, it can be applied to problems with large numbers of variables. As predictor dimension increases, variable selection becomes essential to avoid over-fitting, to provide more accurate predictors and to yield more interpretable parameters. For this reason sparse PLS was developed by H. Chun and S. Keles [9]. The sparse PLS algorithm performs variable selection and dimension reduction simultaneously using an L_1 type variable selection penalty. However, the L_1 penalty used in [9] penalizes each variable independently and this can result in different sets of variables being selected for each PLS component leading to an excessively large number of variables. In this paper we propose a global variable selection approach that penalizes the total number of variables across all PLS components. Put another way, the proposed global penalty guarantees that the selected variables are shared among the PLS components. This results in improved PLS performance with fewer variables. We formulate PLS with

joint sparsity as a variational optimization problem with objective function equal to the univariate PLS criterion with added mixed norm sparsity constraint on the weight matrix. The mixed norm sparsity penalty is the L_1 norm of the L_2 norm on the subsets of variables used by each PLS component. A novel augmented Lagrangian method is proposed to solve the optimization problem and soft thresholding for sparsity occurs naturally as part of the iterative solution. Experiment results show that the modified PLS attains better performance (lower mean squared error, MSE) with many fewer selected predictor variables.

7.2 Partial Least Squares Regression

Partial Least Squares (PLS) methods embrace a suite of data analysis techniques based on algorithms belonging to the PLS family. These algorithms consist of various extensions of the Nonlinear estimation by Iterative PArtial Least Squares (NIPALS) algorithm that was proposed by Herman Wold [101] as an alternative algorithm for implementing a Principal Component Analysis (PCA) [102]. The NIPALS approach was slightly modified by Herman Wold son, Svante, and Harald Martens, in order to obtain a regularized component based regression tool, known as PLS Regression (PLS-R) [8, 103].

Suppose that the data consists of n samples of independent variables $X \in R^{n \times p}$ and dependent variables (responses) $Y \in R^{n \times q}$. In standard PLS Regression the aim is to define orthogonal latent components in R^p, and then use such latent components as predictors for Y in an ordinary least squares framework. The X weights used to compute the latent components can be specified by using iterative algorithms belong to the NIPALS family or by a sequence of eigen-decompositions. The general underlying model is $X = TP' + E$ and $Y = TQ' + F$, where T is the latent component matrix, P and Q are the loading matrices, E and F are the residual terms.

7.2.1 Univariate response

We assume, without loss of generality, that all the variables have been centered in a pre-processing step. For univariate Y, i.e $q = 1$, PLS Regression, also often denoted as PLS1, successively finds X weights $R = [\ \mathbf{r}_1 \quad \mathbf{r}_2 \quad ... \quad \mathbf{r}_K\ ]$ as the solution to the constrained optimization

$$\begin{gathered} \mathbf{r}_k = \arg\max_{\mathbf{r}} \{\mathbf{r}' X'_{(k-1)} Y_{k-1} Y'_{k-1} X_{(k-1)} \mathbf{r}\} \\ s.t.\ \mathbf{r}'\mathbf{r} = 1 \end{gathered} \tag{7.1}$$

where $X_{(k-1)}$ is the matrix of the residuals (i.e., the deflated matrix) from the regression of the X-variables on the first $k-1$ latent components, and $X_0 = X$. Due to the deflation on data after each iteration for finding the weight vector $\mathbf{r}_k$, the orthogonality constraint is satisfied by construction. These weights are then used to find the orthogonal latent components $T = X_{(k-1)}R$. Such components can be also expressed in terms of original variables (instead of deflated variables), i.e., as $T = XW$, where W is the matrix containing the weights to be applied to the original variables in order to exactly obtain the latent components [104].

For a fixed number of components, the response variable Y is predicted in an ordinary least squares regression model where the latent components play the role of the exogenous variables,

$$\arg\min_{Q} \{||Y - TQ'||_2\} = (T'T)^{-1}T'Y.$$

This provides the regression coefficients $\hat{\beta}^{PLS} = W\hat{Q}'$ for the model $Y = X\beta^{PLS} + F$.

Depending on the number of selected latent components the length $\|\hat{\beta}^{PLS}\|_2$ of the vector of the PLS coefficient estimators changes. In particular, de Jong [105] has shown that the sequence of these coefficient vectors have lengths that are strictly increasing as the number of component increases. This sequence converges to the

ordinary least squares coefficient vector and the maximum number of latent components obtainable equals the rank of the X matrix. Thus, by using a number of latent components $K < rank(X)$, PLS-R performs a dimension reduction by shrinking the X matrix. Hence, PLS-R is a suitable tool for problems with data containing many more variables p than observations n.

The objective function in (7.1) can be interpreted as maximizing the squared covariance between Y and the latent component: $corr^2(Y, X_{k-1}\mathbf{r}_k)\text{var}(X_{k-1}\mathbf{r}_k)$. Because the response Y has been taken into account to formulate the latent matrix, PLS usually has better performance in prediction problems than principle component analysis (PCA) does. This is one of the main difference between PLS and principle component analysis (PCA) [106].

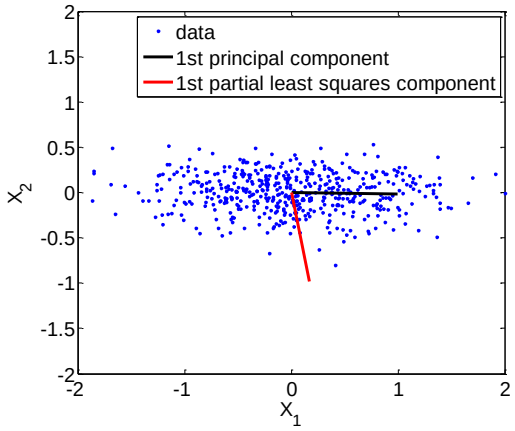

Figure 7.1: A comparison between PCA and PLS: Suppose the response variable is generated to be linear with X_2. The components found by PCA and PLS differ because PLS takes into account the response variables.

7.2.2 Multivariate response

Similarly to univariate response PLS-R, multivariate response PLS-R selects latent components in R^p and R^q , i.e., $\mathbf{t}_k$ and $\mathbf{v}_k$, such that the covariance between $\mathbf{t}_k$ and $\mathbf{v}_k$ is maximized. For a specific component, the sets of weights $\mathbf{r}_k \in R^p$ and $\mathbf{c}_k \in R^q$

are obtained by solving

$$\begin{aligned} \max\{\mathbf{t}'\mathbf{v}\} &= \max\{\mathbf{r}'X'_{k-1}Y_{k-1}\mathbf{c}\} \\ s.t.\ \mathbf{r}'\mathbf{r} &= \mathbf{c}'\mathbf{c} = 1 \end{aligned} \tag{7.2}$$

where $\mathbf{t}_k = X_{(k-1)}\mathbf{r}_k$, $\mathbf{v}_k = Y_{(k-1)}\mathbf{c}_k$, and $X_{(k-1)}$ and $Y_{(k-1)}$ are the deflated matrices associated to X and Y. Notice that the optimal solution $\mathbf{c}_k$ should be proportional to $Y'_{k-1}X_{k-1}\mathbf{r}_k$. Therefore, the optimization in (7.2) is equivalent to

$$\begin{aligned} \max_{\mathbf{r}}\{\mathbf{r}'X'_{k-1}Y_{k-1}Y'_{k-1}X_{k-1}\mathbf{r}\} \\ s.t.\ \mathbf{r}'\mathbf{r} = 1. \end{aligned} \tag{7.3}$$

For each component, the solution to this criterion can be obtained by using a so called PLS2 algorithm. A detailed description of the iterative algorithm as presented by Höskuldsson is in Algorithm 16 [107].

Algorithm 6: PLS2 algorithm

```
for k=1:K do
    initialize r
    X = X_new
    Y = Y_new
    while solution has not converged do
        t = Xr
        c = Y't
        Scale c to length 1
        v = Yc
        r = X'v
        Scale r to length 1
    loading vector p = X't/(t't)
    deflate X_new = X − tp'
    regression b = Y't/(t't)
    deflate Y_new = Y − tb'
    r_k = r
```

In 1993 de Jong proposed a variant of the PLS2 algorithm, called Statistically

Inspired Modification of PLS (SIMPLS), which calculates the PLS latent components directly as linear combinations of the original variables [108]. The SIMPLS was first developed as an optimality problem and solve the optimization

$$\begin{gathered} \mathbf{w}_k = \arg\max_{\mathbf{w}} (\mathbf{w}' X' Y Y' X \mathbf{w}) \\ s.t.\ \mathbf{w}'\mathbf{w} = 1,\ \ \mathbf{w}' X' X \mathbf{w}_j = 0\ for\ j = 1, ..., k-1. \end{gathered} \tag{7.4}$$

Ter Braak and de Jong [109] provided a detailed comparison between the objective functions for PLS2 in (7.3) and SIMPLS in (7.4) and shown that the successive weight vectors $\mathbf{w}_k$ can be derived either from the deflated data matrices or original variables in PLS2 and SIMPLS respectively. Let W^+ be the Moore-Penrose inverse of $W = [\mathbf{w}_1\ \mathbf{w}_2\ ...\ \mathbf{w}_{k-1}]$. The PLS2 algorithm (Algorithm 16) is equivalent to solving the optimization

$$\mathbf{w}_k = \arg\max_{\mathbf{w}} (\mathbf{w}' X' Y Y' X \mathbf{w})$$

$$s.t.\ \mathbf{w}'(I - WW^+)\mathbf{w} = 1, \mathbf{w}' X' X \mathbf{w}_i = 0\ for\ i = 1, ..., k-1.$$

Both NIPALS and SIMPLS have the same objective function but each are maximized under different constraints. NIPALS and SIMPLS are equivalent when Y is univariate, but provide slightly different weight vectors in multivariate scenarios. The performance depends on the nature of the data, but SIMPLS appears easier to interpret since it does not involve deflation of the data sets [108]. We develop our globally sparse PLS based on the SIMPLS optimization formulation.

7.3 Mix Norm Relaxation of Subset Selection

One approach to sparse PLS is to add the L_1 norm of the weight vector, a sparsity inducing penalty, to (7.4). The solution for the first component would be obtained

by solving

$$\mathbf{w}_1 = \arg\max_{\mathbf{w}}(\mathbf{w}'X'YY'X\mathbf{w}) \; s.t. \; \mathbf{w}'\mathbf{w} = 1, \; ||\mathbf{w}||_1 \leq \lambda. \tag{7.5}$$

The addition of the L_1 norm is similar to SCOTLASS (simplified component lasso technique), the sparse PCA proposed by Jolliffe [110]. However, the solution of SCOTLASS is not sufficiently sparse, and the same issue remains in (7.5). Chun and Keles [9] reformulated the problem, promoting the exact zero property by imposing the L_1 penalty on a surrogate of the weight vector instead of the original weight vector [9]. For the first component, they solve the following optimization by alternating between updating $\mathbf{w}$ and $\mathbf{z}$ (block coordinate descent).

$$\mathbf{w}_1, \mathbf{z}_1 = \arg\min_{\mathbf{w},\mathbf{z}}\{-\kappa\mathbf{w}'X'YY'X\mathbf{w}+(1-\kappa)(\mathbf{z}-\mathbf{w})'X'YY'X(\mathbf{z}-\mathbf{w})+\lambda_1||\mathbf{z}||_1+\lambda_2||\mathbf{z}||_2^2\}$$

$$s.t. \; \mathbf{w}'\mathbf{w} = 1$$

As mentioned in the Introduction, this formulation penalizes the variables in each PLS component independently. This thesis proposes an alternative in which variables are penalized simultaneously over all directions. First, we define the global weight matrix, consisting of the K weight vectors, as

$$W = \begin{bmatrix} | & | & & | \\ \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_K \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \mathbf{w}'_{(1)} & - \\ - & \mathbf{w}'_{(2)} & - \\ & \vdots & \\ - & \mathbf{w}'_{(p)} & - \end{bmatrix}.$$

Notice that the elements in a particular row of W, i.e., $\mathbf{w}'_{(j)}$, are all associated with the same predictor variable $\mathbf{x}_j$. Therefore, rows of zeros correspond to variables that are not selected. To illustrate the drawbacks of penalizing each variable independently, as in [9], suppose that each entry in W is selected independently with probability

p_1. The probability that the $(j)_{th}$ variable is not selected becomes $(1-p_1)^K$, and the probability that all the variables are selected for at least one weight vector is $[1-(1-p_1)^K]^p$, which increases as the number of weight vectors K increases. This suggests that for large K the local variable selection approach of [9] may not lead to an overall sparse and parsimonious PLS model. In such cases a group sparsity constraint is necessary to limit the number of selected variables. The jointly sparse global SIMPLS variable selection problem is to find the top K weight vectors that best relate X to Y, while using limited number of variables. This is a subset selection problem that is equivalent to adding a constraint on the L_0 norm of the vector consisting of any norm of each row in W. In other words, counting the number of nonzero rows in W. This leads to the optimization problem

$$\begin{aligned} W &= \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k \\ s.t.\ ||\boldsymbol{\varpi}||_0 \leq t,\ &\mathbf{w}_k'\mathbf{w}_k = 1\ \forall\ k,\ and\ \mathbf{w}_k' X'X\mathbf{w}_i = 0\ \forall\ \ i \neq k \end{aligned} \tag{7.6}$$

in which

$$\boldsymbol{\varpi} = \begin{bmatrix} ||\mathbf{w}_{(1)}||_2 \\ ||\mathbf{w}_{(2)}||_2 \\ \vdots \\ ||\mathbf{w}_{(p)}||_2 \end{bmatrix}.$$

The objective function (7.6) is the summation of the first K terms in the SIMPLS objective, which we refer to as global SIMPLS. Instead of the sequential greedy solution in PLS2 algorithm, the proposed jointly sparse global SIMPLS must solve for the K weight vectors simultaneously. Given the complexity of this combinatorial problem, as is standard optimization practice, we relax the L_0 norm optimization to

a mixed norm structured sparsity penalty [111].

$$W = \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{w}_{(j)}||_2 \\ s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall\ k\ and\ \mathbf{w}_k' X'X\mathbf{w}_i = 0\ \forall\ i \neq k \tag{7.7}$$

The L_2 norm of each row of W promotes grouping entries in W that relate to the same predictor variable, whereas the L_1 norm promotes a small number of groups, as in (7.5).

Suppose we rotate the independent variables by a rotation matrix R_x and the response variables by a rotation matrix R_y. The optimization becomes

$$\begin{aligned} W &= \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k'(XR_x)'(YR_y)(YR_y)'(XR_x)\mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{w}_{(j)}||_2 \\ &= \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' R_x' X'YY'XR_x\mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{w}_{(j)}||_2 \end{aligned} \tag{7.8}$$

$$s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall\ k\ and\ \mathbf{w}_k'(XR_x)'(XR_x)\mathbf{w}_i = \mathbf{w}_k' R_x' X'XR_x\mathbf{w}_i = 0\ \forall\ i \neq k.$$

Since we do not impose sparsity constraints on the response variables, the response variables are invariant to rotation, whereas the independent variables are.

7.4 Algorithmic Implementation for jointly sparse Global SIMPLS

Constrained eigen-decomposition and group variable selection are each well-studied problems for which efficient algorithms have been developed. We propose to solve the optimization (7.7) by augmented Lagrangian methods, which allows one to solve (7.7) by variable splitting iterations. Augmented Lagrangian methods introduce a new variable M, constrained such that $M = W$, such that the row vectors $\mathbf{m}_{(j)}$ of M obey

the same structural pattern as the rows of W:

$$\begin{aligned} &\min_{W,M} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k + \lambda \sum_{j=1}^{p} ||\mathbf{m}_{(j)}||_2 \\ &s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall\ k\ ,\ \mathbf{w}_k' X'X\mathbf{w}_i = 0\ \forall\ i \neq k,\ and\ M = W \end{aligned} \tag{7.9}$$

The optimization (7.9) can be solved by replacing the constrained problem by an unconstrained one with an additional penalty on the Frobenius norm of the difference $M - W$. This penalized optimization can be iteratively solved by a block coordinate descent method that alternates between optimizing over W and over M (See algorithm 6). We initialize the algorithm 6 with $M(0)$ equals to the solution of standard PLS, and $D(0)$ equals to the zero matrix. Once the algorithm converges, the final PLS regression coefficients are obtained by applying the standard PLS regression on the selected variables keeping the same number of components K. The optimization over W can be further simplified to a secular equation problem, whereas the optimization over M can be shown to reduce to solving a soft thresholding operation. As described later in the experimental comparisons section, the parameters λ and K are decided by cross validation.

Algorithm 7: Algorithm for solving the global SIMPLS with variable selection problem using the augmented Lagrangian method

set $\tau = 0$, choose $\mu > 0$, $M(0)$, $W(0)$, $D(0)$;
while *stopping criterion is not satisfied* **do**
  $W(\tau+1) = \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k + \frac{\mu}{2}||W - M(\tau) - D(\tau)||_F^2$
  $s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall k,\ \mathbf{w}_k' X'X\mathbf{w}_i = 0\ \forall\ i \neq k;$
  $M(\tau+1) = \arg\min_{M} \lambda \sum_{j=1}^{p} ||\mathbf{m}_{(j)}||_2 + \frac{\mu}{2}||W(\tau+1) - M - D(\tau)||_F^2;$
  $D(\tau+1) = D(\tau) - W(\tau+1) + M(\tau+1);$

Optimization over W The following optimization in algorithm 6 is a nonconvex quadratically constrained quadratic program (QCQP). The nonconvexity is mainly

due to the equality constraints.

$$W(\tau+1) = \arg\min_{W} -\frac{1}{n^2}\sum_{k=1}^{K} \mathbf{w}_k' X'YY'X\mathbf{w}_k + \frac{\mu}{2}||W - M(\tau) - D(\tau)||_F^2$$

$$s.t.\ \mathbf{w}_k'\mathbf{w}_k = 1\ \forall k,\ \mathbf{w}_k' X'X\mathbf{w}_i = 0\ \forall\ i \neq k$$

We propose solving for the K vectors in W successively by a greedy approach. Let $\mathbf{m}_k$ and $\mathbf{d}_k$ be the columns of the matrices M and D, and $\boldsymbol{\omega}_k = \mathbf{m}_k + \mathbf{d}_k$. The optimization over W becomes

$$\begin{aligned} \mathbf{w}_k(\tau+1) = \arg\min_{\mathbf{w}} -\frac{1}{n^2}\mathbf{w}' X'YY'X\mathbf{w} + \frac{\mu}{2}||\mathbf{w} - \boldsymbol{\omega}_k||_2^2 \\ s.t.\ \ \mathbf{w}'\mathbf{w} = 1,\ \mathbf{w}' X'X\mathbf{w}_i = 0\ \forall\ i < k. \end{aligned} \tag{7.10}$$

Let N be an orthonormal basis for the orthogonal complement of $\{X'X\mathbf{w}_i\}, i < k$. The optimization (7.10) can be solved by the method of Lagrange multipliers. The solution is $\mathbf{w}_k = N(A - \alpha I)^{-1}\mathbf{b}$, in which $A = -\frac{1}{n^2}N'X'YY'XN$, $\mathbf{b} = \frac{\mu}{2}N'\boldsymbol{\omega}_k$ and α is the minimum solution that satisfies $\mathbf{b}'(A - \alpha I)^{-2}\mathbf{b} = 1$. To see this, let $\mathbf{w} = N\tilde{\mathbf{w}}$. The optimization (7.10) can be written as

$$\min \tilde{\mathbf{w}}' A\tilde{\mathbf{w}} - 2\mathbf{b}'\tilde{\mathbf{w}}\ \ s.t.\ \tilde{\mathbf{w}}'\tilde{\mathbf{w}} = 1.$$

Since we assume that $\mathbf{w}$ is a linear combination of the basis vectors in N, the orthogonality conditions in (7.10) are automatically satisfied. Hence these conditions have been dropped in the new formulation. Then using Lagrange multipliers, we can show that the solution as stated above. Suppose there are two solutions of α that satisfy $\mathbf{b}'(A - \alpha I)^{-2}\mathbf{b} = 1$, corresponding to two pairs of solutions to the optimization, $(\tilde{\mathbf{w}}_1, \alpha_1)$ and $(\tilde{\mathbf{w}}_2, \alpha_2)$. It can be shown that

$$(\tilde{\mathbf{w}}_1' A\tilde{\mathbf{w}}_1 - 2\mathbf{b}'\tilde{\mathbf{w}}_1) - (\tilde{\mathbf{w}}_2' A\tilde{\mathbf{w}}_2 - 2\mathbf{b}'\tilde{\mathbf{w}}_2) = \frac{\alpha_1 - \alpha_2}{2}||\tilde{\mathbf{w}}_1 - \tilde{\mathbf{w}}_2||_2^2.$$

Hence, one should select the minimum among all the α's.

The equation $\mathbf{b}'(A-\alpha I)^{-2}\mathbf{b}=1$ is a secular equation, a well studied problem in constrained eigenvalue decomposition [112][113]. We can diagonalize the matrix A as $A=UDU'$, in which D is diagonal with eigenvalues $d_1, d_2, ..., d_p$in decreasing order on the diagonal, and the columns of U are the corresponding eigenvectors. Define

$$g(\alpha)=\mathbf{b}'(A-\alpha I)^{-2}\mathbf{b}=\mathbf{b}'U\begin{bmatrix}\frac{1}{(d_1-\alpha)^2} & & & \\ & \frac{1}{(d_2-\alpha)^2} & & \\ & & \ddots & \\ & & & \frac{1}{(d_p-\alpha)^2}\end{bmatrix}U'\mathbf{b}.$$

Let $\tilde{\mathbf{b}}=U'\mathbf{b}$, then $g(\alpha)=\mathbf{b}'(A-\alpha I)^{-2}\mathbf{b}=\sum_i \frac{\tilde{b}_i^2}{(d_i-\alpha)^2}$, and hence $g(\alpha)=1$ is a secular equation. $g(\alpha)$ increases strictly as α increases from $-\infty$ to d_p, since

$$g'(\alpha)=\sum_i \frac{2\tilde{b}_i^2}{(d_i-\alpha)^3}$$

is positive for $-\infty<\alpha<d_p$. Moreover, given the limits

$$\lim_{\alpha\to-\infty} g(\alpha)=0$$

$$\lim_{\alpha\to d_p^-} g(\alpha)=\infty$$

we can conclude that there is exactly one solution $\alpha<d_p$ to the equation $g(\alpha)=1$, [112]. An iterative algorithm (algorithm 3) is used to solve $g(\alpha)=1$ starting from a point to the left of the smallest eigenvalue d_p [113].

Algorithm 8: iteration for solving secular equation

1 set $\tau=0$, choose $\alpha_0=d_p-\varepsilon_1$;
2 **while** *stopping criterion is not satisfied,* $|g(\alpha_\tau)-1|>\varepsilon_2$ **do**
3 $\quad\alpha_{\tau+1}=\alpha_\tau+2\frac{g^{-1/2}(\alpha_\tau)-1}{g^{-3/2}(\alpha_\tau)g'(\alpha_\tau)}$;

Optimization over M The optimization over M has a closed form solution. Let $\Delta = W(\tau+1) - D(\tau)$, then each row of M is given as $\mathbf{m}_{(j)} = [||\delta_{(j)}|| - \frac{\lambda}{\mu}]_+ \frac{\delta_{(j)}}{||\delta_{(j)}||}$.

7.5 Simulation Experiments

We implement the simulation models in [9]. There are four models all following $Y = X\beta + f$, $n = 100$, and $p = 5000$. We compare five different methods: the standard PLS-R, PLS generalized linear regression proposed by Bastien et al. [114], L_1 penalized PLS-R [9], Lasso [84] and the jointly sparse global SIMPLS-R (or denoted as L_1/L_2 SPLS in the performance comparison tables). All the methods select the parameters by ten fold cross-validation, except for the PLS generalized linear regression, which stops to include an additional component if the new component is not significant. Two i.i.d sets are generated for each trial: one as the training set and one as the test set. Ten trials are conducted for each model, and averaged results are listed in Table 7.1, 7.2, 7.3, and 7.4. The details of the models are as follows:

model 1

$$H_{1j} = \begin{cases} 3 & 1 \le j \le 50 \\ 4 & 51 \le j \le n \end{cases}$$

$$H_{2j} = 3.5$$

$$X_i = \begin{cases} H_1 + \varepsilon_i & 1 \le i \le 50 \\ H_2 + \varepsilon_i & 51 \le i \le p \end{cases}$$

$$\beta = \begin{cases} \frac{1}{25} & 1 \le i \le 50 \\ 0 & 51 \le i \le p \end{cases}$$

ε_i *is* $N(0, I_n)$ *distributed, and* f *is* $N(0, 1.5^2 I_n)$ *distributed.*

model 2

$H_{1j} = 3I(j \leq 50) + 4I(j > 50)$

$H_{2j} = 3.5 + 1.5I(u_{1j} \leq 0.4)$

$H_{3j} = 3.5 + 0.5I(u_{2j} \leq 0.7)$

$H_{4j} = 3.5 - 1.5I(u_{3j} \leq 0.3)$

$H_{5j} = 3.5$

u_{1j}, u_{2j}, u_{3j} *are i.i.d from* $Unif(0,1)$

$X_i = H_j + \varepsilon_i,\ n_{j-1} \leq i \leq n_j,\ j = 1, ..., 5, (n_0, ..., n_5) = (0, 50, 100, 200, 300, p)$

$$\beta = \begin{cases} \frac{1}{25} & 1 \leq i \leq 50 \\ 0 & 51 \leq i \leq p \end{cases}$$

ε_i *is* $N(0, I_n)$ *distributed, and* f *is* $N(0, 1.5^2 I_n)$ *distributed.*

model 3

$H_{1j} = 2.5I(j \leq 50) + 4I(j > 50)$

$H_{2j} = 2.5I(1 \leq j \leq 25, 51 \leq j \leq 75) + 4I(26 \leq j \leq 50, 76 \leq j \leq 100)$

$H_{3j} = 3.5 + 1.5I(u_{1j} \leq 0.4)$

$H_{4j} = 3.5 + 0.5I(u_{2j} \leq 0.7)$

$H_{5j} = 3.5 - 1.5I(u_{3j} \leq 0.3)$

$H_{6j} = 3.5$

u_{1j}, u_{2j}, u_{3j} *are i.i.d from* $Unif(0,1)$

$X_i = H_j + \varepsilon_i,\ n_{j-1} \leq i \leq n_j,\ j = 1, ..., 5, (n_0, ..., n_6) = (0, 25, 50, 100, 200, 300, p)$

$$\beta = \begin{cases} \frac{1}{25} & 1 \leq i \leq 50 \\ 0 & 51 \leq i \leq p \end{cases}$$

ε_i *is* $N(0, I_n)$ *distributed, and* f *is* $N(0, I_n)$ *distributed.*

model 4

$H_{1j} = I(j \leq 50) + 6I(j > 50)$

$H_{2j} = 3.5 + 1.5I(u_{1j} \leq 0.4)$

$H_{3j} = 3.5 + 0.5I(u_{2j} \leq 0.7)$

$H_{4j} = 3.5 - 1.5I(u_{3j} \leq 0.3)$

$H_{5j} = 3.5$

u_{1j}, u_{2j}, u_{3j} *are i.i.d from* $Unif(0,1)$

$X = (X^{(1)}, X^{(2)})$

$X^{(1)}$ *is generated from* $N(0, \Sigma)$, Σ *is from* $AR(1)$ *with* $\rho = 0.9$.

$X_i^{(2)} = H_j + \varepsilon_i,\ n_{j-1} \leq i \leq n_j,\ j = 1, ..., 5, (n_0, ..., n_5) = (0, 50, 100, 200, 300, p - 50)$

$\beta_i = k_j$ *for* $n_{j-1} + 1 \leq i \leq n_j,\ j = 1, ..., 6,$ *where*

$(n_0, ..., n_6) = (0, 10, 20, 30, 40, 50, p),\ (k_1, ..., k_6) = (8, 6, 4, 2, 1, 0)/25$

ε_i *is* $N(0, I_n)$ *distributed, and* f *is* $N(0, 1.5^2 I_n)$ *distributed.*

In most of the simulations, we observe that the proposed jointly sparse globel SIMPLS-R performs the best in most cases in terms of the prediction MSE. In particular, the number of variables and the number of components chosen in jointly sparse global SiMPLS-R are usually less than the L_1 SPLS method. The cross validation time for globally sparse PLS is long, searching over a two-dimensional grids of the number of components and the regularization parameter. However, the performance improves.

7.6 Application 1: Chemometrics

In this section we show experimental results obtained by comparing standard PLS-R, L_1 penalized PLS-R [9] (denoted as L_1 SPLS in the performance table), our proposed jointly sparse global SIMPLS-R (denoted as L_1/L_2 SPLS in the performance table), and Correlated Component Regression [115]. All the methods have

Model 1:	1. PLS-R	2. Bastien	3. L_1 SPLS	4. Lasso	5. L_1/L_2 SPLS
number of comp.	1.4	5	1.9	NaN	1.4
number of variables	5000	1129.4	246.5	40.7	276.1
R^2	0.98	1	0.71	0.59	0.83
MSE	3.14	2.98	3.00	3.23	2.82
Time CV	101.47	0	43.49	51.59	11414
Time analysis	0.89	121.91	0.05	0.04	6.40
Time prediction	0.010	0.011	0.002	0.04	0.002
Total time	102.37	121.92	43.54	51.67	11420

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(5,1)	$5 \neq 1$: 0.5	$5 < 1$: 1.84×10^{-11}	$5 < 1$: 0.01
(5,2)	$5 < 2$: 4.49×10^{-7}	$5 < 2$: 5.46×10^{-5}	$5 < 2$: 0.043
(5,3)	$5 < 3$: 0.19	$5 > 3$: 0.35	$5 < 3$: 0.075
(5,4)	NaN	$5 > 4$: 0.053	$5 < 4$: 0.006

Table 7.1: Simulation Model 1.

Model 2:	1. PLS-R	2. Bastien	3. L_1 SPLS	4. Lasso	5. L_1/L_2 SPLS
number of comp.	2	5	2.3	NaN	1.1
number of variables	5000	1158.4	273.4	15.8	171.7
R^2	0.98	1	0.79	0.39	0.75
MSE	3.18	2.99	2.93	3.09	2.69
Time CV	100.51	0	43.03	53.97	11420
Time analysis	1.28	122.69	0.06	0.04	5.72
Time prediction	0.010	0.011	0.002	0.039	0.001
Total time	101.80	122.70	43.09	54.05	11426

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(5,1)	$5 < 1$: 0.0671	$5 < 1$: 1.13×10^{-14}	$5 < 1$: 8.69×10^{-4}
(5,2)	$5 < 2$: 1.19×10^{-11}	$5 < 2$: 6.91×10^{-9}	$5 < 2$: 0.0046
(5,3)	$5 < 3$: 0.0184	$5 < 3$: 0.1041	$5 < 3$: 0.0344
(5,4)	NaN	$5 > 4$: 0.0119	$5 < 4$: 2.70×10^{-4}

Table 7.2: Simulation Model 2.

Model 3:	1. PLS-R	2. Bastien	3. L_1 SPLS	4. Lasso	5. L_1/L_2 SPLS
number of comp.	1.4	5	1.4	NaN	1.5
number of variables	5000	1156.4	89.2	41.3	60.5
R^2	0.98	1	0.77	0.75	0.73
MSE	1.82	1.48	1.27	1.48	1.25
Time CV	102.61	0	43.84	49.45	11295
Time analysis	1.03	126.08	0.04	0.04	5.48
Time prediction	0.01	0.01	0.001	0.039	0.001
Total time	103.65	126.09	43.88	49.53	11300

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(5,1)	$5 > 1$: 0.4201	$5 < 1$: 2.32×10^{-19}	$5 < 1$: 1.18×10^{-4}
(5,2)	$5 < 2$: 5.53×10^{-6}	$5 < 2$: 1.39×10^{-12}	$5 < 2$: 0.0104
(5,3)	$5 > 3$: 0.3632	$5 < 3$: 0.1697	$5 < 3$: 0.3430
(5,4)	NaN	$5 > 4$: 0.1726	$5 < 4$: 0.0054

Table 7.3: Simulation Model 3.

Model 4:	1. PLS-R	2. Bastien	3. L_1 SPLS	4. Lasso	5. L_1/L_2 SPLS
number of comp.	2	5	2.6	NaN	2.1
number of variables	5000	1118.8	1260.8	9.4	1180.5
R^2	1	1	0.78	0.19	0.91
MSE	2.15	2.29	2.41	2.14	2.36
Time CV	98.16	0	44.31	50.52	12051
Time analysis	1.55	123.79	0.10	0.04	7.97
Time prediction	0.010	0.011	0.007	0.042	0.004
Total time	99.73	123.8	44.41	50.60	12059

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(5,1)	$5 > 1$: 0.4057	$5 < 1$: 9.62×10^{-5}	$5 > 1$: 0.0087
(5,2)	$5 < 2$: 6.82×10^{-5}	$5 > 2$: 0.4618	$5 > 2$: 0.1874
(5,3)	$5 < 3$: 0.2201	$5 < 3$: 0.3918	$5 < 3$: 0.3812
(5,4)	NaN	$5 > 4$: 0.0485	$5 > 4$: 0.0056

Table 7.4: Simulation Model 4.

been applied on the Octane data set (see [104]). The Octane data is a real data set consisting of 39 gasoline samples for which the digitized Octane spectra have been recorded at 225 wavelengths (in nm). The aim is to predict the Octane number, a key measurement of the physical properties of gasoline, using the spectra as predictors. This is of major interest in real applications, because the conventional procedure to calculate the Octane number is time consuming and involves expensive and maintenance-intensive equipment as well as skilled labor.

The experiments are composed of 150 trials. In each trial we randomly split the 39 samples into 26 training samples and 13 test samples. The regularization parameter λ and number of components K are selected by 2-fold cross validation on the training set, while μ is fixed to 2000. The averaged results over the 150 trials are shown in Table 7.5. All the methods but CCR perform reasonably in terms of MSE on the test set. We further show the variable selection frequencies for the first three PLS methods over the 150 trials superimposed on the octane data in Fig. 7.2. In chemometrics, the rule of thumb is to look for variables that have large amplitudes in first derivatives with respect to wavelength. Notice that both L_1 penalized PLS-R and jointly sparse global SIMPLS have selected variables around 1200 and 1350 nm, and the selected region in the latter case is more confined. Box and Whisker plots for comparing the MSE, number of selected variables, and number of components of these three PLS formulations are shown in Fig. 7.3. Comparing our proposed jointly sparse global SIMPLS with standard PLS and L_1 penalized PLS [9], we see that global SIMPLS with joint variable selection attains better performance in terms of MSE, the number of predictors, and the number of components.

7.7 Application 2: Predictive Health Study

In this section we apply the jointly sparse global SIMPLS-R to 3 predictive health challenge studies. The H3N2 challenge study introduced in Chapter V, and an-

Table 7.5: Performance of the global SIMPLS with joint variable selection compared with standard PLS, L_1 penalized PLS and CCR.

methods	MSE	number of var.	number of comp.
PLS-R	0.0564	225	5.5
L_1 SPLS	0.0509	87.3	4.5
L_1/L_2 SPLS	0.0481	38.5	3.8
CCR	0.8284	19.1	6

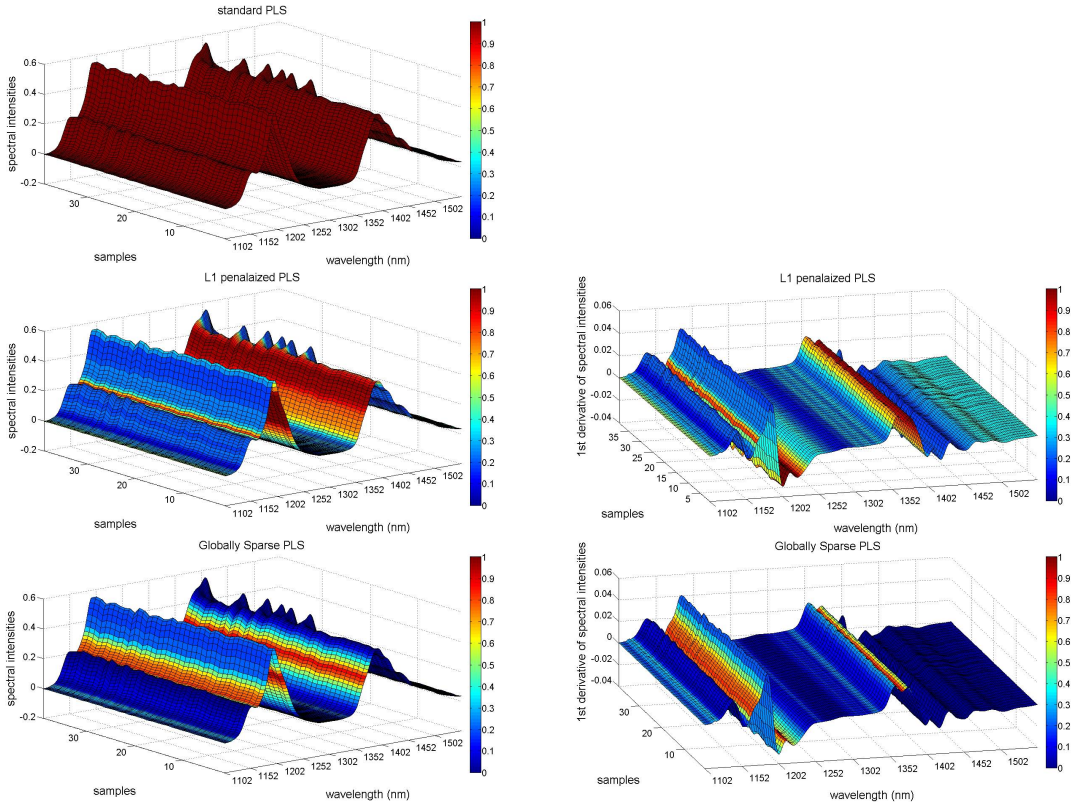

Figure 7.2: Variable selection frequency superimposed on the octane data: The hight of the surfaces represents the exact value of the data over 225 variables for the 39 samples. The color of the surface shows the selection frequency of the variables as depicted on the colorbar.

other two: H1N1 and HRV studies. The H1N1 D3 challenge study consisted of 24 pre-screened volunteers without recent influenza-like illness in the preceding 45 days. These subjects had samples taken 24 hours prior to inoculation of A/Brisbane/59/2007

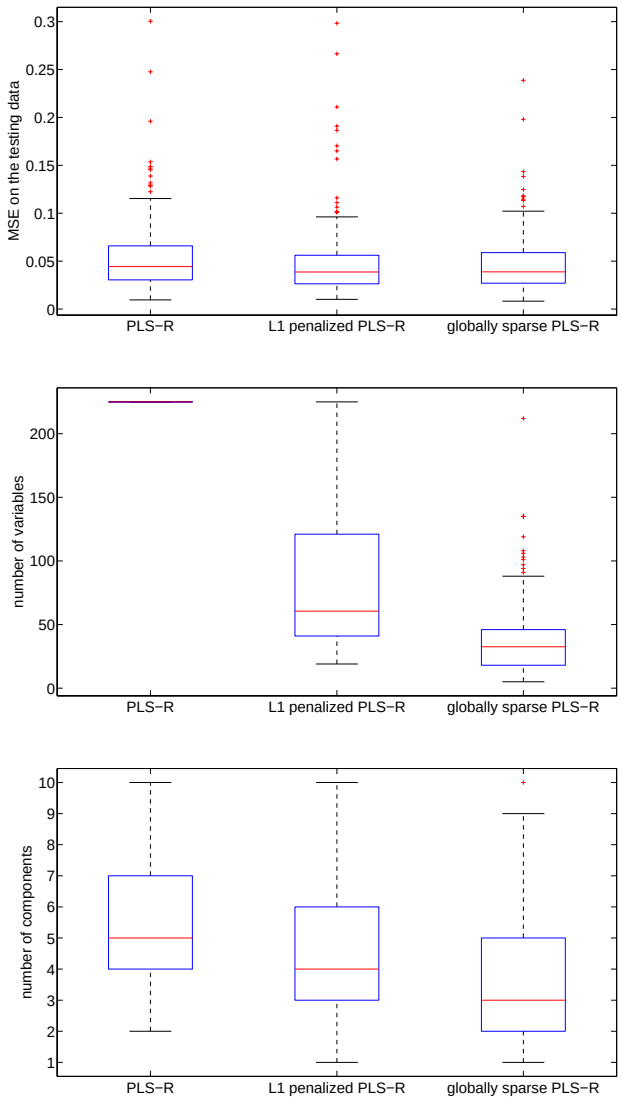

Figure 7.3: The Box and Whisker plot for comparing MSE, and number of selected variables, and number of components on the test samples.

(H1N1) and immediately prior to inoculation. Peripheral blood was taken at baseline, then at 8 hour intervals until 108 hours after inoculation. The H1N1 D4 had 19 pre-screened volunteers. Subjects had two references (29 and 5 hours prior to inoculation of A/Brisbane/59/2007 (H1N1)) then sampled at 8 hour intervals for the initial 142 hours and again at 165 hours. The H1N1 UVA challenge study consisted of 20 pre-screened volunteers without recent influenza-like illness in the preceding 45

days. These subjects had samples taken 12 hours prior to inoculation of HRV and immediately prior to inoculation. Peripheral blood was taken at baseline, then at 4 to 6 hour intervals until 36 hours after inoculation and again at the 3rd and 4th day. The HRV DUKE study had 30 pre-screened volunteers. Subjects had two references as the UVA study then sampled at 4 to 6 hour intervals for the initial 48 hours and then 12 hour intervals unitl 132 hours.

The prediction task in these experiments is to predict the symptom scores based on gene expression. The symptoms are self-reported scores, ranging from 0 to 3. We compare the jointly sparse global SIMPLS-R with standard PLS by leaving one subject out as the test set, and the rest as the training set. The process is repeated until all subjects have been treated as the test set. The number of components for both methods and the regularization parameter in jointly sparse global SIMPLS-R are selected by 2-fold cross validation. Since each subject has multiple samples, we perform the cross validation by splitting by subjects, i.e., no samples from the same subject will appear in both training and tunning sets. The results for H3N2 are listed in Table 7.6 and 7.7. We further restrict the responses to the first 3 symptoms, which are the upper respiratory symptoms in Table 7.8 and 7.9. We notice that restricting the responses to the upper respiratory symptoms improves the performances. The reason may be that these viruses are more closely related to the upper respiratory symptoms than the others. Results on H1N1, HRV UVA, HRV DUKE are listed in Table 7.10, 7.11, 7.12, 7.13, 7.14, 7.15, 7.16, 7.17, 7.19, 7.19, 7.20, and 7.21. In most of the cases, the jointly sparse global SIMPLS-R outperforms the standard PLS-R in terms of prediction MSE, number of components, number of genes.

7.8 Application 3: Agriculture

This application studies the relationship between the wine and growing conditions [116], in which a set of 27 wines from the same chateau (cheval blanc) but made from

	1. PLS	2. SPLS
number of comp.	1.97	1.29
number of genes	12023	57.29
Overall MSE	1.72	1.66

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 7.66×10^{-4}	$1 > 2$: 8.78×10^{-15}	$1 > 2$: 0.0070

Table 7.6: Overall performance of PLS and jointly sparse global SIMPLS applied to 10 H3N2 symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.14	0.14	0.41	0.39	0.38	0.40
Stuffy nose	0.26	0.25	0.50	0.47	0.34	0.36
Sneezing	0.17	0.17	0.38	0.37	0.32	0.34
Sore throat	0.22	0.22	0.14	0.14	0.06	0.05
Earache	0.03	0.03	0.10	0.11	0.19	0.19
Malaise	0.27	0.26	0.30	0.35	0.22	0.22
Cough	0.16	0.15	0.10	0.10	0.05	0.04
Short. of breaths	0.10	0.10	0.15	0.17	0.15	0.15
Headache	0.24	0.22	0.40	0.40	0.29	0.31
Myalgia	0.13	0.12	0.31	0.32	0.32	0.35

Table 7.7: Performance of PLS and jointly sparse global SIMPLS applied to 10 H3N2 symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	4.95	4.79
number of genes	12023	246.63
Overall MSE	0.54	0.49

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 0.0416	$1 > 2$: 2.14×10^{-14}	$1 > 2$: 0.0099

Table 7.8: Overall performance of PLS and jointly sparse global SIMPLS applied to 3 H3N2 symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.13	0.12	0.53	0.47	0.59	0.60
Stuffy nose	0.24	0.23	0.67	0.71	0.54	0.59
Sneezing	0.16	0.14	0.52	0.48	0.50	0.56

Table 7.9: Performance of PLS and jointly sparse global SIMPLS applied to 3 H3N2 symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	4.98	4.95
number of genes	12023	256.63
Overall MSE	0.97	1.02

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 0.2849	$1 > 2$: 2.95×10^{-130}	$1 < 2$: 0.0019

Table 7.10: Overall performance of PLS and jointly sparse global SIMPLS applied to 10 H1N1 symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.16	0.18	0.34	0.44	0.31	0.41
Stuffy nose	0.18	0.20	0.59	0.49	0.56	0.53
Sneezing	0.09	0.10	0.31	0.28	0.37	0.39
Sore throat	0.09	0.09	0.18	0.14	0.19	0.20
Earache	0.04	0.04	0.13	0.12	0.19	0.18
Malaise	0.10	0.10	0.24	0.23	0.26	0.26
Cough	0.03	0.04	0.11	0.10	0.17	0.15
Short. of breaths	0.01	0.01	0.05	0.06	0.14	0.15
Headache	0.15	0.15	0.37	0.36	0.42	0.39
Myalgia	0.11	0.12	0.32	0.32	0.43	0.41

Table 7.11: Performance of PLS and jointly sparse global SIMPLS applied to 10 H1N1 symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	5	4.95
number of genes	12023	105.68
Overall MSE (sum of MSEs)	0.47	0.54

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 0.08	$1 > 2$: 6.33×10^{-146}	$1 < 2$: 1.19×10^{-6}

Table 7.12: Overall performance of PLS and jointly sparse global SIMPLS applied to 3 H1N1 symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.18	0.20	0.50	0.58	0.38	0.42
Stuffy nose	0.20	0.24	0.74	0.62	0.62	0.49
Sneezing	0.10	0.11	0.41	0.37	0.41	0.38

Table 7.13: Performance of PLS and jointly sparse global SIMPLS applied to 3 H1N1 symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	3.45	2.3
number of genes	12023	233.75
Overall MSE (sum of MSEs)	0.91	0.87

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 0.0082	$1 > 2$: 1.11×10^{-64}	$1 > 2$: 0.1098

Table 7.14: Overall performance of PLS and jointly sparse global SIMPLS applied to 8 HRV UVA symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Sneezing	0.06	0.06	0.24	0.24	0.30	0.33
Runny Nose	0.17	0.16	0.45	0.45	0.30	0.32
Nasal Obstruc.	0.32	0.30	0.64	0.57	0.34	0.36
Sore Throat	0.22	0.22	0.47	0.52	0.27	0.33
Cough	0.05	0.05	0.10	0.09	0.05	0.05
Headache	0.02	0.02	0.06	0.08	0.08	0.10
Malaise	0.06	0.05	0.18	0.16	0.16	0.20
Chilliness	0.00	0.00	0.00	0.00	0.00	0.00

Table 7.15: Performance of PLS and jointly sparse global SIMPLS applied to 8 HRV UVA symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	3.5	2.15
number of genes	12023	200.5
Overall MSE	0.53	0.51

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 0.0074	$1 > 2$: 1.04×10^{-63}	$1 > 2$: 0.1975

Table 7.16: Overall performance of PLS and jointly sparse global SIMPLS applied to 3 HRV UVA symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Sneezing	0.06	0.06	0.28	0.30	0.31	0.35
Runny Nose	0.17	0.16	0.51	0.55	0.33	0.36
Nasal Obstruc.	0.30	0.29	0.77	0.73	0.42	0.40

Table 7.17: Performance of PLS and jointly sparse global SIMPLS applied to 3 HRV UVA symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	3.26	1.17
number of genes	12023	112.26
Overall MSE	1.64	1.54

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 2.06×10^{-7}	$1 > 2$: 2.02×10^{-79}	$1 > 2$: 4.9×10^{-4}

Table 7.18: Overall performance of PLS and jointly sparse global SIMPLS applied to 10 HRV DUKE symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.27	0.23	0.49	0.52	0.43	0.38
Stuffy nose	0.33	0.30	0.57	0.54	0.41	0.32
Sneezing	0.14	0.13	0.30	0.33	0.28	0.29
Sore throat	0.16	0.17	0.24	0.22	0.18	0.11
Earache	0.17	0.18	0.31	0.30	0.27	0.19
Malaise	0.27	0.26	0.29	0.35	0.26	0.21
Cough	0.02	0.02	0.06	0.05	0.09	0.07
Short. of breaths	0.23	0.22	0.24	0.17	0.11	0.06
Headache	0.01	0.01	0.03	0.04	0.05	0.06
Myalgia	0.03	0.03	0.07	0.07	0.09	0.07

Table 7.19: Performance of PLS and jointly sparse global SIMPLS applied to 10 HRV DUKE symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

	1. PLS	2. SPLS
number of comp.	2.87	1
number of genes	12023	79.26
Overall MSE	0.71	0.65

p-values of one sided paired t-test:

method	number of comp.	number of var.	MSE
(1,2)	$1 > 2$: 2.38×10^{-9}	$1 > 2$: 7.82×10^{-91}	$1 > 2$: 0.0046

Table 7.20: Overall performance of PLS and jointly sparse global SIMPLS applied to 3 HRV DUKE symptoms scores and gene expression.

	PLS MSE	SPLS MSE	PLS Wy	SPLS Wy	PLS R^2	SPLS R^2
Runny nose	0.26	0.23	0.61	0.65	0.45	0.37
Stuffy nose	0.31	0.29	0.70	0.64	0.43	0.31
Sneezing	0.14	0.13	0.36	0.41	0.28	0.29

Table 7.21: Performance of PLS and jointly sparse global SIMPLS applied to 3 HRV DUKE symptoms scores and gene expression. Wy represents the averaged L_1 norm of the response weight vector. R^2 represents the R square value on the training set.

Set 1	Set 2	Set 3
9MC, 8SC, 7FC	9FC, 8MC, 7SC	9SC, 8FC, 7MC
9SS, 8FS, 7MS	9MS, 8SS, 7FS	9FS, 8MS, 8SS
9FG, 8MG, 7SG	9SG, 8FG, 7MG	9MG, 8SG, 7FG

Table 7.22: Data sets used as validation sets. S: cabernet sauvignon, F: cabernet franc, M: merlot, G:gravel, S: sand, C: clay, 7: 1997, 8: 1998, 9:1999

three different varieties of grape (cepage): cabernet sauvignon, cabernet franc and merlot noir; for three different years: 1997, 1998 and 1999; and for three different soils: clay, sand and gravel were collected. The reverse hieteronuclear 2D NMR (HMBC) was then measured from each total phenolic contents (TPC), a chemometric technique. We apply ANOVA to preprocess the data in [116], and reduce the dimension of the data to $p = 21938$. Since there are triple replicates for each kind of wine, we follow the splitting of the data in [116], listed in Table 7.22.

To map the categorical responses to the space in R, we map the classes onto the verteces of an equilateral triangle. We apply the standard PLS regression (denoted as PLS), the jointly sparse global SIMPLS regression (denoted as SPLS in the performance table) and the multi-class sparse SVM in Chapter V to the dataset. The multi-response PLS-R is modeled to predict all the 3 responses, cepage, soil and year, together, whereas the single response PLS-R builds a model for each response. Sparse SVM treats each of the response as a 3 class classification problem. Results are shown in Table 7.23 and 7.24.

method	Cepage error rate	Soil error rate	Year error rate	number of variables	number of components
PLS	0.037	0.111	0.074	21938	6
SPLS	0.037	0.124	0.136	4324	8.333

Table 7.23: Results of wine classification by multi-response PLS methods.

Cepage			
method	error rate	number of variables	number of components
PLS	0.037	21938	2.667
SPLS	0.074	787.333	3.667
sparse SVM	0.049	101.667	NA
Soil			
method	error rate	number of variables	number of components
PLS	0.111	21938	4
SPLS	0.222	433	3.667
sparse SVM	0.086	2443.3	NA
Year			
method	error rate	number of variables	number of components
PLS	0.074	21938	4
SPLS	0.124	5772	4
sparse SVM	0.074	1392.7	NA

Table 7.24: Results of wine classification by single-response PLS and SVM methods.

This experiment on agriculture data raises an interesting question about the use of PLS regression in categorical responses. In most of the comparisons, sparse SVM outperforms both the PLS-R and jointly sparse global SIMPLS-R. This may suggest that we should map the categorical responses to continuous responses in different ways, or design PLS for categorical responses. We discuss the extensions further in Chapter VIII.

7.9 Conclusion

The formulation of the SIMPLS objective function with an added group sparsity penalty greatly reduces the number of variables used to predict the response. This suggests that when multiple components are desired, the variable selection technique should take into account the sparsity structure for the same variables among all the components. Our proposed jointly sparse global SIMPLS algorithm is able to achieve as good or better performance with fewer predictor variables and fewer components as compared to competing methods. It is thus useful for performing dimension reduction and variable selection simultaneously in applications with large dimensional data but comparatively few samples ($n < p$).

The ADMM algorithm splits the optimization into the global SIMPLS eigen-decomposition problem and the soft-thresholding for sparsity constraints. This suggests that we can impose more complicated regularization tailored for each application and solve the optimization by ADMM. For example, in the chemometric application, the data is smooth over the wavelengths and we can apply wavelet shrinkage on the data or include a total variation regularization to encourage smoothness. The sparsity constraints can be imposed on the wavelet coefficients if wavelet shrinkage is applied, or together with the total variation regularization. The equivalence of soft wavelet shrinkage and total variation regularization was discussed in [117].

CHAPTER VIII

Conclusion and Future Work

This thesis consists of four main topics of high dimensional small sample size leaning motivated by learning from electrocardiograms: the binary classification with group structured sparsity constraint, the multi-class classification with sparsity constraint for variable selection, the uneven margin support vector machine for imbalanced learning, and the globally sparse partial least squares regression.

In Chapter II, we have shown the value of the electrograms from ICDs, and developed quantitative measurement to describe the spatial resolution of ECGs and EGMs. We have also demonstrated that automated learning of the origin of the VT is possible with ECGs, which can reduce the surgery time during ablation procedure. The work suggested that learning from these high dimensional data and sometimes with imbalanced sample sizes requires more sophisticated methods. The high dimensional data can be easily overfitted, and if the underlying model is sparse, then imposing sparsity constraints to the original statistical problem can be a solution to this problem.

Chapter III is a review of optimization for group structured variable selection, in which we have discussed the augmented Lagrangian and ADMM algorithm. These techniques enable us to develop algorithms for the sparse statistical learning in the chapters that follow. In Chapter IV, we proposed the SVM with group structured variable selection. Application to 3D cell Microscopy showed significant improve-

ments as compared with standard SVM without sparsity constraints. In Chapter V, we have formulated the optimization for multi-block multi-class classification with structured variable selection. Applying our algorithm developed based on augmented Lagrangian to predictive health problems, we have shown the benefit of structured variable selection, such as improving performance, using less predictor variables, and providing insights to the underlying model.

Chapter VI is a study of another aspect of small sample size issue: the imbalanced classification problems. Here, the sample size is small for one of the classes. The decision boundary of classical learning methods can be greatly skewed because the minority class is outnumbered by the majority class. We have reviewed the common strategies developed for these problems, and concluded that the α-classification calibrated loss function performs better than those that are not. The additional margin parameter provides the ability to adapt to the imbalanced data.

In Chapter VII, we again looked for sparse solution to supervised methods, the partial least squares regression. With structured variable selection technique, we were able to select the minimum set of predictor variables that are used for all the components. Improved performance is attained with less variables and components. This suggested standard PLS methods may face the problem of overfitting in the high dimensional data.

Some possible future work motivated by this thesis includes:

(1) Extension of PLS regression to categorical responses. In Chapter VII, we did experiments on continuous response (Octane data), ordinal responses (symptoms) and categorical responses (wine data). One would expect that classification techniques should perform better than regression on categorical data. Indeed, in the experiment on wine data, SVM performs better on the average than ordinary PLS. How to extend PLS methodology to categorical responses remains an open question. We propose developing methodology for probabilistic PLS discriminant analysis. Some relevant

work in the literature includes [118][119] [120][121]. Although PLS was designed for continuous response variables, the latent model successfully solve the colinear issues in high-dimensional data, which may be applicable to categorical responses. This will enable us to perform dimension reduction, variable selection and classification simultaneously. The difficulty is that by modeling the response variables by multinomial distribution, the probabilistic PLS does not reduce a simple eigen-decomposition as in the Gaussian case [118].

(2) Global SIMPLS algorithm. We solve the global SIMPLS optimization by greedy methods. One possible extension is to develop closer approximations for quadratic matrix programming [122] [123] [124]. The global SIMPLS optimization can be written as the same format in [124]. Beck was able to solve the quadratic matrix programming optimization with K constraints [124], whereas the global SIMPLS has $K + C_2^k$ constraints. Extending the approaches in [124] may improve the approximation. This will provide the technique to understand the difference between global SIMPLS and SIMPLS, and better solution for the jointly sparse global SIMPLS. Developing performance bounds for the greedy algorithm is another possibility to evaluate the algorithm.

(3) Multi-block PLS regression with structured variable selection. We formulated the global SIMPLS with joint sparsity constraints. The extension from two blocks to multi-blocks may be of interest, especially when integrating different biomedical data becomes important, [125]. However, the number of regularization parameter increases, and tunning the parameters becomes difficult. One possible solution is to select the regularization parameters by optimizing the objective function, instead of minimizing the prediction error, e.g., MSE.

(4) Convergence of ADMM methods. ADMM has been successfully applied to statistical learning problems with structured sparsity, which is important as data dimension increases. As discussed in Chapter III, the linear convergence rate is guar-

anteed for strong convex functions. Studying the conditions under which convergence is guaranteed, and the rate of convergence will be an important topic. This will provide useful guidelines for whether solving the optimization with ADMM methods.

(5) Prescreening in multi-class classifications. We noticed that prescreening the variables by pairwise binary classifiers for the multi-class classification improved the performance. This was true for both simulation and real data experiments in Chapter V. A theoretical work can be motivated by this observation. A possible direction is to modify the generalized hinge loss function for multi-class SVMs.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

[2] C. Ducroz, J. Olivo-Marin, and A. Dufour. Spherical harmonics based extraction and annotation of cell shape in 3d time-lapse microscopy sequences. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 6619–6622. IEEE, 2011.

[3] K. Crammer and Y. Singer. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *J. Machine Learning Research*, 2:265–292, 2002.

[4] T. Hastie, R. Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[5] H. H. Zhang, Y. Liu, Y. Wu, and J. Zhu. Variable selection for the multicategory SVM via adaptive sup-norm regularizaion. *Electronic Journal of Statistics*, 2:149–167, 2008.

[6] I.T. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

[7] H. Wold. Soft modeling by latent variables; the non-linear iterative partial least squres appraoch. In J. Gani, editor, *Perspectives in Probability and Statistics*, pages 117–142. Adademic Press, London, 1975.

[8] S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the pls method. *Lecture Notes in Mathematics*, 973:286–293, 1983.

[9] H. Chun and S. Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J.R. Statist. Soc.B*, 72:3–25, 2010.

[10] Y. Li and J. Shawe-Taylor. The svm with uneven margins and chinese document categorization. In *Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC17)*, pages 216–227. MIT Press, 2003.

[11] C. Yang, J. Yang, and J. Wang. Margin Calibration in SVM class-imbalanced learning. *Neurocomputing*.

[12] C.X. Ling and C. Li. Data Mining for Direct Marketing: Problems and Solutions. *International Conference on Knowledge Discovery and Data Mining*, 1998.

[13] F. Vilariño, P. Spyridonos, J. Vitrià, and P. Radeva. Experiments with svm and stratified sampling with an imbalanced problem: Detection of intestinal contractions. *Pattern Recognition and Image Analysis*, pages 783–791, 2005.

[14] M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 179–186. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[15] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[16] Y. Li, K. Bontcheva, and H. Cunningham. Using uneven margins svm and perceptron for information extraction. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 72–79. Association for Computational Linguistics, 2005.

[17] W.S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 448, 2003.

[18] H. Masnadi-Shirazi and N. Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive svms. In *Proceedings of the International Conference on Machine Learning*, pages 204–213, 2010.

[19] C. Scott. Calibrated surrogate losses for classification with label-dependent costs. *arXiv preprint arXiv:1009.2718*, 2010.

[20] J.E. Poole, G.W. Johnson, A.S. Hellkamp, J. Anderson, D.J. Callans, M.H. Raitt, R.K. Reddy, F.E. Marchlinski, R. Yee, T. Guarnieri, et al. Prognostic importance of defibrillator shocks in patients with heart failure. *New England Journal of Medicine*, 359(10):1009–1017, 2008.

[21] A.J.J. Wood and F. Morady. Radio-frequency ablation as treatment for cardiac arrhythmias. *New England Journal of Medicine*, 340(7):534–544, 1999.

[22] D.P. Zipes, J.P. Dimarco, P.C. Gillette, W.M. Jackman, R.J. Myerburg, S.H. Rahimtoola, J.L. Ritchie, M.D. Cheitlin, A. Garson, R.J. Gibbons, et al. Guidelines for clinical intracardiac electrophysiological and catheter ablation procedures: a report of the american college of cardiology/american heart association task force on practice guidelines (committee on clinical intracardiac electrophysiologic and catheter ablation procedures), developed in collaboration with the north american society of pacing and electrophysiology. *Journal of the American College of Cardiology*, 26(2):555–573, 1995.

[23] R.M. Berne and M.N. Levy. *Physiology*. Mosby, St. Louis, 1983.

[24] W.G. Stevenson and E. Delacretaz. Radiofrequency catheter ablation of ventricular tachycardia. *Heart*, 84(5):553–559, 2000.

[25] F. Bogun, M. Taj, M. Ting, H.M. Kim, S. Reich, E. Good, K. Jongnarangsin, A. Chugh, F. Pelosi, H. Oral, et al. Spatial resolution of pace mapping of idiopathic ventricular tachycardia/ectopy originating in the right ventricular outflow tract. *Heart Rhythm*, 5(3):339–344, 2008.

[26] J.M. Miller, F.E. Marchlinski, A.E. Buxton, and M.E. Josephson. Relationship between the 12-lead electrocardiogram during ventricular tachycardia and endocardial site of origin in patients with coronary artery disease. *Circulation*, 77(4):759–766, 1988.

[27] 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.

[28] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[29] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[30] Z.Q. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

[31] T. Goldstein, B. O'Donoghue, and S. Setzer. Fast alternating direction optimization methods. *CAM report*, pages 12–35, 2012.

[32] E. Esser. Applications of lagrangian-based alternating direction methods and connections to split bregman. *CAM report*, 9:31, 2009.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[34] J. Eckstein and D.P. Bertsekas. On the douglasrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[35] M. Figueiredo, J.M. Bioucas-Dias, and M.V. Afonso. Fast frame-based image deconvolution using variable splitting and constrained optimization. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 109–112. IEEE, 2009.

[36] V. Vapnik. *The nature of statistical learning theory*. springer, 1999.

[37] S.G. Wilson. *Digital modulation and coding*. Prentice-Hall, Inc., 1995.

[38] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

[39] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

[40] Y.J. Lee and O.L. Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational optimization and Applications*, 20(1):5–22, 2001.

[41] T. Zhou, D. Tao, and X. Wu. Nesvm: a fast gradient method for support vector machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 679–688. IEEE, 2010.

[42] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[43] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[45] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[46] G.M. Fung and O.L. Mangasarian. A feature selection newton method for support vector machine classification. *Computational optimization and applications*, 28(2):185–202, 2004.

[47] O.L. Mangasarian. Exact 1-norm support vector machines via unconstrained convex differentiable minimization. *Journal of Machine Learning Research*, 7(2):1517–1530, 2006.

[48] G.X. Yuan, K.W. Chang, C.J. Hsieh, and C.J. Lin. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *The Journal of Machine Learning Research*, 9999:3183–3234, 2010.

[49] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(2):1391, 2005.

[50] M.Y. Park and T. Hastie. *Regularization path algorithms for detecting gene interactions*. Department of Statistics, Stanford University, 2006.

[51] H. Zou. An improved 1-norm svm for simultaneous classification and variable selection. In *Eleventh international conference on artificial intelligence and statistics*. Citeseer, 2007.

[52] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2004.

[53] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[54] L. Wang, J. Zhu, and H. Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589, 2006.

[55] H.H. Zhang, J. Ahn, X. Lin, and C. Park. Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1):88–95, 2006.

[56] Y. Liu, H. Helen Zhang, C. Park, and J. Ahn. Support vector machines with adaptive lq penalty. *Computational Statistics & Data Analysis*, 51(12):6380–6394, 2007.

[57] Y. Liu and Y. Wu. Variable selection via a combination of the l 0 and l 1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

[58] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

[59] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[60] C.J. Hsieh, K.W. Chang, C.J. Lin, S.S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on Machine learning*, volume 951, pages 408–415, 2008.

[61] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[62] A.T. Puig, A. Wiesel, and A.O. Hero. A multidimensional shrinkage-thresholding operator. In *Statistical Signal Processing, 2009. SSP'09. IEEE/SP 15th Workshop on*, pages 113–116. IEEE, 2009.

[63] R. Ananthakrishnan and A. Ehrlicher. The forces behind cell movement. *Int. J Biol Sci.*, 3(5).

[64] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3D shape descriptors. *Proc. Eurographics Symp. on Geometry Process.*, 2003.

[65] R.J. Morris, R.J. Najmanovich, A. Kahraman, and J.M. Thornton. Real spherical harmonic expansion coefficients as 3d shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, 21(10):2347–2355, 2005.

[66] K. Khairy, J. Foo, and J. Howard. Shapes of Red Blood Cells: Comparison of 3D Confocal Images with the Bilayer-couple Model. *Cell. and Mol. Bioeng.*, 1:173–181, 2008.

[67] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. and Machine Intell.*, 18(6):607–616, 1996.

[68] R. Thibeaux, A. Dufour, P. Roux, M. Bernier, A.C. Baglin, P. Frileux, J.C. Olivo-Marin, N. Guillén, and E. Labruyère. Newly visualized fibrillar collagen scaffolds dictate entamoeba histolytica invasion route in the human colon. *Cellular Microbiology*, 2012.

[69] H.H. Zhang, Y. Liu, Y. Wu, and J. Zhu. Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, 2:149–167, 2008.

[70] K. Wolf, S. Alexander, V. Schacht, L.M. Coussens, U.H. von Andrian, J. van Rheenen, E. Deryugina, and P. Friedl. Collagen-based cell migration models in vitro and in vivo. In *Seminars in cell & developmental biology*, volume 20, pages 931–941. Elsevier, 2009.

[71] D. Shao, W.J. Rappel, and H. Levine. Computational model for cell morphodynamics. *Physical review letters*, 105(10):108104, 2010.

[72] F. de Chaumont, S. Dallongeville, N. Chenouard, N. Hervé, S. Pop, T. Provoost, V. Meas-Yedid, P. Pankajakshan, T. Lecomte, Y. Le Montagner, et al. Icy: an open bioimage informatics platform for extended reproducible research. *Nature Methods*, 9(7):690–696, 2012.

[73] A. Dufour, R. Thibeaux, E. Labruyere, N. Guillen, and J.C. Olivo-Marin. 3-d active meshes: Fast discrete deformable models for cell tracking in 3-d time-lapse microscopy. *Image Processing, IEEE Transactions on*, 20(7):1925–1937, 2011.

[74] M.S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. *Advances in multiresolution for geometric modelling*, pages 157–186, 2005.

[75] L. Shen and F. Makedon. Spherical mapping for processing of 3D closed surfaces. *Image and Vision Computing*, 24(7):743–761, 2006.

[76] S. Knerr, L. Personnaz, G. Dreyfus, J. Fogelman, A. Agresti, M. Ajiz, A. Jennings, F. Alizadeh, F. Alizadeh, J. Haeberly, et al. Single-layer learning revisited: a stepwise procedure for building and training a neural network. *Optimization Methods and Software*, 1:23–34, 1990.

[77] U.H.G. Kreßel. Pairwise classification and support vector machines. In *Advances in kernel methods*, pages 255–268. MIT Press, 1999.

[78] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the seventh European symposium on artificial neural networks*, volume 4, pages 219–224, 1999.

[79] E.J. Bredensteiner and K.P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1):53–79, 1999.

[80] Y. Guermeur. Combining discriminant models with new multi-class svms. *Pattern Analysis & Applications*, 5(2):168–179, 2002.

[81] Y. Liu and X. Shen. Multicategory ψ-learning. *Journal of the American Statistical Association*, 101(474):500–509, 2006.

[82] L. Wang and X. Shen. On l 1-norm multiclass support vector machines. *Journal of the American Statistical Association*, 102(478):583–594, 2007.

[83] lldiko E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[84] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[85] L. Wang and X. Shen. On $l1 - norm$ Multi-class Support Vector Machines: Methodology and Theory. *Journal of the American Statistical Association*.

[86] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[87] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[88] D.P. Bertsekas. *Nonlinear Programming*. Cambridge, MA.

[89] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

[90] S.S. Keerthi, S. Sundararajan, K.W. Chang, C.J. Hsieh, and C.J. Lin. A sequential dual method for large scale multi-class linear svms. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 408–416. ACM, 2008.

[91] T.D. Querec, R.S. Akondy, E.K. Lee, W. Cao, H.I. Nakaya, D. Teuwen, A. Pirani, K. Gernert, J. Deng, B. Marzolf, et al. Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature immunology*, 10(1):116–125, 2008.

[92] H.I. Nakaya, J. Wrammert, E.K. Lee, L. Racioppi, S. Marie-Kunze, W.N. Haining, A.R. Means, S.P. Kasturi, N. Khan, G.M. Li, et al. Systems biology of vaccination for seasonal influenza in humans. *Nature immunology*, 12(8):786–795, 2011.

[93] Y. Huang, A.K. Zaas, A. Rao, N. Dobigeon, P.J. Woolf, T. Veldman, N.C. Øien, M.T. McClain, J.B. Varkey, B. Nicholson, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS genetics*, 7(8):e1002234, 2011.

[94] R. Chen, G.I. Mias, J. Li-Pook-Than, L. Jiang, H.Y.K. Lam, R. Chen, E. Miriami, K.J. Karczewski, M. Hariharan, F.E. Dewey, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.

[95] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 2004.

[96] I. Steinwart. Consistency of Support Vector Machines and Other Regularized Kernel Classifiers. *IEEE Transactions on Information Theory*, 51(1), 2005.

[97] H. Martens and M. Martens. Validation of pls regression models in sensory science by extended cross-validation. *PLS'99*, 1999.

[98] D. Rossouw, C. Robert-Granié, and P. Besse. A sparse pls for variable selection when integrating omics data. *Genetics and Molecular Biology*, 7(1):35, 2008.

[99] H. Chun and S. Keleş. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*, 182(1):79–90, 2009.

[100] D. Chung, S. Keles, et al. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1):17, 2010.

[101] H. Wold. Nonlinear estimation by iterative least squares procedures. In *Research papers in statistics*. Wiley: New York, 1966.

[102] H. Hotelling. Analysis of a complex of statistical variables into components. *Journal of Educational Psychology*, 24, 1933.

[103] S. Wold, A. Ruhe, H. Wold, and W. Dunn. The collinearity problem in linear regression. the PLS approach to generalised inverses. *Journal of Scientific Statistical Computing, SIAM*, 5:735–743, 1984.

[104] M. Tenenhaus. *La Régression PLS: théorie et pratique*. Ed. Technip, Paris, 1998.

[105] S. De Jong. PLS shrinks. *Journal of Chemometrics*, 9 (4):323–326, 1995.

[106] A. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, pages 32–44, 2006.

[107] A. Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1988.

[108] S. de Jong. Simpls: an alternative approach to partial least squares regression. *Chemom Intell Lab Syst*, pages 251–263, 1993.

[109] C.J.F. ter Braak and S. de Jong. The objective function of partial least squares regression. *Journal of Chemometrics*, pages 41–54, 1998.

[110] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *J. Computnl Graph. Statist.*, pages 531–547, 2003.

[111] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

[112] W. Gander, Gene. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear algebra and its applications*, 1989.

[113] A. Amir Beck and M. Teboulle. Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares. *SIAM Journal on Matrix Analysis and Applications*, pages 425–445, 2006.

[114] P. Bastien, V.E. Vinzi, and M. Tenenhaus. Pls generalised linear regression. *Computational Statistics & Data Analysis*, 48(1):17–46, 2005.

[115] J. Magidson. Correlated component regression: A prediction/classification methodology for possibly many features. In *Proceedings of the American Statistical Association*, 2010.

[116] S. Masoum, D.J.R. Bouveresse, J. Vercauteren, M. Jalali-Heravi, and D.N. Rutledge. Discrimination of wines based on 2d nmr spectra using learning vector quantization neural networks and partial least squares discriminant analysis. *Analytica chimica acta*, 558(1):144–149, 2006.

[117] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk. On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. *SIAM Journal on Numerical Analysis*, 42(2):686–713, 2004.

[118] W. Buntine and A. Jakulin. Discrete component analysis. *Subspace, Latent Structure and Feature Selection*, pages 1–33, 2006.

[119] G. Russolillo and C.N. Lauro. A proposal for handling categorical predictors in pls regression framework. *Classification and Multivariate Analysis for Complex Data Structures*, pages 343–350, 2011.

[120] W. Buntine. Variational extensions to em and multinomial pca. *Machine Learning: ECML 2002*, pages 23–34, 2002.

[121] U. Thissen, M. Pepers, B. Üstün, WJ Melssen, and LMC Buydens. Comparing support vector machines to pls for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*, 73(2):169–179, 2004.

[122] K. Anstreicher and H. Wolkowicz. On lagrangian relaxation of quadratic matrix constraints. *SIAM Journal on Matrix Analysis and Applications*, 22(1):41–55, 2000.

[123] P.H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[124] A. Beck. Quadratic matrix programming. *SIAM Journal on Optimization*, 17(4):1224–1238, 2007.

[125] A. Tenenhaus. Variable selection for generalized regularized canonical correlation analysis. In *7th International Conference on Partial Least Squares and Related Methods*, 2012.

ABSTRACT

Statistical Learning for Sample-Limited High-dimensional Problems with Application to Biomedical Data

by

Tzu-Yu Liu

Chair: Alfred O. Hero

Co-Chair: Clayton D. Scott

With advancing technology comes the need to extract information from increasingly high-dimensional data, whereas the number of samples is often limited or even acquired from imbalanced populations. This thesis develops strategies for classification and prediction in high-dimensional but poorly sampled problems arising in computational biology and medicine. These strategies are presented in 6 chapters. In Chapter II Support Vector Machine (SVM) classifiers are applied to localizing ventricular tachycardia from electrocardiographical data. In Chapters III, IV, V and VII optimization-driven structured sparsity algorithms are developed. In Chapter VI a class of uneven margin SVMs is proposed for learning binary classifiers with imbalanced training populations.

The major part of this thesis is focused on group structured sparsity constrained statistical learning for sample-limited high-dimensional problems. Variable selection consists of reducing the dimension to a few important variables that contain most of the information necessary for discriminating between classes or for prediction of continuous responses. This can potentially avoid overfitting problems, improve gen-

eralizability of the predictors and provide better interpretation. Novel algorithms based on the augmented Larangian and ADMM methods are developed for various statistical learning problems with group structured sparsity penalty: binary SVMs with application to 3D cell microscopy data to discover important shape information for characterizing highly deformable cells; multi-class SVMs with application to gene expression analysis to improve disease prediction rate and control irrelevant patient variations; PLS regression with application to chemometrics, medicine, and agriculture applications. These applications demonstrate the benefit of sparsity constrained optimization approaches to high dimensional problems with limited data.