

DESIGN AND ANALYSIS ISSUES
IN HIGH DIMENSION, LOW SAMPLE SIZE PROBLEMS

by

SANDRA ESI SAFO

(Under the Direction of Jeongyoun Ahn and Kevin K. Dobbin)

ABSTRACT

Advancement in technology and computing power have led to the generation of data with enormous amount of variables when compared to the number of observations. These types of data, also known as high dimension, low sample size, are plagued with different challenges that either require modifications of existing traditional methods or development of new statistical methods. One of these challenges is the development of Sparse methods that use only a fraction of the variables. Sparse methods have been shown to perform better at making predictions on real high dimensional problems, hence justifying their studies and use in practice. This dissertation considers three novel methods for designing and analyzing high dimensional studies. We first propose new sample size method to estimate the number of samples required in a training set when allocating new entities into two groups. The methodology exploits the structural similarity between logistic regression prediction and errors-in-variables models. Secondly, we consider the problem of assigning future observations to known classes using linear discriminant analysis. We propose a new classification approach of generalizing existing binary linear discriminant methods to multi-class methods. Our methodology utilizes the equivalence between discriminant subspace using Fisher's

linear discriminant analysis and basis vectors of between class scatter. We apply the proposed method to two sparse methods. Thirdly, a general framework that results in sparse vectors for many multivariate statistical methods is developed. The framework uses the relationship between many multivariate statistical problems and generalized eigen value problem. We illustrate this framework with two multivariate statistical methods- linear discriminant analysis for classifying new entities into more than two groups, and canonical correlation analysis for studying associations between two different high dimensional data types. The effectiveness of the proposed methods in this dissertation is evaluated by various simulated processes and real data analyses on microarray and RNA sequencing (RNA-seq) data.

INDEX WORDS: High dimensional data; Sample size; Lasso; Classification; Regularized logistic regression; Conditional score; Measurement error; Linear discriminant analysis; Multi-class discrimination; Singular value decomposition; Sparse discrimination; Generalized eigenvalue problem; Sparse canonical correlation analysis

DESIGN AND ANALYSIS ISSUES
IN HIGH DIMENSION, LOW SAMPLE SIZE PROBLEMS

by

SANDRA ESI SAFO

B.A., University of Ghana, 2006

M.Sc., University of Akron, 2009

M.S., University of Georgia, 2011

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

© 2014

Sandra Esi Safo

All Rights Reserved

DESIGN AND ANALYSIS ISSUES
IN HIGH DIMENSION, LOW SAMPLE SIZE PROBLEMS

by

SANDRA ESI SAFO

Approved:

Major Professors: Jeongyoun Ahn
Kevin K. Dobbin

Committee: Nicole Lazar
Jaxk Reeves
Xiao Song

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

DEDICATION

To God

To my husband Kwadwo

To my son Nathan

To my mom Eva and dad Agya Addo

To my sister Adjoa and brother Kobbie

To my relatives and friends

You have made this possible. Love you all.

ACKNOWLEDGMENTS

I thank the Almighty God and my Lord Jesus for the strength and wisdom to complete this dissertation. My most sincere gratitude go to my major professors, Dr. Jeongyoun Ahn and Dr. Kevin K. Dobbin for their immense wisdom and guidance, encouragement, and support throughout the period of my dissertation research. Their patience and kindness during the entire period especially the time of my pregnancy goes beyond the call of duty.

I also want to express my indebtedness to my committee members, Dr. Nicole Lazar, Dr. Jaxk Reeves, and Dr. Xiao Song for the time spent in reading my work. I truly appreciate your excellent and constructive questions, as well as insightful recommendations that helped in the writing of this dissertation. My appreciation also go to Lily Wang, the faculty, staff and graduate students at the Department of Statistics.

Special thanks are due to my mom, dad and siblings for their constant encouragement and support throughout the period of my graduate studies. Finally, I want to say a big thank you to my husband, Kwadwo, for being my rock and shoulder in my darkest times. His daily love and support, more especially during the birth of our son, has helped me in completing this dissertation in a timely manner; I could not have done this without him.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
 CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
1.1 INTRODUCTION	1
1.2 LITERATURE REVIEW	4
1.3 OUTLINE OF DISSERTATION	22
1.4 REFERENCES	24
2 SAMPLE SIZE DETERMINATION FOR REGULARIZED LOGISTIC REGRESSION- BASED CLASSIFICATION	29
2.1 INTRODUCTION	30
2.2 METHODS	34
2.3 SIMULATION STUDIES	43
2.4 REAL DATASET ANALYSES	48
2.5 DISCUSSION	51
2.6 REFERENCES	54
3 GENERAL SPARSE MULTI-CLASS LINEAR DISCRIMINANT ANALYSIS	59
3.1 INTRODUCTION	60

3.2	Sparse Multi-class Linear Discrimination	62
3.3	Empirical Studies	67
3.4	Discussion	78
3.5	References	79
4	Sparse Analysis for High Dimensional Data	82
4.1	Introduction	83
4.2	The Substitution Method	84
4.3	Substitution for Sparse Linear Discriminant Analysis . .	88
4.4	Substitution for Sparse Canonical Correlation Analysis	89
4.5	Discussion	116
4.6	References	118
5	Conclusion	121
6	Supplement	125
	Bibliography	146

LIST OF FIGURES

2.2.1 Summary of results of simulations	47
3.3.1 Basis simulation test errors	69
3.3.2 Settings I and II variable selection properties of basis methods compared . .	70
3.3.3 Settings III and IV variable selection properties of basis methods compared .	71
3.3.4 Class boundary of LDA on real dataset	76
4.4.1 Estimated maximum canonical correlation coefficient	103
4.4.2 Variables selected by α in Settings I and II	104
4.4.3 Variables selected by α in Settings III and IV	105
4.4.4 Variables selected by β in Settings I and II	106
4.4.5 Variables selected by β in Settings III and IV	107
4.4.6 Matthew's correlation coefficients	108
4.4.7 Distribution of genes and copy number variations for breast cancer data . . .	112
4.4.8 Distribution of gene expressions selected on chromosome one	113
4.4.9 Estimated gene expression variates vs CNV variates	114
4.4.10 CNV canonical vectors compared	114
6.0.1 Simulated datasets results	126
6.0.2 Nested cross-validation procedure	127
6.0.3 Converting between a logistic slope and the classification error rate	128

LIST OF TABLES

2.2.1 Estimates of the asymptotic slope β_∞ and corresponding accuracy acc_∞ . . .	42
2.2.2 Evaluation of the sample size estimates from AR(1) and identity covariances	43
2.2.3 Clinical covariate simulations	44
2.2.4 Table of sample size estimates	45
2.2.5 Resampling studies	46
3.3.1 Microarray datasets summary statistics	74
3.3.2 Samples per class for microarray and RNA-seq datasets	74
3.3.3 Classification accuracy of basis methods compared	75
3.3.4 Variable selection of basis methods on real datasets compared	75
4.4.1 Comparison of substitution method to PMD	113
6.0.1 Mixture normal simulations results	130
6.0.2 Comparison of LC and EIV asymptotic slope	130
6.0.3 Evaluation of the sample size estimates from identity covariance	131
6.0.4 Evaluation of the sample size estimates from CS covariance	131
6.0.5 Resampling studies	132
6.0.6 Effect size for different covariance structures and number of informative features	140

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Modern technology and computing power have facilitated the generation of data that deviate from the typical. The typical data have the number of samples or observations, n , far more than the number of variables or features, p . We are now able to collect data with gigantic number of variables in relation to a small number of observations studied. Such data types are known as high dimension, low sample size (HDLSS). HDLSS (also referred to as “large p , small n ”) problems arise in many fields. For instance, in genomics, a single DNA microarray experiment produces tens of thousands of gene expressions compared to the few samples (usually in the hundreds) in the study. Data from image and text analyses have enormous amount of variables compared to the observations. Many more examples of HDLSS data are found in Donoho (2000), Hastie et al. (2009) and Dudoit et al. (2002).

These types of data are plagued with different challenges in their analyses that either require modifications of the existing traditional methods or development of new statistical methods. Many of the available low dimensional methods that rely on multivariate analysis break down here. These methods were developed under the assumption of $p \ll n$ and one of the main reasons for their breakdown is that there are not enough samples to better estimate the underlying covariance structure. Considering the colossal nature of the variables, a statistical challenge facing the statistics community is the development of methodologies that use some of the many high dimensional variables. These methodologies, referred to

as Sparse methods have been shown to perform better at making predictions on real high dimensional problems, and hence worthy of being studied and used in practice.

In the analyses of HDLSS data, if there are class labels with two or more distinct classes, one may be interested in how best to allocate new samples into already existing classes using only a fraction of the variables. A rule for the assignment of an unclassified entity to one of two or more groups is known as a discriminant or allocation rule, and the process of allocating new entities into already existing class is known as classification. In other fields, such a rule is called a classifier. Discriminant analysis is used, for instance, in medical diagnosis to assign a new patient into one of two or more existing disease classes before a microscopic examination for the actual cause of a disease. The development of sparse classifiers have been shown to be statistically and clinically relevant. A critical question in using these classifiers is whether a better classifier can be developed from a larger training set size and, if so, how large the training set should be. This is important because of the costly and often times complicated clinical procedures in obtaining additional samples, making it useful to estimate the performance of a classifier for future larger training set sizes. Also, conclusions on disease classification made on an inadequate sample size may be statistically unsound and medically aggravating. The design part of the dissertation examines this sample size question via regularized logistic regression classifiers and errors-in-variables models.

The analysis part of the dissertation considers several sparse multivariate methods in HDLSS. In the study of HDLSS problems, if there are one or more different HDLSS data types for the same set of samples, an interest may be in the individual analysis of each data type or in the joint study of the different data types. Individual analysis of a HDLSS data type with class labels via discriminant analysis will yield directions of maximal separation among the classes. This can further be used for classification of future observations into one of the existing classes. For different HDLSS data types on the same observations, joint analysis of these data are becoming more prevalent since researchers are able to measure more than one

different high dimensional data types on the same set of samples. As an example, biologists may obtain DNA methylation data, structural variation data (e.g copy number changes) and gene expression data on the same set of patients. Each type of data assayed provides them with a different snapshot of the biological system of the samples. Instead of studying the data individually, they may perform integrative analysis of the different data types to identify interactions or associations between them. For instance, joint study of copy number changes and gene expression data may reveal regions where the DNA of a patient have been amplified or deleted and which contribute to the development and progression of disease genomes. In this part of the dissertation, we first propose a methodology that can only be used in individual modeling of high dimensional data where the goal is to classify new entities into already existing classes. A new method of generalizing binary problems into multi-class is proposed. Secondly, a unified approach of obtaining sparse solution vectors to individually or jointly model different high dimensional data types through linear discriminant analysis and canonical correlation analysis respectively is proposed.

A motivating example to be used throughout the dissertation will be the allocation of the *Drosophila melanogaster* (Fly) data of Graveley et al. (2011) into existing stages, using the characteristic variables associated with each fly. The fly dataset is comprised of about 22,000 gene expression measurements on $n = 147$ flies that are grouped into four classes. Class 1 consists of all embryos; Class 2 consists of all larvae; Class 3 consists of all white prepupae and Class 4 consists of all adult flies. The allocation problem can be generally stated as follows: for K known classes, given a new sample with gene measurements, we want to be able to predict with much accuracy what stage the fly belongs. For the design issue, we regroup the four classes into two classes: Class 1 consists of all the embryos and some adult and white prepupae (WPP); Class 2 consists of all the larvae and a mix of adults and WPP, and state the problem as follows: for $K = 2$ known classes, can a better classifier be developed using regularized logistic regression and a larger training set, and if so, how

large should the training set be? Our sample size method will be used to study the adequacy of the training samples used in the original study.

1.2 LITERATURE REVIEW

1.2.1 SAMPLE SIZE DETERMINATION FOR REGULARIZED LOGISTIC REGRESSION-BASED CLASSIFICATION

The objective of sample size requirements in training classifiers has not received much attention in the literature. Only some few papers focus on this objective, but most of these works were either developed under low dimensional settings or imposed some distributional assumptions on the data (Lachenbruch, 1968; Dobbin and Simon, 2007). Interestingly enough, there is only one paper in the literature (Mukherjee et al., 2003) for classifier development in HDLSS that does not make any assumptions on the distribution of the high dimensional variables. However, the approach proposed by the authors uses parametric learning curves in estimating sample size requirements.

Lachenbruch (1968) pioneered the development of sample size estimates for classifier training. His sample size method was developed under multivariate normal theory in the low dimensional setting. For two classes with samples n_1 and n_2 respectively, and using discriminant function, he asked the question “How large should n_1 and n_2 be for the discriminant function to have an error rate within η of the optimum value?” The objective function under consideration was the difference between the expected error rate and the optimal (also known as Bayes) error rate. He showed that the sample size needed depended on the number of variables, the desired tolerance, η , and Mahalanobis distance between the class means.

Mukherjee et al. (2003) developed a general sample size method that can be used with different classification functions. Their method is based on resampling repeatedly from a pilot dataset, and using learning curves to study classification accuracy as a function of

training set size. They considered the inverse power-laws model $e(n) = an^{-\alpha} + b$, where $e(n)$ is the expected error rate, a , the learning rate, α , the decay rate and b the Bayes error rate, which is the minimum achievable error rate. The model parameters were estimated through a minimization procedure and the desired sample size was extrapolated or interpolated to achieve a desired error rate.

Dobbin and Simon (2007) used parametric probability modelling for the required sample size. They considered the objective function similar to Lachenbruch (1968) to develop the sample size required so that the expected probability of correct classification (PCC) was within some tolerance of the optimal probability of correct classification. Their approach depended heavily on mathematical simplifications and assumptions. Despite the simplicity of their approach, a potential draw back is that the expected PCC could vary highly especially when the sample size used in the training set is very small, which may affect sample size estimates. Also, the tuning parameter is not dealt adequately in Dobbin and Simon (2007) since they assume pre-specification.

In this dissertation, we develop a sample size method that is appropriate for studies using high dimensional data with corresponding binary response vector, with the goal of developing a classification rule to predict class labels or memberships. The sample size is chosen so that when the classifier is applied on an out of sample data from the same population, it will produce an expected logistic slope that is within a user-specified tolerance of the slope of an optimally trained classifier. Our definition of optimal is the slope of the logistic regression classifier or misclassification error as $n \rightarrow \infty$. The expected logistic slope is the average of the logistic slopes arising from using finite training sets. Using a similar objective function as in Lachenbruch (1968), we ask the question “can we find sample size estimate that will ensure that the expected logistic slope is within some τ of the optimal logistic slope?” We treat the difference between these slopes as prediction error. This prediction error is analogous to measurement error and we use errors-in-variables techniques to recover the sample size

estimates. The regularized logistic regression classifier is used in our sample size method because measurement error techniques are more widely developed for logistic regression than other classifiers such as discriminant analysis, support vector machines and many more. Also, logistic regression is used more often in biostatistical applications where binary responses occur quite frequently.

Our sample size method has the following advantages. The method does not assume a multivariate normal model for the data and therefore can be used for a more general population. It does not use parametric learning curve extrapolation to estimate the asymptotic model performance, as in Mukherjee et al. (2003). Instead, asymptotic performance is estimated directly using errors-in-variables regression methods on a pilot dataset. Unlike Dobbin and Simon (2007) where the tuning parameter is specified a priori to model building, in the proposed sample size method, feature selection is incorporated into model building and is data driven.

A brief review of measurement error models follows next. A more extensive review is found in the Supplement.

1.2.2 MEASUREMENT ERROR MODELS

Many areas in Statistics have models that are defined in terms of some variables, say X , that are sometimes not directly observable or accurately ascertainable. For example, blood pressure in cardiovascular disease studies are typically subject to measurement error because of imperfect instruments. In such instances, we obtain substitutes, say W , which is a measurement of the true value of X . Substituting W for X can complicate the statistical analysis of the observed data when inferences need to be made about a model defined in terms of X , and therefore ignoring the measurement error can lead to substantial bias (Carroll et al., 2006). Under the classical additive measurement error model, instead of X , one observes the

model

$$W = X + U, \tag{1.2.1}$$

where the observed variable is the true variable plus measurement error. Here, W is an unbiased measure of X and $E(U|X) = 0$. The error structure for U is either homoscedastic or heteroscedastic.

Several errors-in-variables methods have been developed for logistic regression models, including simulation extrapolation, SIMEX (Cook and Stefanski, 1994), conditional score (Stefanski and Carroll, 1987), corrected score (Nakamura, 1990) and its variant approximate corrected score (Novick and Stefanski, 2002), consistent functional methods (Huang and Wang, 2001), projected likelihood ratio (Hanfelt and Liang, 1995), and quasi-likelihood (Hanfelt and Liang, 1995).

SIMEX is a simulation and extrapolation based method for parameter estimation when the measurement error variance is known or can be estimated. In the simulation step of SIMEX, the method adds additional independent measurement errors in increasing order to the existing data, W and computes estimates from the contaminated data. In the extrapolation step, one extrapolates back to the case of no measurement error. The key step in the SIMEX method is the extrapolation step. Cook and Stefanski (1994) showed that, under some fairly general conditions, one may find a function of the measurement errors, that when extrapolated to the case of no measurement error, the true parameter is obtained approximately. This function is however, not often known and is usually estimated using a fitted polynomial regression model, thus making SIMEX an approximately consistent method.

Since SIMEX results in approximately consistent estimators, Stefanski and Carroll (1987) subsequently proposed a fully consistent estimator, the conditional scores function, for normal errors in covariates. In this method, one conditions the response vector on a sufficient statistics to eliminate nuisance parameters arising from the error prone variables, resulting in consistent estimators. It was shown that the estimating function which, is a

solution to the conditional log-likelihood of the response given the sufficient statistic, was unbiased in the presence of measurement error. The authors noted that even though the estimating function is unbiased, multiple zero-crossings may exist and that they are not guaranteed to converge.

Nakamura (1990) proposed the corrected score approach to correct for the effect of measurement error on score functions without making any distributional assumptions about the true covariates. The method finds a score function of the observed data which is unbiased for the true-data score function; finding the corrected score function is mathematically challenging. Novick and Stefanski (2002) proposed a Monte Carlo simulation based method, the approximate corrected score function, to deal with models that did not yield exact corrected score functions. The simulation based method however, is computationally expensive, and requires programming software with complex number capabilities (Carroll et al., 2006).

Noting the distributional assumption about the measurement errors and convergence issues for conditional scores, Huang and Wang (2001) proposed the consistent functional methods to eliminate bias when variables are measured with error. The key idea here is to find a correction-amenable estimating function when there is no measurement error, and then construct parametric and nonparametric correction estimation functions when measurement error is present. The consistent functional method is most valuable in large-scale studies (Huang and Wang, 2001), which are currently very rare in high dimensional data from sequencing or microarrays. Hanfelt and Liang (1995) also proposed likelihood based methods—quasi-likelihood and projected likelihood methods, to accompany conditional scores method (Stefanski and Carroll, 1987). These methods serve to eliminate multiple root issues, and hence convergence problems from using estimating functions. The authors noted that projected likelihood may not always exist, and in particular did not exist for logistic regression models.

We note that conditional scores method are computationally tractable and relatively easy to implement, and have shown good finite-sample performance (Carroll et al., 2006). A detailed description of the conditional scores function is found in the Supplement.

1.2.3 MULTI-CLASS LINEAR DISCRIMINANT ANALYSIS

Discriminant analysis is popularly used for many classification problems. Fisher's (Fisher, 1936) linear discriminant analysis (LDA) is a popular method of choice for discriminant analysis. Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, $\mathbf{x} \in \mathbb{R}^p$ be a $p \times n$ data matrix consisting of p variables and n observations. Suppose that each observation belongs to one of K classes. For a two class prediction problem, LDA finds the linear combination of the feature vector which maximizes the separation between the classes while the variation within the classes is kept as small as possible. This leads to finding a nonzero vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ that maximizes the generalized Rayleigh quotient for the pair $(\mathbf{M}, \mathbf{S})$

$$\boldsymbol{\beta}^* = \max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}}, \quad (1.2.2)$$

where $\mathbf{M}$ and $\mathbf{S}$ are between and within class matrices respectively, and are defined as

$$\mathbf{S} = \sum_{k=1}^K \sum_{j=1}^n (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k)^T, \quad \mathbf{M} = \sum_{k=1}^K n_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^T.$$

Here, $\hat{\boldsymbol{\mu}}_k$ is the sample mean vector for Class k , $\hat{\boldsymbol{\mu}}$ is the combined class mean vector and is defined as $\hat{\boldsymbol{\mu}} = (1/n) \sum_{k=1}^K n_k \hat{\boldsymbol{\mu}}_k$ with n_k being the number of samples in Class k . Notice that the vector $\boldsymbol{\beta}$ can be re-scaled without affecting the ratio in (1.2.2). One can thus choose $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} = 1$. Hence, Fisher's LDA for a two class problem becomes finding $\hat{\boldsymbol{\beta}}$ that solves the optimization problem

$$\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} = 1. \quad (1.2.3)$$

For $K > 2$ class problem, additional constraints that ensure that current solution vectors are uncorrelated with previous solution vectors are imposed. Generally, for a K class problem,

we have the optimization problem

$$\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^T \mathbf{M} \boldsymbol{\beta}_k \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \mathbf{S} \boldsymbol{\beta}_k = 1, \quad \boldsymbol{\beta}_l^T \mathbf{S} \boldsymbol{\beta}_k = 0 \quad \forall l < k, \quad k = 1, 2, \dots, K-1 \quad (1.2.4)$$

where $K-1$ is the rank of $\mathbf{M}$.

We will show that the optimization problem (1.2.4) results in a generalized eigenvalue problem with a solution being the eigenvalue-eigenvector pair of $\mathbf{S}^{-1}\mathbf{M}$. First, from Lagrangian multipliers, we have

$$L(\boldsymbol{\beta}, \alpha) = \boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta} - \alpha(\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} - 1).$$

Differentiating the Lagrangian with respect to $\boldsymbol{\beta}$ and setting to zero gives

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 2\mathbf{M}\boldsymbol{\beta} - 2\alpha\mathbf{S}\boldsymbol{\beta} = 0,$$

which leads to the generalized eigenvalue problem

$$\mathbf{M}\boldsymbol{\beta} = \alpha\mathbf{S}\boldsymbol{\beta}. \quad (1.2.5)$$

We next show that the solution to (1.2.5) are the eigenvalue-eigenvector pair of $\mathbf{S}^{-1}\mathbf{M}$ and that these satisfy the orthogonality constraints $\boldsymbol{\beta}_l^T \mathbf{S} \boldsymbol{\beta}_k = 0$ in (1.2.4). From (1.2.5) we have

$$\mathbf{M}\boldsymbol{\beta} = \alpha\mathbf{S}\boldsymbol{\beta} \implies \mathbf{M}\boldsymbol{\beta} = \alpha\mathbf{S}^{1/2}\mathbf{S}^{1/2}\boldsymbol{\beta} \implies \mathbf{M}\mathbf{S}^{-1/2}\mathbf{w} = \alpha\mathbf{S}^{1/2}\mathbf{w}, \quad (1.2.6)$$

where we have set $\mathbf{w} = \mathbf{S}^{1/2}\boldsymbol{\beta}$. Now pre-multiplying equation (1.2.6) by $\mathbf{S}^{-1/2}$ results in $\mathbf{S}^{-1/2}\mathbf{M}\mathbf{S}^{-1/2}\mathbf{w} = \alpha\mathbf{w}$, which is equivalent to finding the pair $(\alpha, \mathbf{w})$ that solves

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}^{-1/2} \mathbf{M} \mathbf{S}^{-1/2} \mathbf{w}}{\mathbf{w}^T \mathbf{w}}.$$

The k th maximum of this ratio is $\tilde{\alpha}_k$, the k th largest eigenvalue of $\mathbf{S}^{-1/2}\mathbf{M}\mathbf{S}^{-1/2}$ (page 80, 6th ed., Richard and Wichern (2007)) and it occurs when $\mathbf{w} = \bar{\boldsymbol{\beta}}_k$, the normalized eigenvector associated with $\tilde{\alpha}_k$. Because $\bar{\boldsymbol{\beta}}_k = \mathbf{w} = \mathbf{S}^{1/2}\boldsymbol{\beta}_k$ or $\boldsymbol{\beta}_k = \mathbf{S}^{-1/2}\bar{\boldsymbol{\beta}}_k$, we have that

the orthogonality constraints in (1.2.4) are satisfied since

$$\begin{aligned}
\boldsymbol{\beta}_l^T \mathbf{S} \boldsymbol{\beta}_k &= (\mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_l)^T \mathbf{S} \mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k \\
&= \bar{\boldsymbol{\beta}}_l^T \mathbf{S}^{-1/2} \mathbf{S} \mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k \\
&= \bar{\boldsymbol{\beta}}_l^T (\mathbf{S}^{-1/2} \mathbf{S}^{1/2}) (\mathbf{S}^{1/2} \mathbf{S}^{-1/2}) \bar{\boldsymbol{\beta}}_k \\
&= \bar{\boldsymbol{\beta}}_l^T \bar{\boldsymbol{\beta}}_k \\
&= 0 \quad \text{because } \bar{\boldsymbol{\beta}}_l \perp \bar{\boldsymbol{\beta}}_k.
\end{aligned}$$

Now, since $\tilde{\alpha}_k$ and $\bar{\boldsymbol{\beta}}_k$ are eigenvalue-eigenvector pair of $\mathbf{S}^{-1/2} \mathbf{M} \mathbf{S}^{-1/2}$, we have that

$$\mathbf{S}^{-1/2} \mathbf{M} \mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k = \tilde{\alpha}_k \bar{\boldsymbol{\beta}}_k$$

and pre-multiplying by $\mathbf{S}^{-1/2}$ gives

$$\mathbf{S}^{-1} \mathbf{M} (\mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k) = \tilde{\alpha}_k (\mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k) \implies \mathbf{S}^{-1} \mathbf{M} \boldsymbol{\beta}_k = \tilde{\alpha}_k \boldsymbol{\beta}_k.$$

Thus, $\mathbf{S}^{-1} \mathbf{M}$ have the same eigenvalues, $\tilde{\alpha}_k$, as $\mathbf{S}^{-1/2} \mathbf{M} \mathbf{S}^{-1/2}$ and eigenvectors, which we denote as $\tilde{\boldsymbol{\beta}}_k$, given by $\mathbf{S}^{-1/2} \bar{\boldsymbol{\beta}}_k$. Hence, the solution to the generalized eigenvalue problem (1.2.5) are the $K - 1$ orthogonal eigenvectors $\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_{K-1}$ that corresponds to the $K - 1$ nonzero eigenvalues, $\tilde{\alpha}_1 \geq \dots \geq \tilde{\alpha}_{K-1}$ of $\mathbf{S}^{-1} \mathbf{M}$.

The *discriminant scores* are defined to be $\mathbf{u}_l = \tilde{\boldsymbol{\beta}}_l^T \mathbf{X}$, $l = 1, \dots, K - 1$ and in particular $\tilde{\boldsymbol{\beta}}_1$ maximizes (1.2.2). Let $\tilde{\mathbf{B}} = [\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_{K-1}]$ be a concatenation of the $K - 1$ eigenvectors of $\mathbf{S}^{-1} \mathbf{M}$. A new observation $\mathbf{z} = (z_1, \dots, z_p)^T$ is assigned to the population whose mean score is closest to $\mathbf{z}^T \tilde{\mathbf{B}}$ (nearest centroid); that is assign $\mathbf{z}$ to Class k if the distance from $\mathbf{z}^T \tilde{\mathbf{B}}$ to $\hat{\boldsymbol{\mu}}_k^T \tilde{\mathbf{B}}$ is minimum, i.e.,

$$\min_k \text{dist}_k(\mathbf{z}^T \tilde{\mathbf{B}}, \hat{\boldsymbol{\mu}}_k^T \tilde{\mathbf{B}}), \quad (1.2.7)$$

where

$$\text{dist}_k(\mathbf{z}^T \tilde{\mathbf{B}}, \hat{\boldsymbol{\mu}}_k^T \tilde{\mathbf{B}}) = \sum_{l=1}^{K-1} ((\mathbf{z} - \hat{\boldsymbol{\mu}}_k)^T \tilde{\boldsymbol{\beta}}_l)^2$$

is the squared Euclidean distance in terms of the discriminant scores. When there are two classes to classify, the LDA solution vector $\tilde{\boldsymbol{\beta}}$ reduces to

$$\tilde{\boldsymbol{\beta}} \propto \mathbf{S}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2). \quad (1.2.8)$$

Here, the population version of the solution vector is obtained by replacing the sample covariance matrix $\mathbf{S}$ and sample mean $\hat{\boldsymbol{\mu}}_i$ with their corresponding population parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_i$ respectively. Fisher's LDA does not impose any distributional assumption on the data variables and hence may be used for general populations. For a two class problem, the classifier (1.2.8) is exactly the same as the maximum likelihood discriminant rule for multivariate normal class distribution with the same covariance matrix (Mardia et al., 2003).

In HDLSS, it is a known fact that the original LDA suffers from singularity of the sample covariance matrix (Bickel and Levina, 2004; Ahn and Marron, 2010). Various regularized versions of LDA have been proposed over the recent years, most of which are intended for high dimensional applications. A popular modification of the original LDA is to fix the singularity of the sample covariance matrix by applying a ridge type regularization (Friedman, 1989).

Some modifications of the original LDA results in discriminant vectors with many entries that are zero. These sparse vectors have been shown to produce better classification accuracy in high dimensional classification studies (Dudoit et al., 2002). A lot of work has been done in the literature in the area of sparse LDA in recent years (see for instance Qiao et al. (2009); Clemmensen et al. (2011); Zou and Hastie (2005); Witten and Tibshirani (2011), and the references there in). Most of these methods enforce sparsity by incorporating lasso (l_1) or elastic net (l_1 and l_2) penalties in the LDA optimization problem (1.2.3). The lasso-based sparse methods can only select up to n variables, which may not be enough when the true structure of the solution vector is not too sparse.

Other sparse LDA methods are motivated by the solution (1.2.8) directly, rather than modifying the LDA optimization problem (1.2.3). Cai and Liu (2011) noted that $\boldsymbol{\beta}$ solves the equation $\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and proposed to directly estimate $\boldsymbol{\beta}$, by using a similar approach

to the Dantzig selector (Candes and Tao, 2007). Shao et al. (2011) assumed that both the common covariance matrix Σ and the mean difference vector $\mu_1 - \mu_2$ are sparse and suggested hard thresholding entries of the sample mean difference and off-diagonals of the sample covariance matrix; their approach does not necessarily yield a sparse estimate for β .

In the analysis part of this dissertation in Chapter 3, we present a general method with which one can generalize a binary LDA method to multi-class. We take two popular sparse binary LDA methods developed by Cai and Liu (2011) and Shao et al. (2011) to demonstrate our approach. The proposed approach utilizes the fact that the canonical subspace generated by original LDA solution is equivalent to the subspace that would be generated by LDA using basis vectors of between-class scatter, $\mathbf{M}$, instead of the mean difference vector. It produces a low-dimensional canonical subspace generated by sparse discriminant vectors. We will discuss briefly some of the popular approaches to sparse LDA. What follows is a discussion of a HDLSS regularization method, the Dantzig selector, which will serve as the building block for the sparse methods we propose in this dissertation in Chapters 3 and 4.

1.2.3.1 THE DANTZIG SELECTOR: STATISTICAL ESTIMATION WHEN $p \gg n$

Candes and Tao (2007) proposed the Dantzig selector (DS) for sparse signal recovery and model selection in HDLSS. They considered sparse estimation of regression coefficients in multiple regression analysis. For a noiseless regression problem

$$\mathbf{y} = \mathbf{X}\beta, \tag{1.2.9}$$

where $\mathbf{X}$ is a $n \times p$ matrix and $\mathbf{y}$ is a vector in $\mathbb{R}^n$, because $p > n$, the regression equation (1.2.9) is underdetermined and has many solutions, which makes estimating β reliably from $\mathbf{y}$ seemingly impossible. Now, if one assumes that β has some structure in the sense that it is sparse, so that only some of its entries are nonzero making the search for solutions feasible, the goal becomes finding the most sparsest solution among all sparse representations of β .

Candes and Tao (2005) showed that the sparse solution can be recovered by solving the convex program

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta},$$

where $\|\mathbf{x}\|_1$ is the l_1 norm of the vector $\mathbf{x}$, and is defined as $\sum_{i=1}^p |x_i|$.

In the situation where the data are noisy so that Gaussian noise term ϵ is added to (1.2.9), the normal equations

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$$

yielding the classical least squares estimator becomes nonstandard since $p > n$ and needs to be estimated reliably. Candes and Tao (2007) proposed to estimate $\boldsymbol{\beta}$ by l_1 minimization of $\boldsymbol{\beta}$ subject to l_∞ bound on the residuals $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}\|_\infty \leq \lambda_p \sigma, \quad (1.2.10)$$

where $\|\mathbf{x}\|_\infty$ is the l_∞ norm of the vector $\mathbf{x}$ and is defined as $\max_i |x_i|, i = 1, \dots, p$, and where $\lambda_p > 0$ and σ is the error standard deviation. The optimization problem is convex and can easily be solved using linear programming, making the estimation procedure computationally tractable when compared to other regularization techniques. They theoretically justified the oracle property of variable selection consistency by showing that the estimator (1.2.10) selects the best subset of variables. The estimator (1.2.10) was called the *Dantzig selector* in memory of the father of linear programming, George B. Dantzig, and also to emphasize that the convex program (DS) is effectively a variable selection technique.

1.2.3.2 SPARSE LINEAR DISCRIMINANT ANALYSIS

Clemmensen et al. (2011) proposed a method of extending LDA to HDLSS that yields sparse discriminant vectors. Their approach was based on optimal scoring technique which recasts LDA problem (1.2.4) as a regression problem by turning categorical variables into quantitative variables via a sequence of scorings. Let $\mathbf{Y}$ be a $K \times n$ indicator matrix of class

labels having the entry one if the observation belongs to class k and zero otherwise. Then Fisher's problem (1.2.4) in regression form is

$$\begin{aligned} & \min_{\boldsymbol{\beta}_k, \boldsymbol{\theta}_k} \|\mathbf{Y}^T \boldsymbol{\theta}_k - \mathbf{X}^T \boldsymbol{\beta}_k\|^2 \quad \text{subject to} \\ & \frac{1}{n} \boldsymbol{\theta}_k^T \mathbf{Y} \mathbf{Y}^T \boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^T \mathbf{Y} \mathbf{Y}^T \boldsymbol{\theta}_l = 0, \quad \text{for all } l < k, \end{aligned} \quad (1.2.11)$$

where $\boldsymbol{\theta}_k$ is a K -vector of scores and $\boldsymbol{\beta}_k$ is a p -vector of variable coefficients. The nonsparse solution vector $\hat{\boldsymbol{\beta}}_k$ was made sparse by adding l_1 and l_2 penalties to the objective function in (1.2.11):

$$\begin{aligned} & \min_{\boldsymbol{\beta}_k, \boldsymbol{\theta}_k} \{ \|\mathbf{Y}^T \boldsymbol{\theta}_k - \mathbf{X}^T \boldsymbol{\beta}_k\|^2 + \eta \boldsymbol{\beta}_k^T \boldsymbol{\Omega} \boldsymbol{\beta}_k + \lambda \|\boldsymbol{\beta}_k\|_1 \} \\ & \text{subject to} \quad \frac{1}{n} \boldsymbol{\theta}_k^T \mathbf{Y} \mathbf{Y}^T \boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^T \mathbf{Y} \mathbf{Y}^T \boldsymbol{\theta}_l = 0, \quad \text{for all } l < k, \end{aligned} \quad (1.2.12)$$

where $\lambda > 0$ and $\eta > 0$ are l_1 and l_2 penalty parameters respectively, and $\boldsymbol{\Omega}$ is a positive definite matrix. The l_2 penalty parameter shrinks coefficients in $\hat{\boldsymbol{\beta}}_k$ towards zero to avoid overfitting of the high dimensional data. However, it does not produce sparse $\hat{\boldsymbol{\beta}}_k$ hence the addition of the l_1 penalty function. The $\hat{\boldsymbol{\beta}}_k$ that solves equation (1.2.12) was referred to as the k th SLDA discriminant vector. Once the sparse discriminant direction vectors have been obtained, a new observation $\mathbf{z}$ may be assigned to the closest population in the space of discriminant scores using (1.2.7).

1.2.3.3 PENALIZED LINEAR DISCRIMINANT ANALYSIS

Witten and Tibshirani (2011) also approached the LDA problem via Fisher's framework. For $\mathbf{Y}$, a $K \times n$ indicator matrix of class labels, it can be shown that the solution $\hat{\boldsymbol{\beta}}_k$ to Fisher's problem (1.2.4) also solves

$$\max_{\boldsymbol{\beta}_k} \{ \boldsymbol{\beta}_k^T \mathbf{M}^k \boldsymbol{\beta}_k \} \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \tilde{\mathbf{S}} \boldsymbol{\beta}_k \leq 1, \quad (1.2.13)$$

where $\tilde{\mathbf{S}}$ is a positive definite covariance matrix for $\mathbf{S}$, $\mathbf{M}^k = \frac{1}{n} \mathbf{X} \mathbf{Y}^T (\mathbf{Y} \mathbf{Y}^T)^{-\frac{1}{2}} \mathbf{P}_k^\perp (\mathbf{Y} \mathbf{Y}^T)^{-\frac{1}{2}} \mathbf{Y} \mathbf{X}^T$ with $\mathbf{P}_k^\perp$ defined as an orthogonal projection matrix into the space orthogonal to $(\mathbf{Y} \mathbf{Y}^T)^{-\frac{1}{2}} \mathbf{Y} \mathbf{X}^T \hat{\boldsymbol{\beta}}_l$, for all $l < k$. The nonsparse solution vector $\hat{\boldsymbol{\beta}}_k$ was made sparse by applying l_1 or fused lasso penalty to (1.2.13):

$$\max_{\boldsymbol{\beta}_k} \{ \boldsymbol{\beta}_k^T \mathbf{M}^k \boldsymbol{\beta}_k - \lambda_k \sum_{j=1}^p |\hat{\sigma}_j \beta_{kj}| \} \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \tilde{\mathbf{S}} \boldsymbol{\beta}_k \leq 1, \quad (1.2.14)$$

where $\hat{\sigma}_j$ is the pooled within-class standard deviation for feature j . The k th solution vector $\hat{\boldsymbol{\beta}}_k$ was called the k th penalized LDA- L_1 emphasizing the use of l_1 penalty function. The resulting optimization problem is nonconvex so they suggested using minorization-maximization technique that allows one to solve problem (1.2.14) efficiently with convex penalties. Once the penalized discriminant direction vectors have been obtained, a new observation $\mathbf{z}$ may be assigned to the closest population in the space of discriminant scores using (1.2.7).

1.2.3.4 SPARSE LINEAR DISCRIMINANT ANALYSIS VIA THRESHOLDING

Shao et al. (2011) obtained sparse LDA vectors directly from LDA solution in (1.2.8) rather than modifying the LDA optimization problem in (1.2.3). They assumed that the common population covariance matrix $\boldsymbol{\Sigma}$ is sparse as well as the population mean difference vector $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, which we denote as $\boldsymbol{\delta}$, and suggested hard thresholding the sample versions $\mathbf{S}$ and $\hat{\boldsymbol{\delta}}$ separately. Specifically, using ideas borrowed from Bickel and Levina (2004), they replaced the off-diagonal elements of $\mathbf{S}$ with

$$\hat{\sigma}_{jl} I(|\hat{\sigma}_{jl}| > t_n), \quad t_n = M_1 \sqrt{\log p / \sqrt{n}},$$

where $\hat{\sigma}_{jl}$ is the (j, l) th component of $\mathbf{S}$, $I(A)$ is the indicator variable, which is 1 if A holds and 0 otherwise, and $M_1 > 0$ is a tuning parameter. Sparsity on $\boldsymbol{\delta}$ was achieved by

$$\hat{\delta}_j I(|\hat{\delta}_j| > a_n), \quad a_n = M_2 \left(\frac{\log p}{n} \right)^\alpha, \quad j \neq l,$$

where $\hat{\delta}_j$ is the j th component of $\hat{\boldsymbol{\delta}}$, $M_2 > 0$ and $\alpha \in (0, 1/2)$ are tuning parameters. Let $\tilde{\mathbf{S}}$ and $\tilde{\boldsymbol{\delta}}$ be the thresholded versions of $\mathbf{S}$ and $\hat{\boldsymbol{\delta}}$ respectively. Then, $\hat{\boldsymbol{\beta}}$ is estimated as $\tilde{\mathbf{S}}^{-1}\tilde{\boldsymbol{\delta}}$ and is used in the classification rule in (1.2.7). We note that by making $\mathbf{S}$ and $\hat{\boldsymbol{\delta}}$ sparse, the LDA vector $\hat{\boldsymbol{\beta}}$, which is the product of $\tilde{\mathbf{S}}^{-1}$ and $\tilde{\boldsymbol{\delta}}$ is not necessarily sparse.

For multi-class classification problem with $K > 2$, Shao et al. (2011) considered all pairwise combinations of classes and proposed to classify $\mathbf{z}$ to the dominant class. This pairwise approach potentially has some drawbacks. First, this method suffers from computational burden if the number of classes K is large. Second, there are likely instances where a class cannot be assigned for some observations when there is no dominant class in the comparisons. Third, this method cannot produce a $K - 1$ dimensional canonical subspace, which is useful for graphical display of the classes.

1.2.3.5 BINARY CLASS DISCRIMINATION VIA LINEAR PROGRAMMING

Cai and Liu (2011) observed that the optimal LDA solution $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ solves the equation $\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\delta}$ and proposed to directly estimate $\boldsymbol{\beta}$ by using a similar approach to the Dantzig selector (Candes and Tao, 2007) in (1.2.10). In HDLSS, the singularity of $\mathbf{S}$ causes the solution to be degenerate. As a refinement, a ridge-type modification where a small multiple, $\rho > 0$, of the identity matrix is added to the covariance matrix is usually implemented. Let $\mathbf{S}_\rho = \mathbf{S} + \rho\mathbf{I}$ be the ridge-corrected sample covariance. A value of ρ suggested by Cai and Liu (2011) is $\rho = \sqrt{\log(p)/n}$. The LDA solution vector $\boldsymbol{\beta}$ was made sparse by solving the optimization problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|\mathbf{S}_\rho\boldsymbol{\beta} - \hat{\boldsymbol{\delta}}\|_\infty \leq \lambda, \quad (1.2.15)$$

where $\lambda \in \mathbb{R}^+$ is a nonzero tuning parameter that controls how many coefficients in $\boldsymbol{\beta}$ are set to zero. Since the objective function and constraints in (3.2.3) are linear in $\boldsymbol{\beta}$, the solution $\hat{\boldsymbol{\beta}}$ can be found by linear programming. The generalization to multi-class was carried out in

a similar way as in Shao et al. (2011). A new data $\mathbf{z}$ was allocated to the dominating class from the pairwise comparisons, making it have the limitations discussed in section (1.2.3.4).

1.2.4 SPARSE HIGH DIMENSION, LOW SAMPLE SIZE METHODS

Most traditional multivariate methods revolve around the common theme of projecting high dimensional data onto basis vectors, which are a few meaningful directions spanning a lower dimensional subspace. For instance, for a K class linear discriminant analysis problem, these meaningful directions are $K - 1$ vectors with maximal discriminatory power and are less than the number of available variables. It so happens that many of these direction vectors are solutions of a generalized eigenvalue (GEV) problem.

The GEV problem for the pair of matrices $(\mathbf{M}, \mathbf{S})$ is the problem of finding a pair $(\alpha, \mathbf{v})$ such that

$$\mathbf{M}\mathbf{v} = \alpha\mathbf{S}\mathbf{v} \tag{1.2.16}$$

where $\mathbf{M}, \mathbf{S} \in \mathbb{R}^{p \times p}$ are usually symmetric matrices, $\mathbf{v} \neq \mathbf{0}$ and $\alpha \in \mathbb{R}$. The pair $(\alpha, \mathbf{v})$ that solves the GEV problem is called the generalized eigenvalue-eigenvector pair. Some popular and widely used data analysis methods that results from the GEV problem are principal component analysis (PCA), Fisher's linear discriminant analysis (LDA), canonical correlation analysis (CCA), and multivariate analysis of variance (MANOVA). In PCA, the principal components are directions of maximal variance in a given multivariate data set. Also, the discriminant vectors in Fisher's LDA separates the sample classes as much as possible while ensuring that the variation within each class is smallest as possible. In CCA where association between two sets of variables is of interest, the canonical variates are directions of maximal correlation. Despite the popularity of these methods, one main drawback is the lack of sparsity. They have the limitation that their solution vector $\mathbf{v}$ is a weighted combination of all available variables, making it difficult to interpret the results often times.

Sparse representations usually have physical interpretations in practice, and since they have been shown to have good prediction performance in many high dimensional studies (Dudoit et al., 2002), it is a good enough reason to study and use them in practice. For example, in credit card analysis by financial service companies, the two sets of variables in CCA may be family characteristics (family size, income, expenditure, other debts etc.) and credit card usage (amount spent each month, minimum payment per month, account balance, number of credit cards etc). Sparsity in family characteristics will allow to select the most important components affecting the family's overall credit card usage.

Several approaches have been discussed in the literature to make $\mathbf{v}$ sparse. An ad hoc way to achieve sparsity on $\mathbf{v}$ is to set the loadings with absolute values smaller than a threshold to zero. This approach, though simple and conceptually intuitive, can be potentially misleading in various respects (Cadima and Jolliffe, 1995). A popular approach for sparse $\mathbf{v}$ is to solve a variation (Sriperumbudur et al., 2011) of a GEV problem

$$\max_{\mathbf{v}} \{ \mathbf{v}^T \mathbf{M} \mathbf{v} : \mathbf{v}^T \mathbf{S} \mathbf{v} = 1 \} \quad (1.2.17)$$

for the specific multivariate problem while imposing lasso penalty (Tibshirani, 1994), adaptive lasso (Zou, 2006), fused lasso (Tibshirani et al., 2005), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), to mention but a few. For example, Jolliffe et al. (2003) proposed ScoTLASS, a procedure that obtains sparse vectors by directly imposing l_1 constraint on PCA. Witten and Tibshirani (2011) applied lasso and fused lasso on Fisher's LDA problem to obtain sparse vectors that results in maximal separation between classes and minimal variation within classes. These lasso-based approaches have a major limitation of selecting the size of the sample as the maximum number of variables in $\mathbf{v}$, which is too restrictive for some high dimensional studies. Some other methods in the literature achieve sparsity differently. Sriperumbudur et al. (2011) obtained sparse solutions by constraining the cardinality of $\mathbf{v}$ with negative log-likelihood of a Student's t-distribution instead of the popular lasso penalty and its variants.

In Chapter 4, we propose a different way of obtaining sparse $\mathbf{v}$ from the GEV problem (1.2.16) directly rather than the variation of (1.2.17) and demonstrate its' use on LDA and CCA. We describe briefly the concept of CCA and review the literature on some popular sparse CCA.

1.2.4.1 SPARSE CANONICAL CORRELATION ANALYSIS

Canonical Correlation Analysis (Hotelling, 1936) is a statistical method used to study inter-relations between two sets of variables. CCA finds a weighted combination of all available variables within one set of variables and a weighted combination of all the other variables in the other set such that they have maximum correlation. Suppose that we have two $n \times p$ and $n \times q$ data matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively. The traditional CCA finds $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that solves the optimization problem

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} = 1 \quad \text{and} \quad \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} = 1, \quad (1.2.18)$$

where $\boldsymbol{\Sigma}_{xy}$, $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are cross-covariance and covariance matrices respectively. The optimization problem (1.2.18) may be solved by applying singular value decomposition (SVD) on the matrix

$$K = \boldsymbol{\Sigma}_{xx}^{-1/2} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1/2}.$$

Parkhomenko et al. (2009) achieved sparsity of the canonical vectors by considering a sparse SVD of the matrix K . In particular, a soft-thresholding technique similar to lasso (Tibshirani, 1994) was applied iteratively to the left and right singular vectors of the SVD of K . The update and univariate soft-thresholding function for $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} \boldsymbol{\alpha}_c &= K \boldsymbol{\beta}_d && \text{update of } \boldsymbol{\alpha} \\ \left(|\boldsymbol{\alpha}_{c_j}| - \frac{1}{2} \lambda_\alpha \right)_+ \text{Sign}(\boldsymbol{\alpha}_{c_j}), \quad j = 1, 2, \dots, p &&& \text{soft-threshold for sparse } \boldsymbol{\alpha} \\ \hat{\boldsymbol{\alpha}} &= \frac{\boldsymbol{\alpha}_c}{\|\boldsymbol{\alpha}_c\|_2} && \text{normalize sparse solution,} \end{aligned}$$

where subscripts c and d denote current and previous iterations respectively, $f_+ = f$ if $f > 0$ and $f_+ = 0$ if $f \leq 0$, and λ_α is a tuning parameter that controls the cardinality of zeros in α_c . The update and univariate soft-thresholding function for β is similar. In their algorithm, Σ_{xx} and Σ_{yy} are replaced with their diagonal estimates to make the matrices nonsingular.

1.2.4.2 SPARSE CANONICAL CORRELATION ANALYSIS VIA ELASTIC NET

Waaijenborg et al. (2008) enforced sparsity on the canonical vectors by recasting the CCA problem into regression and applying elastic net penalty (Zou and Hastie, 2005), a combination of l_1 and l_2 penalties to the regression problem. Let $\omega = \mathbf{X}\alpha$ and $\xi = \mathbf{Y}\beta$ be the weighted combinations of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Because the objective of CCA is to maximize correlation between ω and ξ , the initial canonical covariates ξ and ω are respectively regressed on $\mathbf{X}$ and $\mathbf{Y}$ to estimate a new set of weights. With this new set of weights, new canonical covariates are obtained, which are in turn regressed on $\mathbf{X}$ and $\mathbf{Y}$. The process is repeated until the weights converge, yielding the first pair of canonical variates with maximum correlation. Mathematically, current sparse estimates $\hat{\alpha}_c$ and $\hat{\beta}_c$ are obtained by performing multiple regression using elastic net:

$$\begin{aligned} \min_{\alpha_c} & \alpha_c^T \left(\frac{\mathbf{X}^T \mathbf{X} + \lambda_{2X} \mathbf{I}}{1 + \lambda_{2X}} \right) \alpha_c - 2 \hat{\xi}_c^T \mathbf{X} \alpha_c + \lambda_{1X} \|\alpha_c\|_1 \\ \min_{\beta_c} & \beta_c^T \left(\frac{\mathbf{Y}^T \mathbf{Y} + \lambda_{2Y} \mathbf{I}}{1 + \lambda_{2Y}} \right) \beta_c - 2 \hat{\omega}_c^T \mathbf{Y} \beta_c + \lambda_{1Y} \|\beta_c\|_1, \end{aligned} \quad (1.2.19)$$

where $\hat{\omega}_c = \mathbf{X}\hat{\alpha}_d$ and $\hat{\xi}_c = \mathbf{Y}\hat{\beta}_d$ with subscripts c and d denoting current and previous iterations respectively, λ_1 and λ_2 are lasso and ridge penalties respectively. The minimization problem in (1.2.19) may be computationally expensive since it has four tuning parameters that need to be optimized in order to estimate $\hat{\alpha}$ and $\hat{\beta}$. To make it computationally tractable, Waaijenborg et al. (2008) proposed to set $\lambda_{2X} \rightarrow \infty$ and $\lambda_{2Y} \rightarrow \infty$, which results in univariate

soft-thresholding:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_c &= \left(|\boldsymbol{\xi}_c^T \mathbf{X}_j| - \frac{1}{2} \lambda_{1X} \right)_+ \text{Sign}(\boldsymbol{\xi}_c^T \mathbf{X}_j), \quad j = 1, \dots, p \\ \hat{\boldsymbol{\beta}}_c &= \left(|\boldsymbol{\omega}_c^T \mathbf{Y}_j| - \frac{1}{2} \lambda_{1Y} \right)_+ \text{Sign}(\boldsymbol{\omega}_c^T \mathbf{Y}_j), \quad j = 1, \dots, q.\end{aligned}$$

It is observable from (1.2.20) that only the optimal l_1 penalties have to be chosen now.

1.2.4.3 SPARSE CANONICAL CORRELATION ANALYSIS VIA PENALIZED MATRIX DECOMPOSITION

Witten et al. (2009) proposed penalized matrix decomposition (PMD), a method that decomposes a random matrix using SVD and then applies penalty functions to the singular vectors to achieve sparsity. Let the SVD of the random matrix $\boldsymbol{\Sigma}_{xy}$ be $\mathbf{A}\mathbf{\Lambda}\mathbf{B}^T$, and let $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$ be the k th left and right singular vectors respectively and let λ_k be the k th singular value. It is known that for $r \leq K = \min(p, q)$, the rank r approximation of the random matrix $\hat{\boldsymbol{\Sigma}}_{xy} = \sum_{k=1}^r \lambda_k \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^T$ minimizes $\|\boldsymbol{\Sigma}_{xy} - \hat{\boldsymbol{\Sigma}}_{xy}\|_F^2$, where $\|\cdot\|_F$ is the squared Frobenius norm, which for a matrix $\mathbf{F}$ is defined as trace of $(\mathbf{F}^T \mathbf{F})$. The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that minimize the Frobenius norm of the difference matrix $\boldsymbol{\Sigma}_{xy} - \hat{\boldsymbol{\Sigma}}_{xy}$ subject to penalties on $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ also solve the canonical correlation optimization problem

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\alpha}^T \boldsymbol{\alpha} \leq 1, \boldsymbol{\beta}^T \boldsymbol{\beta} \leq 1, P_1(\boldsymbol{\alpha}) \leq c_1, P_2(\boldsymbol{\beta}) \leq c_2. \quad (1.2.20)$$

Here, $K = 1$, P_1 and P_2 are penalty functions which were chosen to be lasso or fused lasso. The optimization problem (1.2.20) was termed the (rank-one) PMD because the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that solve the optimization problem are the first regularized singular vectors of $\boldsymbol{\Sigma}_{xy}$. The optimization problem is biconvex and may be solved iteratively by fixing $\boldsymbol{\beta}$ and solving for $\boldsymbol{\alpha}$ and conversely fixing $\boldsymbol{\alpha}$ and solving for $\boldsymbol{\beta}$. With $\boldsymbol{\beta}$ fixed and assuming that both P_1 and P_2 are lasso penalty functions, the solution $\hat{\boldsymbol{\alpha}}$ to (1.2.20) is obtained by univariate soft-thresholding

$\Sigma_{xy}\beta$ and normalizing:

$$\begin{aligned}\tilde{\alpha} &= (|(\Sigma_{xy}\beta)_j| - \Delta_1)_+ \text{Sign}((\Sigma_{xy}\beta)_j), \quad j = 1, \dots, p \\ \hat{\alpha} &= \frac{\tilde{\alpha}}{\|\tilde{\alpha}\|_2}\end{aligned}$$

where $\Delta_1 = 0$ if $\|\tilde{\alpha}\|_1 \leq \lambda_\alpha$; otherwise Δ_1 is chosen to be a positive constant such that $\|\tilde{\alpha}\|_1 = \lambda_\alpha$. Also, with α fixed, the solution $\hat{\beta}$ is obtained in a similar way. The canonical covariates arising from this was termed $\text{PMD}(L_1, L_1)$, emphasizing the use of lasso penalties to obtain both sparse $\hat{\alpha}$ and $\hat{\beta}$.

Witten et al. (2009) also proposed $\text{PMD}(L_1, FL)$ for problems where features in one set of variables are assumed ordered in some meaningful way by implementing fused lasso penalty on the variables in that set. The fused lasso penalizes the l_1 norm of both coefficients and their successive differences. Assuming that the features in $\mathbf{Y}$ are ordered in some meaningful way, β was made sparse by solving

$$\hat{\beta} = \min_{\beta} \left\{ \frac{1}{2} \|\Sigma_{xy}^T \alpha - \beta\|^2 + \lambda_\alpha \sum_{j=1}^q |\beta_j| + \lambda_\beta \sum_{j=1}^q |\beta_j - \beta_{j-1}| \right\}, \quad j = 1, 2, \dots, q.$$

The canonical covariates arising from this was termed $\text{PMD}(L_1, FL)$, which also emphasize the use of lasso penalty on one set of variables and fused lasso penalty on the other set of variables.

1.3 OUTLINE OF DISSERTATION

In Chapter 2, we consider sample size requirements for developing classifiers in high dimensional studies. We propose a nonparametric sample size method that requires a pilot dataset. The novelty in our method is the use of regularized logistic regression to reduce dimensionality, errors-in-variables methods to estimate asymptotic performance as $n \rightarrow \infty$, and the inclusion of clinical covariates to effectively characterize individual subjects. The sample size method will help clinicians who are designing new studies, those wanting to evaluate

the adequacy of existing studies, or those who want to incorporate clinical covariates to already existing studies. We also make available MATLAB program for implementation of our method.

In Chapter 3, a sparse discriminant function for classifying new samples into more than two classes is considered. A novel method that generalizes binary LDA methods to multi-class is proposed. The methodology exploits the structural relationship between basis vectors of the between class scatter and Fisher's LDA. We apply our method to two existing binary LDA methods. The work is motivated by the uprise of multi-class classification problems and the popularity of LDA as a classification tool. Simulation studies have been carried out to compare the classification accuracy and selectivity of our method to other existing methods. Various types of real data examples including RNA-seq data have been used to show the efficacy of our method.

In Chapter 4, a general framework for producing sparse high dimensional vectors for many multivariate statistical problems is discussed. The methodology capitalizes on a core idea of extracting meaningful direction vectors spanning a lower dimensional subspace and their relationships with generalized eigenvalue problems. A demonstration of the use of the method is carried out on two multivariate statistical problems - linear discriminant analysis and canonical correlation analysis, to obtain sparse solution vectors. Simulation processes and real data applications reveal superior performance of the proposed method in comparison to existing methods.

1.4 REFERENCES

- Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.
- Bickel, P. J. and Levina, E. (2004). Some theory for fisher’s linear discriminant function, naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.
- Cadima, J. and Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–212.
- Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.
- Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2nd edition.
- Clemmensen, L., Hastie, T., Witten, D., and Ersbll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.
- Cook, J. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

- Dobbin, K. K. and Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional dna microarray data. *Biostatistics*, 8(1):101–117.
- Donoho, D. L. (2000). Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality.
- Dudoit, S., Fridlyand, J., and Terence, P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.
- Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011). The developmental transcriptome of drosophila melanogaster. *Nature*, 471:473–479.
- Hanfelt, J. J. and Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82(3):pp. 461–477.

- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
- Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, pages 312–377.
- Huang, Y. and Wang, C. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, 96:1469–1482.
- Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.
- Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics*, 24(4):823–834.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (2003). *Multivariate Analysis*. Academic Press.
- Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., , and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying dna microarray data. *Journal of Computational Biology*, 10(2):119–142.
- Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1):127–137.
- Novick, S. J. and Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481.
- Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8.

- Qiao, Z., Zhou, L., and Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data. *IAENG International Journal of Applied Mathematics*, 39(1):48–60.
- Richard, J. A. and Wichern, W. D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 6th edition.
- Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*., 39:1241–1265.
- Sriperumbudur, B., Torres, D. A., and Lanckriet, R. (2011). A majorization-minimization approach to sparse generalized eigenvalue problem. *Journal of Machine Learning*, 85:3–39.
- Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74:703–716.
- Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.
- Waaijenborg, S., de Witt Hamer, P. C. V., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7.
- Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

- Witten, D. M., Tibshirani, R. J., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

CHAPTER 2

SAMPLE SIZE DETERMINATION FOR REGULARIZED LOGISTIC REGRESSION-BASED CLASSIFICATION¹

¹Sandra Safo, Xiao Song, and Kevin K. Dobbin (2014) Sample Size Determination for Training Cancer Classifiers from Microarray and RNAseq Data. *Submitted to The Annals of Applied Statistics*.

ABSTRACT

A method for estimating the number of samples needed for regularized logistic regression when the objective is to develop a classifier is presented. Regularized methods, such as the lasso, are widely used for high dimensional classification. A sample size n is adequate if the developed predictor’s expected performance is close to the optimal performance as $n \rightarrow \infty$. Optimal performance can be in terms of logistic regression slope or misclassification error. The new method requires a pilot dataset, or simulated dataset if no pilot exists. Errors-in-variables (EIV) regression techniques are used to estimate the asymptotic model and resampling methods to estimate EIV regression inputs. Comparisons with an existing method are examined using simulated data, resampled microarray data, and RNA-seq data. Software to implement the method is made available.

KEYWORDS: Sample size, Lasso, Classification, Regularized logistic regression, Conditional score, High dimensional data, Measurement error

2.1 INTRODUCTION

Regularized regression methods, such as the lasso, are common in the analysis of high dimensional data (Bi et al., 2014; Zwiener et al., 2014; Moehler et al., 2013; Zhang et al., 2013). Regularized logistic regression is often used to classify patients into different groups, such as those who will versus will not respond to a targeted therapy. While development of classifiers is a long process (Pfeffer et al., 2009; Hanash et al., 2011; Simon, 2010), a critical step in that process is determining the sample size necessary to adequately train a classifier from high dimensional data.

There are several methods for sample size determination to identify features associated with a class distinction while controlling the false discovery rate (FDR) and related quantities

(Pawitan et al., 2005; Li et al., 2005; Shao and Tseng, 2007; Liu and Hwang, 2007; Jung et al., 2005; Pounds and Cheng, 2005; Tibshirani, 2006). Identification of features related to the class distinction is part of regularized regression, but there is not a direct relationship between FDR control, statistical power, and classifier performance. A sample size adequate for identifying features may or may not be adequate for classifier development. Similarly, there is a rich theory on misclassification error bounds and sample size in machine learning (Vapnik and Chervonenkis, 1971), but worst-case methods developed from VC theory, such as the probably approximately correct learning framework (e.g., Bishop (2006)), are likely to result in sample size estimates that are excessively conservative.

There are very few sample size methods focused specifically on the objective of classifier training. One of the first was developed by Lachenbruch (1968). His sample size method guarantees a mean error rate within a given tolerance of the optimal (also known as Bayes) error rate. This approach assumes a discriminant analysis and homoscedastic normal model. In high dimensions, classification algorithms are typically more complex than discriminant analysis. Mukherjee et al. (2003) developed a generic sample size method that can be used with more complicated classifiers. Their method is based on resampling repeatedly from a pilot dataset, and developing the classifier each time. Error rates are estimated for different training sample sizes, and then the asymptotic performance is extrapolated using parametric regression. There is some precedent for this approach (Duda et al., 2000). Dobbin and Simon (2007) and Dobbin et al. (2008) used a parametric approach to sample size estimation based on a high dimensional multivariate normal model. Advantages of this method are that it is simple and can be used when no pilot dataset is available. Drawbacks of the method are the assumption of a parametric model and the need to prespecify a stringency parameter used for feature selection. de Valpine et al. (2009) developed a model-based approach to this problem using simulation from a multivariate normal and approximating equations. Unlike these previous works, the approach presented in this dissertation does not assume

a multivariate normal model for the high dimensional data and does not use a parametric learning curve extrapolation to estimate the asymptotic model performance. Asymptotic performance is estimated directly using errors-in-variables regression methods. This general approach is similar to one presented in Dobbin and Song (2013) for Cox regression.

Errors-in-variables (EIV) methods for logistic regression include simulation extrapolation (SIMEX, Cook and Stefanski (1994)), conditional score (Stefanski and Carroll (1987)), consistent functional methods (Huang and Wang, 2001), approximate corrected score (Novick and Stefanski, 2002), projected likelihood ratio (Hanfelt and Liang, 1995), and quasi-likelihood (Hanfelt and Liang, 1995). We discuss each approach briefly. The SIMEX EIV method adds additional measurement error to the data, establishes the trend, and extrapolates back to the no error model using a fitted polynomial regression; as discussed in Cook and Stefanski (1994), evaluating the adequacy of a fitted regression requires judgment and is not automatic. The subjective fitting step can complicate algorithm implementation and Monte Carlo evaluation of performance. So we do not focus on SIMEX, although we do find SIMEX useful in settings where other EIV methods do not perform well. The consistent functional method is most valuable in large-scale studies (Huang and Wang, 2001), which are currently very rare in high dimensions with sequencing or microarrays. The logistic model does not fit the corrected score smoothness assumptions; also, the Monte Carlo corrected score method is not consistent for logistic regression and implementation of the method requires programming software with complex number capabilities (Carroll et al., 2006). Conditional score methods, on the other hand, are computationally tractable and relatively easy to implement, and have shown good finite-sample performance (Carroll et al., 2006). We found the quasi-likelihood method more stable than the conditional score, so we used this closely related approach.

A practical question when using a sample size method that will be based on a pilot dataset, rather than a parametric model, is whether the pilot dataset is large enough. If the

pilot dataset is too small, then no classifier developed on it may be statistically significantly better than chance, which can be assessed with a permutation test (e.g., Mukherjee et al. (2003)). But, even if the classifier developed on the pilot dataset is better than chance, the pilot dataset can still be too small to estimate the asymptotic performance as $n \rightarrow \infty$ well. This latter is a more complex question. But because it is practically important, guidelines are developed here for evaluating the pilot dataset size.

Our approach is based on resampling from a pilot dataset, or from a simulated dataset if no pilot is available. Resampling is used to estimate the logistic regression slopes for different sample sizes, and the prediction error variances. Cross-validation (CV) (e.g., Geisser, 1993) is a well-established method for obtaining nearly unbiased estimates of logistic regression slopes. Because regularized regression already contains a cross-validation step for parameter tuning, estimating the logistic regression slope by cross-validation requires nested (double) cross-validation (e.g., Davison and Hinkley (1997)). An inner cross-validation loop selects the penalty parameter value, which is then used in the outer loop to obtain the cross-validated classification scores. We also found it necessary to center and rescale individual CV batches, and repeat the CV 20-50 times to denoise the estimates. This process is termed repeated, centered, scaled cross-validation (RCS-CV). To estimate prediction error variances, the leave-one-out bootstrap (LOOBS) (Efron and Tibshirani, 1997) can be used. Modification of standard LOOBS is needed because of the cross-validation step embedded in the regularized regression. To avoid information leak, the prediction error variance is estimated by the leave-one-out nested case-cross-validated (LOO-NCCV-BS) bootstrap (Varma and Simon, 2006). The same centering and scaling steps added for CV were also added to the LOO-NCCV-BS. We call this CS-LOO-NCCV-BS.

Regularized regression for high dimensional data is a very active area of current research in statistics. Common methods include the lasso (Tibshirani, 1994), adaptive lasso (Zou, 2006), elastic net (Zou and Hastie, 2005), among many others (Fan and Li, 2001; Meier et

al., 2008; Zhou et al. 2009; Zhu and Hastie, 2004). In this paper, the focus of the simulation studies is on the lasso logistic regression, with selection of the penalty parameter via the cross-validated error rate. Our sample size methodology can be used with other regularized logistic regression methods, but may require modifications, particularly if additional layers of resampling are involved (e.g., the adaptive lasso).

The paper is organized as follows: Section 2.2 presents the methodology. Section 2.3 presents the results of simulation studies. Section 2.4 presents the results of real data analysis and resampling studies. Section 2.5 presents discussion and conclusions.

2.2 METHODS

2.2.1 THE PENALIZED LOGISTIC REGRESSION MODEL

Each individual in a population $\mathcal{P}$ belongs to one of two classes, $\mathcal{C}_0$ and $\mathcal{C}_1$. For individual i , let $Y_i = 0$ if $i \in \mathcal{C}_0$ and $Y_i = 1$ if $i \in \mathcal{C}_1$. One wants to predict Y_i based on observed high dimensional data $g_i \in \mathbb{R}^p$ and clinical covariates $z_i \in \mathbb{R}^q$. A widely used model for this setting is the linear logistic regression model,

$$\pi(g_i, z_i) = P(Y_i = 1 | g_i, z_i) = \{1 + \text{Exp}[-\alpha - \delta'z_i - \gamma'g_i]\}^{-1} \quad (2.2.1)$$

where $\alpha \in \mathbb{R}^1$, $\delta \in \mathbb{R}^q$, and $\gamma \in \mathbb{R}^p$ are population parameters.

The negative log-likelihood, given observed data (y_i, z_i, g_i) for $i = 1, \dots, n$ is

$$L(\alpha, \delta, \gamma) = - \sum_{i=1}^n \{y_i \ln[\pi(g_i, z_i)] + (1 - y_i) \ln[1 - \pi(g_i, z_i)]\}.$$

To estimate parameters and reduce the dimension of g_i , a regularized regression is often fit. Coefficients are set to zero using the penalized negative log-likelihood function

$$L_{\text{penalized}}(\alpha, \delta, \gamma) = L(\alpha, \delta, \gamma) + \sum_{k=1}^p \lambda_k f(\gamma_k), \quad (2.2.2)$$

where λ_k are penalty parameters, and f is a loss function. If $f(\gamma_k) = |\gamma_k|$ and $\lambda_k \equiv \lambda > 0$, then the result is lasso logistic regression (Tibshirani, 1994). The first step of the lasso is to

estimate the penalty parameter λ , which is typically done by cross-validation. The clinical covariates z_i are not part of the feature selection process in Equation 2.2.2, but they can be added to that process if desired. The regularized regression estimates are the solutions to:

$$(\hat{\alpha}, \hat{\delta}, \hat{\gamma}) = \min_{\alpha, \delta, \gamma} L_{\text{penalized}}(\alpha, \delta, \gamma). \quad (2.2.3)$$

The minimum can be found by the coordinate descent algorithm (Friedman et al., 2008).

2.2.2 PREDICTED CLASSIFICATION SCORES

Consider a training set and independent validation set. The training set is

$$T_j = \{(y_1, z_1, g_1), \dots, (y_n, z_n, g_n)\}$$

and the validation set is

$$V_k = \{(y_1^v, z_1^v, g_1^v), \dots, (y_m^v, z_m^v, g_m^v)\}.$$

The minimization in Equation 2.2.3 based on the dataset T_j produces estimates $(\hat{\alpha}_j, \hat{\delta}_j, \hat{\gamma}_j)$.

The model is applied to the validation set V_k resulting in estimated scores

$$\left\{ \hat{\alpha}_j + \hat{\delta}_j z_i^v + \hat{\gamma}_j g_i^v \right\}_{i=1}^m.$$

Let $W_{ij}^u = \hat{\gamma}_j g_i^v$ be the high dimensional part of the predicted classification score for individual i in the validation set. Let $X_i^u = \gamma' g_i^v$ and note that we can write $W_{ij}^u = X_i^u + U_{ij}^u$ where $U_{ij}^u = (\hat{\gamma}_j - \gamma)' g_i^v$. (The u superscripts denote unstandardized variables, in contrast to standardized versions presented below.) The model of Equation 2.2.1 can be written in the form:

$$P(Y_i^v = 1 | z_i, g_i) = \{1 + \text{Exp}[-\alpha - \delta' z_i^v - X_i^u]\}.$$

Note that, unlike the standard logistic regression model, the variable X_i^u does not have a slope parameter multiple. We could develop the model in its present form, but it will simplify presentation if we make it look more like the standard model.

Define $\mu_x = E_{\mathcal{P}}[X_i^u] = \int \gamma' g f(g) d\mu$ as the mean of the $\gamma' g_i$ taken across the target population $\mathcal{P}$, where the high dimensional vectors have density f with respect to a measure μ . Similarly, define $\sigma_x^2 = Var_{\mathcal{P}}(X_i^u)$. If these exist, then we can standardize the scores

$$X_i = \frac{X_i^u - \mu_x}{\sigma_x}, U_{ij} = \frac{U_{ij}^u}{\sigma_x}, W_{ij} = \frac{W_{ij}^u - \mu_x}{\sigma_x} = X_i + U_{ij} \quad (2.2.4)$$

resulting in the EIV logistic regression model,

$$\begin{aligned} P(Y_i^v = 1 | z_i, X_i) &= \{1 + \exp[-\alpha - \delta' z_i^v - \mu_x - \sigma_x X_i]\} \\ &= \{1 + \exp[-\alpha_x - \delta' z_i^v - \beta_{\infty} X_i]\} \end{aligned}$$

where $\alpha_x = \alpha + \mu_x$, $\beta_{\infty} = \sigma_x$. Note that $E_{\mathcal{P}}[X_i] = 0$ and $Var_{\mathcal{P}}(X_i) = 1$. This is the EIV model of Carroll et al. (2006). With these adjustments, we can apply EIV methods in a straightforward way.

Suppose we repeatedly draw training sets T_j at random from the population $\mathcal{P}$, resulting in $T_1, T_2, \dots$. Each time we apply the developed predictor to the validation set V_k , producing a vector of error values $U_1, U_2, \dots$ where $U_t = (U_{1t}, \dots, U_{mt})'$. Define $E_n[U_t] = \lim_{t_0 \rightarrow \infty} \frac{1}{t_0} \sum_{t=1}^{t_0} U_t$, and $Var_n(U_t) = \lim_{t_0 \rightarrow \infty} \frac{1}{t_0-1} \sum_{t=1}^{t_0} (U_t - E_n[U_t])(U_t - E_n[U_t])'$. The derivation of the conditional score method is based on an assumption that the U_t are independent and identically distributed Gaussian with $E_n[U_t] = 0$ and $Var_n(U_t) = \Sigma_{uu}$ where Σ_{uu} is a positive definite matrix. This assumption can be divided into three component parts:

- 1:** The $E_n[U_{ij}|g_i] = 0$ for $i = 1, \dots, m$. Equivalently, $E_n[W_{ij}|g_i] = X_i$, so that the estimated values are unbiased estimates of the population values. Intuitively, if n_{train} is large enough to develop a good classifier, then this assumption should be approximately true. However, if n_{train} is much too small, then the estimated scores may be more or less random and not centered at the true values – so that this assumption would be violated. But the assumption is required for identifiability (Dobbin and Song, 2013).

This shows that some model violation may be expected for our approach as the sample size gets small.

- 2:** The U_{ij} have finite variance. This would be true if $g_i'Var_n(\hat{\gamma}_j)g_i < \infty$ for each i . So, if the regularized linear predictor $\hat{\gamma}_j$ has finite second moments for training samples of size n , the condition would be satisfied.
- 3:** The vector $(U_{1j}, \dots, U_{mj})$ is multivariate normal. This means that given $G_{mat} = (g_1, \dots, g_m)$, $(\hat{\gamma}_j - \gamma)'G_{mat}$ is multivariate normal. This would be true if $\hat{\gamma}_j$ were multivariate normal, and may be approximately true if conditions under which $\hat{\gamma}_j$ converges to a normal distribution are satisfied (e.g. Buhlmann and van de Geer (2011)).

To further simplify the model, we assume $Var(U_j) = \sigma_n^2 R_n$ where R_n is a correlation matrix; in other words, we assume the prediction error variance is the same for each individual i .

2.2.3 DEFINING THE OBJECTIVE

Define β_j as the slope (associated with the W_{ij}) from fitting a logistic regression of Y_i on (z_i, W_{ij}) across the entire population $\mathcal{P}$. The tolerance is:

$$Tol(n) = |\beta_\infty - E_n[\beta_j]|.$$

Under regularity conditions the tolerance will be finite and $|E_n[\beta_j]| < |\beta_\infty|$, and $\lim_{n \rightarrow \infty} Tol(n) = 0$ (Supplement Section 5.1). Note that it is possible to have $|E_n[\beta_j]| > \beta_\infty$ in logistic EIV (Stefanski and Carroll, 1987). Let t_{target} be the targeted tolerance. The targeted sample size n_{target} is the solution to:

$$n_{target} = \min\{n | Tol(n) \leq t_{target}\}.$$

2.2.4 ESTIMATION

Resampling is used to search for n_{target} nonparametrically. This section outlines each step in the estimation process. More detailed descriptions appear in the supplement.

2.2.4.1 ESTIMATION FOR THE FULL PILOT DATASET

Let n_{pilot} be the size of the pilot dataset. The parameter $\beta_{n_{pilot}} = E_{n_{pilot}}[\beta_j]$ defined in Section 2.2.3 can be estimated by cross-validation (e.g., Geisser, 1993). Regularized logistic regression requires specification of a penalty parameter (λ in Equation 2.2.2). Selecting this penalty parameter once using the whole dataset results in biased estimates of predicted classification performance (Ambroise and McLachlan, 2002; Simon et al. (2003)). Therefore, a nested (double) cross-validation is required (see, e.g., Davison and Hinkley (1997)). An inner loop is used to select the penalty parameter λ ; then that penalty parameter is used in the outer loops to obtain the cross-validated classification scores. Because the split of the dataset into 5 subsets may impact the resulting nested CV slope estimate, we suggest the RCS-CV method; RCS-CV is defined as repeating the cross-validation 20-50 times, centering and scaling each cross-validated batch, and using the mean of these 20-50 cross-validated slopes as the estimate. Centering and scaling of the cross-validated batches is needed to reduce error variance due to instability in the lasso regression parameter estimates (not shown). We recommend 5-fold cross-validation.

The cross-validated scores provide an estimate of the slope for a training sample of size n_{pilot} , which we can denote $\hat{\beta}_{n_{pilot}}$. We want to apply errors-in-variables regression to estimate the tolerance, $Tol(n_{pilot})$, and for that we also need an estimate of the error variance, $\sigma_{n_{pilot}}^2 = Var_{n_{pilot}}(U_{ij})$. The leave-one-out bootstrap (e.g., Efron and Tibshirani (1997)) can be used to estimate $\sigma_{n_{pilot}}^2$. Because tuning parameters must be selected in regularized regression, a nested, case-cross-validated leave-one out bootstrap (LOO-NCCV-BS) is required (see, e.g., Varma and Simon (2006)). Letting $W_{ij,bs}$ represent these bootstrap scores for $i = 1, \dots, n_{pilot}$ and $j = 1, \dots, b_0$, where b_0 is the number of bootstraps for each left-out case, then the estimate of $\sigma_{n_{pilot}}^2$ is

$$\hat{\sigma}_{n_{pilot}}^2 = \frac{1}{n_{pilot}(b_0 - 1)} \sum_{i=1}^{n_{pilot}} \sum_{j=1}^{b_0} (W_{ij,bs} - \bar{W}_{i,\cdot,bs})^2$$

where $\overline{W}_{i,\cdot,bs} = \frac{1}{b_0} \sum_{j=1}^{b_0} W_{ij,bs}$. As with the CV described in the previous paragraph, one needs to standardize the cross-validated bootstrap batches to have mean zero and variance 1. This is the CS-LOO-NCCV-BS procedure. Note that in practice the leave-one-out bootstrap is performed using a single bootstrap and collating the results appropriately, which reduces the computation cost (Davison and Hinkley, 1997).

Now the $\hat{\sigma}_{n_{pilot}}^2$ is “plugged in” to a univariate EIV logistic regression which also uses the nested CV predicted classification scores as the W_{ij} in Equation 2.2.4. The conditional score method of Stefanski and Carroll (1987), with the (Hanfelt and Liang (1997), Eqn. 3) modification is used to estimate the asymptotic slope β_∞ associated with the X_i . Briefly, if we write the logistic density of Equation 2.2.3 in the canonical generalized linear model form

$$f(y_i) = \text{Exp} \left\{ \frac{y_i(\alpha + \delta' z_i + \beta_\infty X_i) - b(\alpha + \delta' z_i + \beta_\infty X_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

where the functions are $a(\phi) = 1$, $b(x) = \ln(x)$, $c(y_i, \phi) = 0$; then letting $\theta = (\alpha, \delta, \beta_\infty)'$ the conditional score function for θ has the form,

$$\sum_i \begin{pmatrix} (y_i - E[y_i | A_{\theta_i}]) \\ z_i(y_i - E[y_i | A_{\theta_i}]) \\ \tilde{x}_i(y_i - E[y_i | A_{\theta_i}]) \end{pmatrix}$$

where $A_{\theta_i} = W_{ij} + y_i \Psi \beta_\infty$, $\tilde{x}_i$ is an estimator of X_i based A_{θ_i} , and $\Psi = \text{Var}_{n_{pilot}}(U_{ij})a(\phi)$.

The conditional score method produces $\hat{\beta}_\infty$. The tolerance is then estimated with $\widehat{Tol}(n_{pilot}) = |\hat{\beta}_\infty - \hat{\beta}_{n_{pilot}}|$.

2.2.4.2 ESTIMATING TOLERANCE FOR SUBSETS OF THE PILOT DATASET

Typically, $\widehat{Tol}(n_{pilot})$ will be larger or smaller than t_{target} , the targeted tolerance. In either case, more information about the relationship between $Tol(n)$ and n is needed to estimate n_{target} . Such information can be obtained by subsampling from the pilot dataset. We suggest 7 subsets with a range of sizes be taken from the pilot dataset. Each subset should be large

enough, as defined in Section 2.2.5 below. For example, $n_{pilot} \times k/7$ for $k = 1, \dots, 7$ can be used. More typically, if the pilot dataset is not as large, then one may use $(n_{pilot}/2) + k/6 * (n_{pilot}/2)$ for $k = 0, \dots, 6$. If $n_{pilot}/2$ is not large enough, then the pilot set is probably inadequate.

For each subset size less than n_{pilot} , call them $n_1^*, \dots, n_6^*$, take a random sample from the full dataset without replacement. Then apply the procedure described for the full pilot dataset to each subset and obtain $\widehat{Tol}(n_k^*)$, $k = 1, \dots, 6$. The only modification we suggest to the original RCS-CV procedure is that for the 20-50 repetitions, take a different random sample each time.

2.2.4.3 ESTIMATION OF $\hat{n}_{target}$

Analysis of a pilot or simulated dataset produces sample sizes $n_1^* < n_2^* < \dots < n_7^*$, and corresponding tolerance estimates $\hat{t}_1 = \widehat{Tol}(n_1^*), \dots, \hat{t}_7 = \widehat{Tol}(n_7^*)$. Fit the Box-Cox regression model $(\hat{t}_i^\lambda - 1)/\lambda = \delta_0 + \delta_1 n_i^* + \epsilon_i$ to obtain $\hat{\lambda}$, and define $\hat{h}(x) = (x^{\hat{\lambda}} - 1)/\hat{\lambda}$. Then fit with least squares $n_i^* = \eta + \zeta \hat{h}(\hat{t}_i)$ which produces $\hat{\eta}$ and $\hat{\zeta}$. Finally, if t_{target} is the desired tolerance, the sample size is:

$$\hat{n}_{target} = \hat{\eta} + \hat{\zeta} \hat{h}(t_{target}).$$

As discussed above, we recommend estimating each tolerance 20-50 times by repeated random sampling, say $\hat{t}_{1,1}, \dots, \hat{t}_{1,20}$ and estimating $\hat{t}_1 = \frac{1}{20} \sum_{i=1}^{20} \hat{t}_{1,i}$.

2.2.5 ARE THERE ENOUGH SAMPLES IN THE PILOT DATASET?

The nested resampling methods in our approach require there be adequate numbers in the subsets. If there are n_{pilot} in the pilot dataset, then a bootstrap sample will contain on average $0.632 \times n_{pilot}$ unique samples. A 5-fold case cross-validation of the bootstrap sample will result in $0.2 \times 0.632 \times n_{pilot} = 0.13 \times n_{pilot}$ in a validation set. Since the validation set scores will be normalized to have mean zero and variance 1, we recommend at least 80 samples in the

training set to ensure at least 10 samples in these cross-validated sets. If the class prevalence is imbalanced, this number should be increased. In particular, we recommend:

Condition 1: If n_{pilot} is the size of the pilot dataset, then $n_{pilot} \times 0.13 \times \pi_{lowest} \geq 5$, where π_{lowest} is the proportion from the under-represented class.

The conditional score methods will not work well as the error variance gets large. Since the conditional score methods are repeated 20-50 times for each subset size in the RCS-CV procedure, the stability of these estimates can be evaluated. Therefore, the following guideline is advised:

Practical guideline 1: If the conditional score errors-in-variables regression estimates display instability for any subsample size, use quadratic SIMEX errors-in-variables regression instead. An example of instability would be $\left| mean(\hat{\beta}_{\infty}) / s.d.(\hat{\beta}_{\infty}) \right| < .5$, where the mean and standard deviation are taken across the 20-50 replicates.

Resampling-based approaches to sample size estimation require that the relationship between the asymptotic model and the estimated model can be adequately estimated from the pilot dataset. Trouble can arise if the learning pattern displayed on the pilot set changes dramatically for sample sizes larger than the pilot dataset. For example, there may be no classification signal detectable with 3 samples per class, but one detectable with 50 samples per class. So a pilot dataset of 6 would lead to the erroneous conclusion that the asymptotic error rate is 50%, and any resulting sample size estimates would likely be erroneous. Similarly, the learning process can be uneven, so that the asymptotic error rate estimate increases or decreases as the sample size increases. This latter can happen when some subset of the features have smaller effects than others and are only detected for larger sample sizes. To guard against this in simulations, at least, we found that the following guideline is useful:

Condition 2: The predictor needs to find the important features related to the class distinction with power at least 85%.

Our simulation-based software program checks the empirical power for this condition. In the context of resampling from real data, it is not clear how one could verify this assumption empirically. But it may be possible to evaluate the effect size associated with this power by a parametric bootstrap.

2.2.6 TRANSLATING BETWEEN LOGISTIC SLOPE AND MISCLASSIFICATION ACCURACY

If there are no clinical covariates in the model, then the misclassification error rate for the asymptotic model is (e.g., Efron (1975))

$$P(Y_i = 1 \text{ and } \alpha + \beta_\infty X_i \leq 0) + P(Y_i = 0 \text{ and } \alpha + \beta_\infty X_i > 0) \\ = \int_{-\infty}^{-\alpha/\beta_\infty} \frac{e^{\alpha+\beta_\infty X_i}}{1 + e^{\alpha+\beta_\infty X_i}} f_x(x) dx + \int_{-\alpha/\beta_\infty}^{\infty} \frac{1}{1 + e^{\alpha+\beta_\infty X_i}} f_x(x) dx$$

where $f_x(x)$ is the marginal density of the asymptotic scores across the population $\mathcal{P}$. By definition, these scores have mean zero and variance one. If we further assume the scores are Gaussian, then the misclassification rate can be estimated with

$$\sum_{i=1}^{m_0} \left[\frac{e^{\alpha+\beta_\infty x_i}}{1 + e^{\alpha+\beta_\infty x_i}} \right] 1(x_i \leq -\alpha/\beta_\infty) + \sum_{i=1}^{m_0} \left[\frac{1}{1 + e^{\alpha+\beta_\infty x_i}} \right] 1(x_i > -\alpha/\beta_\infty)$$

where $1(A)$ is the indicator function for event A , and m_0 is a number of Monte Carlo simulations, and $x_1, \dots, x_{m_0}$ are drawn from the distribution $x_i \sim Normal(0, 1)$. If covariates are added to the model, then the conditional distribution of $x_i|z_i$ needs to be used for the Monte Carlo. If x_i is independent of z_i , then the x_i could be generated from a standard normal, and the Monte Carlo equations modified in the obvious way. In the Supplement is a graph of the no covariate case relationship.

2.3 SIMULATION STUDIES

For the simulation studies, high dimensional data were generated from multivariate normal distributions, both a single multivariate normal and a mixture multivariate normal with

Table 2.2.1: Estimates of the asymptotic slope β_∞ and corresponding accuracy acc_∞ evaluated by simulations. n_{pilot} is the number of samples in the pilot dataset. The covariance structure “Cov” are: AR1 is block autoregressive order 1 in 3 blocks of size 3 (9 informative features) with parameter 0.7; Iden. is identity with 1 block of 1 (1 informative feature). Total of $p = 500$ features; all noise features independent standard normal. Summary statistics based on 200 Monte Carlo. More results appear in the Supplement.

n_{pilot}	Cov	Logist. slope		Class. Acc.		mean $\hat{\sigma}_n^2$	mean $\hat{\beta}_n$
		β_∞	mean $\hat{\beta}_\infty$	acc_∞	$\widehat{acc}_\infty$ mean		
300	AR1	2.0	2.07	0.778	0.783	0.43	1.49
400	AR1	2.0	2.01	0.778	0.779	0.35	1.62
300	AR1	3.0	3.04	0.836	0.838	0.32	2.31
400	AR1	3.0	2.93	0.836	0.834	0.27	2.47
300	AR1	4.0	3.95	0.871	0.869	0.28	3.06
400	AR1	4.0	3.88	0.871	0.868	0.23	3.26
300	AR1	5.0	3.77	0.894	0.865	0.25	3.71
400	AR1	5.0	4.81	0.894	0.891	0.21	3.99
300	Iden.	2.0	2.05	0.778	0.781	0.23	1.87
400	Iden.	2.0	2.01	0.778	0.778	0.19	1.90
300	Iden.	3.0	3.02	0.836	0.836	0.17	2.85
400	Iden.	3.0	2.98	0.836	0.835	0.14	2.87
300	Iden.	4.0	3.97	0.871	0.870	0.14	3.72
400	Iden.	4.0	3.92	0.871	0.869	0.12	3.75
300	Iden.	5.0	4.94	0.894	0.893	0.14	4.50
400	Iden.	5.0	4.86	0.894	0.891	0.12	4.55

homoscedastic variance. Both multivariate normal settings performed similarly (see Supplement), so we just present one in the paper. The covariance matrices were identity, compound symmetric (CS), and autoregressive order 1 (AR(1)), as indicated. Class labels were generated from the linear logistic regression model of Equation 2.2.1. Categorical clinical covariate data, when included in simulations, were generated from a distribution with equal probability assigned to each of three categories, where categories are correlated with class labels.

The asymptotic slope parameter β_∞ must be estimated. Table 2.2.1 presents a simulation to evaluate the bias and variance of the asymptotic slope parameter estimate $\hat{\beta}_\infty$. Also

Table 2.2.2: Evaluation of the sample size estimates from AR(1) and identity covariances. The number in the pilot dataset is 400. $\beta_\infty = 4$. Identity covariance had one informative feature, and AR(1) had nine informative features in a block structure of 3 blocks of size 3 with correlation parameter 0.7. Estimates evaluated using 400 Monte Carlo simulations with the estimated sample size. The mean tolerance from the 400 simulations, and the proportion of the 400 within the specified tolerance are given in the rightmost two columns. The dimension is $p=500$.

Cov.	t_{target}	$\hat{n}$	Mean MC tol	% of MC within tol
AR1	0.10	1742	0.09	64%
AR1	0.20	986	0.19	62%
AR1	0.30	715	0.27	67%
AR1	0.40	573	0.34	71%
AR1	0.50	484	0.43	72%
AR1	0.60	424	0.49	75%
AR1	0.70	380	0.57	77%
Identity	0.10	509	0.09	79%
Identity	0.20	322	0.10	87%
Identity	0.30	242	0.15	87%
Identity	0.40	194	0.16	92%
Identity	0.50	162	0.21	90%
Identity	0.60	139	0.25	93%
Identity	0.70	121	0.31	91%

presented are the corresponding estimates of asymptotic classification accuracy $\widehat{acc}_\infty$. As can be seen from the table, this approach does well overall at estimating the asymptotic performance for these pilot dataset sample sizes (300 and 400), asymptotic slopes (2,3,4,5), multivariate normal high dimensional data, covariance matrix structures (Identity, CS (supplement) and AR1), and numbers of informative features (1 and 9). There is some small bias apparent as the slope becomes large ($\beta_\infty = 5$), probably reflecting the fact that large slopes are problematic for EIV logistic regression.

The tolerance associated with the estimated sample size should be within the user-targeted tolerance. To test this, sample sizes were calculated by applying the method to

Table 2.2.3: Clinical covariate simulations. One clinical covariate with 3 levels which are associated with the class distinction. The identity covariance and an asymptotic true slope of $\beta_\infty = 4$. The dimension is $p=500$. See text for more information.

t_{target}	$\hat{n}$	Mean MC tol.	% of MC within tol.
0.10	592	0.07	80%
0.20	416	0.09	88%
0.30	334	0.12	92%
0.40	284	0.14	91%
0.50	249	0.15	95%
0.60	223	0.17	96%
0.70	203	0.19	96%

simulated pilot datasets. Then, these sample size estimates were assessed by performing very large pure Monte Carlo studies. Table 2.2.2 presents sample size estimates from our method and sample statistics from the Monte Carlo (MC) simulations. The mean tolerances from the MC are all within the targeted tolerance, indicating that the estimated sample sizes do achieve the targeted tolerance. The method tends to produce larger sample size estimates than required with 62%-93% of the true tolerances within the target (rightmost column). Note that our method guarantees that the expected slope is within the tolerance, but not that the actual slope is within the tolerance; this latter would be a stronger requirement.

Implementation of our approach in the presence of clinical covariates was evaluated. Table 2.2.3 shows results when a clinical covariate is included into the setting. In this case the clinical covariate is also associated with the class distinction; in particular, in Equation 2.2.1, $\delta = Ln(2)$ and $z_i \in \{-1, 0, 1\}$, with 1/3rd probability assigned to each value. As can be seen by comparison with Table 2.2.2, the addition of the clinical covariate increases the required sample sizes. For example, the estimated sample size for a tolerance of 0.20 increases 29%, from 322 to 416. This increase reflects correlation between the clinical covariate and

Table 2.2.4: Table of sample size estimates. “EIV” is the method presented in this paper. “LC” is Mukherjee et al.’s (2003) method. For the LC method, constrained the optimization to the feasible region using the L-BFGS-B algorithm in optim in R v.3.0.2. “Truth” values obtained from pure Monte Carlo. Datasets are: 1) Identity Cov, nPilot=300, slope=3; 2) AR1 covariance, nPilot=400, slope=3; 3) AR1 covariance, nPilot=400, slope=4; 4) Identity covariance, nPilot=300, slope=4; 5) CS covariance, npilot 400, slope 3; 6) CS covariance, nPilot=400, slope=4.

Dataset	Tolerance									
	Tol=0.1					Tol=0.2				
	EIV	LC	Truth	err %	err %	EIV	LC	Truth	err %	err %
				EIV	LC				EIV	LC
1	383	1,407	293	31%	380%	251	598	152	65%	293%
2	682	9,341	1,023	-33%	813%	540	3,249	706	-24%	360%
3	1,742	44,840	1,633	7%	2,646%	986	26,062	950	4%	2,643%
4	509	2,904	398	28%	629%	369	1,639	160	131%	924%
5	685	7,750	1,027	-33%	655%	542	2,870	696	-22%	312%
6	1,460	4,908	1,579	-8%	211%	855	1,596	940	-9%	70%

the class labels. The pure Monte Carlo evaluations in Table 2.2.2 show that the method does still produce adequate sample size estimates in the presence of the clinical covariate.

Figure 2.2.1 is a summarization of results from all the different simulation studies. Negative values on the y-axis mean the sample size was overestimated, and positive values mean the sample size was underestimated. As can be seen in the figure, the sample size estimates are mostly adequate or conservative. When the estimated sample size required is smaller than the pilot dataset (x-axis values are negative), the resulting tolerance estimates are adequate or conservative; intuitively, identifying a sample size smaller than the pilot dataset should be relatively easy. When the estimated sample size required is larger than the pilot dataset, the method continues to perform well overall. The exceptions are in the cases of compound symmetric and AR1 covariance with a small slope of 3; in these cases, the y-values are positive, indicating anti-conservative sample size estimates. The problem here seems to be the power to detect the features. For the compound symmetric simulations, the empirical

Table 2.2.5: Resampling studies. Dataset is the dataset used for resampling. Rep is the replication number of 5 independent random subsamples (without replacement) of size n_{Pilot} . n_{Full} is the size of the full dataset. Classes for the Shedden dataset were Alive versus Dead. Classes for Rosenwald dataset were Germinal-Center B-Cell lymphoma type versus all others. $err(n_{Full})$ is estimated from 200 (50 for Shedden) random cross-validation estimations on the full dataset using different partitions each time, and this serves as the gold standard error rate for n_{Full} . $\widehat{err}(n_{Full})$ is the estimated error rate for the full dataset based on LC method or EIV method. Similarly, $\widehat{err}(\infty)$ is the asymptotic error rate based on the LC method or EIV method. The first column is the dataset, “R” for Rosenwald and “S” for Shedden. For the Shedden dataset, used conditional score EIV; for Rosenwald dataset, we used quadratic SIMEX EIV because the criterion for conditional score was violated (Section 2.2.5).

Data	Rep	nPilot	nFull	err (n_{Full})	LC method		EIV Method		nFull err %	
					$\widehat{err}$ (n_{Full})	$\widehat{err}(\infty)$	$\widehat{err}$ (n_{Full})	$\widehat{err}(\infty)$	LC	EIV
R	1	100	240	0.1129	0.0855	0.0729	0.1344	0.1135	-25%	19%
R	2	100	240	0.1129	0.0611	0.0435	0.1078	0.0933	-46%	-5%
R	3	100	240	0.1129	0.0298	0.0089	0.0771	0.0691	-74%	-32%
R	4	100	240	0.1129	0.1443	0.1270	0.1396	0.1379	28%	24%
R	5	100	240	0.1129	0.0682	0.0480	0.0864	0.0783	-40%	-23%
mean				0.1129	0.0778	0.0601	0.1091	0.0984	-31%	-3%
S	1	200	443	0.4207	0.4638	0.4634	0.4347	0.4347	10%	3%
S	2	200	443	0.4207	0.4496	0.4481	0.4154	0.4151	7%	-1%
S	3	200	443	0.4207	0.4300	0.4258	0.2778	0.2778	2%	-34%
S	4	200	443	0.4207	0.4166	0.4126	0.3550	0.3550	-1%	-16%
S	5	200	443	0.4207	0.4159	0.4117	0.2907	0.2894	-1%	-31%
mean				0.4207	0.4352	0.4323	0.3548	0.3544	3%	-16%

bootstrap power was $7.67/9=85.2\%$, and for AR1 simulations the power was 84.7% . Both are near the cut off of the 85% power criterion developed in Section 2.2.5 above. Still, overall, the method seems to perform well.

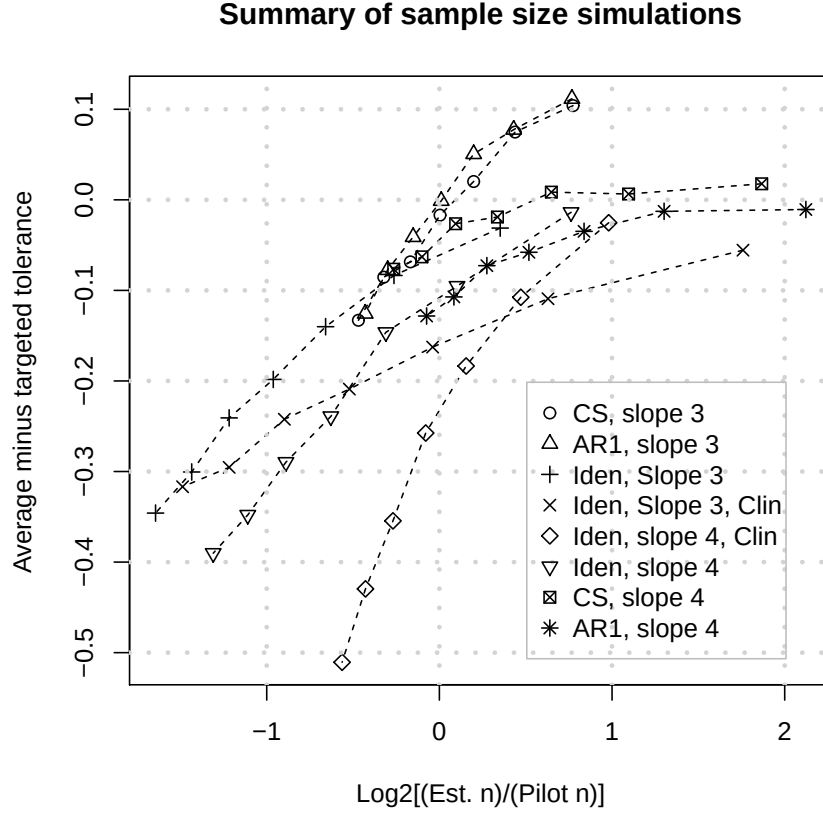


Figure 2.2.1: Summary of results of simulations. The x-axis is the base 2 logarithm of the ratio of the estimated training sample size required divided by the pilot training sample size used. The y-axis is the average tolerance estimated from pure Monte Carlo simulations minus the targeted tolerance.

2.3.0.1 LC vs. EIV IN SIMULATIONS

We compared our resampling-based method and the resampling-based method of Mukherjee et al. (2003) to a pure Monte Carlo estimation of the truth in simulations. We will denote their method by LC (for learning curve) and our method by EIV (for errors-in-variables). Table 2.2.4 shows a comparison of the two methods under a range of simulation settings. In

these simulations, our method may have an advantage because the logistic regression model was used to generate the response data. Tolerances of 0.1 and 0.2 were considered since these are associated with larger training sample sizes than the pilot dataset. Comparing the percentage error of the sample size estimate to an estimate based on pure Monte Carlo, one can see that the learning curve method has an error an order of magnitude or more larger than our method. The LC method tends to consistently overestimate the sample size in these simulations. In sum, the EIV method estimates were closer to the true sample size values than the LC method estimates across all of these simulations.

2.4 REAL DATASET ANALYSES

Any real dataset developed from physical experiments will be imperfectly represented in simulated models. Therefore, we evaluate robustness to real data violations by studying the performance of the LC and EIV methods on resampled real datasets.

2.4.1 RESAMPLING STUDIES OF MICROARRAY DATASETS

The purpose of a resampling study is to compare estimates from a procedure to a resampling-based “truth.” Since adequate sample sizes are unknown on these datasets, it was not feasible to compare sample size estimates to any corresponding estimated true values. But we can compare the error rate estimated from a subset of the dataset, to an independent estimate based on cross-validation on the whole dataset, and see whether LC or EIV is closer to this “truth.”

We applied the EIV and the LC method to the dataset of Rosenwald et al. (2002). The classes were germinal center B-cell lymphoma versus all other types of lymphoma. We subsetted 5 “pilot datasets” of size 100 at random from this dataset. For each of these “pilot datasets,” we estimated the performance when $n = 240$ are in the training set. Then, we could compare the estimated performance to a “gold standard” resampling-based performance on

the full dataset. Results are shown in Table 2.2.5. As can be seen from the table, the EIV method has better mean performance in terms of estimating the full dataset error than the LC method. Here, the differences are less dramatic than the sample size differences; this may be due to the sensitivity of sample size methods to relatively small changes in asymptotic error rates, or to the underlying data distribution. Both methods show some variation in error rate estimates across the five datasets.

We next applied both methods to the lung cancer dataset of Shedden et al. (2008), where the classes were based on survival status at last follow-up. In this case, the two methods produce similar results. The LC method was slightly better on average than the EIV with conditional score based on percentage error (rightmost columns of table); but the conditional score criterion in Section 2.2.5 was exceeded on 3 of the 5 datasets, and if quadratic SIMEX is used then the LC and EIV are almost identical (Supplement). This is a very noisy problem and classification accuracy based on a training set of all 443 samples is only estimated to be around 56%-60%. Both methods indicated that the full dataset error rate is very close to the optimal error rate across all 5 subset analyses.

2.4.2 RNASEQ APPLICATIONS

We performed a proof of principle study to see if these methods could be applied effectively to RNA-seq data. First, note that RNA-seq data after being processed may be in the form of counts (e.g., from the Myrna algorithm), but are more often in the form of continuous values (e.g., normalized Myrna data, or FPKM fragments-per-kilobase of exon per million fragments mapped from Cufflinks or other software). Therefore, linear models with continuous high dimensional predictors are reasonable to use for RNA-seq data. But it is important to check that the processed data appear reasonably Gaussian and, if not, to transform the data.

We applied the LC and EIV methods to the *Drosophila melanogaster* data of Graveley et al. (2011). Processed data were downloaded from ReCount database (Frazee et al., 2011).

Variables with more than 50% missing data were removed. Remaining data were truncated below at 1.5, and log-transformed. Low variance features were filtered out, resulting in $p=500$ dimensions. Since this was a highly controlled experiment with large biological differences between the fly states, some class distinctions resulted in separable classes. Logistic regression is not appropriate for perfectly separated data. Samples were split into two classes: Class 1 consisted of all the embryos and some adult and white prepupae (WPP); Class 2 consisted of all the larvae and a mix of adults and WPP. The class sizes were 82 and 65. A principal component plot is shown in the Supplement. The dataset consisted of a total of $n_{pilot} = 147$ data points. Technical replicates in the data created a clustering pattern visible in principal components plots. This type of clustering is often observed in real datasets due to disease subgroupings. We did not attempt to adjust the analysis for the technical replicates. The resulting EIV method equation for the sample size was,

$$\hat{n} = 105.73 - 14.25 \left(\frac{t^{-0.3434} - 1}{(-0.3434)} \right).$$

For tolerances of 0.1, 0.05 and 0.02, sample size estimates were 156, 180, 223, respectively. The cross-validated accuracy, averaged over 10 replications, was 91%. Based on the $\hat{\beta}_{\infty} = 4.55$, the optimal accuracy is 88.5%, and the full dataset accuracy is 88%, corresponding to $\hat{\beta}_{147} = 4.42 = 4.55 - \widehat{Tol}(147)$. The conditional score was used for the EIV method. The LC method curve was $err = 0.075 + 1.252 \times n^{-0.9122321}$. The asymptotic accuracy estimate is 92.5%, corresponding to $\beta_{\infty} = 7.2$, and the estimated accuracy when $n=147$ is 91.2%, corresponding to $\beta_n = 6.1$. The LC sample size estimates for tolerances of 0.10, 0.05 and 0.02 were 1,383, 8,361, and 13,342, respectively. As with the simulation studies, the LC method estimates are much larger than the EIV estimates.

2.5 DISCUSSION

In this paper a new sample size method for training regularized logistic regression-based classifiers was developed. The method exploits a structural similarity between logistic prediction and errors-in-variables regression models. The method was shown to perform well when an adequate pilot dataset is available. Methods for assessing the adequacy of a pilot dataset were developed. If no adequate pilot dataset is available, the method can be used with Monte Carlo samples from a parametric simulation. The method was shown to perform well, and was compared with an existing method both on simulated datasets, resampled datasets and on an RNA-seq dataset.

We compared our method to a previously developed generic method. Our method provided better sample size estimates on simulated data, and seemed to provide more reasonable estimates on the RNA-seq data. This comparison is not quite “fair” to the method of Mukherjee et al. (2003) though since our method assumes the lasso logistic regression is used for model development, whereas the Mukherjee et al. (2003) method does not make this assumption. A future direction of this work is to compare these methods under other regularized regression models.

An important issue in using either the LC or EIV method is the fitting of the curve that produces the final sample size estimate. In the LC method, as described in Mukherjee et al. (2003), a constrained least squares optimization must be performed on a nonlinear regression model. Constrained optimization methods like the L-BFGS-B algorithm used in the application of the Mukherjee et al. (2003) method in this paper may produce different solutions than standard, unconstrained least squares optimization methods such as Nelder-Mead. In contrast, the Box-Cox algorithm and linear regression fitting used by our approach is more straightforward to implement. Because our method does not need to “extrapolate to infinity” as the typical learning curve method requires, the regression model is chosen that fits the best in the vicinity of the data points. This simplifies the fitting procedure albeit at

the cost of the errors-in-variables regression step. For both methods, it is advisable to look at the final plot of the fitted line and the data points as a basic regression diagnostic.

The reader may have noted that the variance parameter $\sigma_n^2 = \text{Var}_n(U_{ij})$ is estimated by bootstrapping the pilot dataset. But the variance is defined as a variance across independent training sets of size n in the population. Since the bootstrap datasets will have overlap, obviously there is potential bias in the bootstrap estimation procedure. Whether the bootstrap could be modified to reduce this bias is a potential area for future work.

If more than two classes are present in the data, then simple regularized logistic regression is no longer an appropriate analysis strategy. In order to apply our method in that setting, regularized methods for more than two classes would need to be developed; for example, regularized multinomial or ordinal logistic regression methods. Also, corresponding errors-in-variables methods for these multi-class logistic regression methods would be needed. It appears that both these would be pre-requisites to such an extension.

If classes are completely separable in the high dimensional space, then regularized logistic regression is not advisable because the logistic regression slope will be undefined and the logistic fitting algorithms will become unstable. The approach presented in this paper cannot be used in that context.

In this paper we have focused simulations on settings with equal prevalence from each class. If the class prevalences are unequal, then the method can still be applied as presented in the paper – as was done in the applications to the real datasets for example. However, if the imbalance is large (e.g., 90% versus 10%), then the training set size required by our Condition 1 in Section 2.2.5 would likely be excessive.

2.6 REFERENCES

- Bi, X., Rexer, B., Arteaga, C. L., Guo, M., and Mahadevan-Jansen, A. (2014). Evaluating her2 amplification status and acquired drug resistance in breast cancer cells using raman spectroscopy. *J Biomed Opt*, 19.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, New York.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2nd edition.
- Cook, J. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.
- Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, New York.
- de Valpine, P., Bitter, H., Brown, M., and Heller, J. (2009). A simulation-approximation approach to sample size planning for high-dimensional classification studies. *Biostatistics*, 10:424–435.
- Dobbin, K. K. and Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional dna microarray data. *Biostatistics*, 8(1):101–117.
- Dobbin, K. K., Zhao, Y., and Simon, R. M. (2008). How large a training set is needed to develop a classifier for microarray data. *Clinical Cancer Research*, pages 108–114.

- Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley, New York.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.
- Frazee, A., Langmead, B., and Leek, J. (2011). Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 12(449).
- Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011). The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479.
- Hanash, S., Baik, C., and Kallioniemi, O. (2011). Emerging molecular biomarkers – blood-based strategies to detect and monitor cancer. *Nature Reviews Clinical Oncology*, 8:142–150.
- Hanfelt, J. J. and Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82(3):pp. 461–477.
- Hanfelt, J. J. and Liang, K. Y. (1997). Approximate likelihoods for generalized linear errors-in-variables models. *Journal of the Royal Statistical Society. Series B*, 59:627–637.

- Huang, Y. and Wang, C. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, 96:1469–1482.
- Jung, S., Bang, H., and Young, S. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6:157–169.
- Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics*, 24(4):823–834.
- Li, S. S., Bigler, J., Lampe, J., Potter, J., and Feng, Z. (2005). Fdr-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine*, 15:2267–2280.
- Liu, P. and Hwang, J. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23:739–746.
- Moehler, T., Seckinger, A., Hose, D., Andrulis, M., Moreaux, J., Hielscher, T., Willhauck-Fleckenstein, M., Merling, A., Bertsch, U., Jauch, A., Goldschmidt, H., Klein, B., and Schwartz-Albiez, R. (2013). The glycome of normal and malignant plasma cells. *PLoS One*, 8:e83719.
- Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying dna microarray data. *Journal of Computational Biology*, 10(2):119–142.
- Novick, S. J. and Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481.

- Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024.
- Pfeffer, U., Romeo, F., Noonan, D. M., and Albin, A. (2009). Prediction of breast cancer metastasis by genomic profiling: where do we stand. *Clinical Exp Metastasis*, 26:547–558.
- Pounds, S. and Cheng, C. (2005). Sample size determination for the false discovery rate. *Bioinformatics*, 21:4263–4271.
- Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltner, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947.
- Shao, Y. and Tseng, C. H. (2007). Sample size calculation with dependence adjustment for fdr-control in microarray studies. *Statistics in Medicine*, 26:4219–4237.
- Simon, R. (2010). Clinical trials for predictive medicine: new challenges and paradigms. *Clinical Trials*, 7:516–524.
- Simon, R., Radmacher, M., Dobbin, K., and McShane, L. (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95:14–18.
- Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74:703–716.

- Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7:106.
- Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. theory of probability and its applications. *Theory of Probability and its Applications*, 16:264–280.
- Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91.
- Zhang, J. X., Song, W., Chen, Z. H., Wei, J. H., Liao, Y. J., Lei, J., Hu, M., Chen, G. Z., Liao, B., Lu, J., Zhao, H. W., Chen, W., He, Y. L., Wang, H. Y., Xie, D., and Luo, J. H. (2013). Prognostic and predictive value of a microrna signature in stage ii colon cancer: a microrna expression analysis. *Lancet Oncology*, 14:1295–1306.
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.
- Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming rna-seq data to improve the performance of prognostic gene signatures. *PLoS One*, 8:e85150.

CHAPTER 3

GENERAL SPARSE MULTI-CLASS LINEAR DISCRIMINANT ANALYSIS¹

¹Sandra Safo, Jeongyoun Ahn (2014+) General Sparse Multi-class Linear Discriminant Analysis.
To be submitted.

ABSTRACT

Discrimination with high dimensional data is often more effectively done with sparse methods that use a small fraction of predictors rather than using all the available ones. In recent years, some effective sparse discrimination methods based on Fisher’s linear discriminant analysis (LDA) have been proposed for binary class problems and extensions to the multi-class problems are in those works. However, drawbacks for the suggested way of extension include instances where the methods are unable to assign class labels, and also are computationally expensive for large number of classes. We propose an approach to generalize a binary LDA solution for multi-class problems while avoiding the limitations of the existing methods. Simulation studies with various settings confirms effectiveness of the proposed approach, as well as real data examples including a next generation sequencing data set and text classification examples.

KEYWORDS: High Dimension, Low Sample Size; Linear Discriminant Analysis; Multi-class Discrimination; Singular Value Decomposition; Sparse Discrimination.

3.1 INTRODUCTION

Classification is the process of allocating new entities into one of two or more existing classes. Fisher’s linear discriminant analysis (LDA) (Fisher, 1936) has been extensively studied and popular for many classification problems. If there are two class labels to predict, the population LDA discriminant vector, denoted by β , is expressed as

$$\beta \propto \Sigma^{-1}(\mu_1 - \mu_2), \quad (3.1.1)$$

where $\mu_i, i = 1, 2$ is a underlying mean of each class, respectively. The estimated discriminant vector $\hat{\beta}$ is calculated with estimated parameters based on the sample.

Even with the advent of high dimensional data, particularly, high-dimension-low-sample-size (HDLSS) data, LDA remains to be one of the most popular methods of choice for various classification problems, such as face recognition (Lu et al., 2003). However, it is well known that the original LDA suffers from singularity of the sample covariance matrix when applied to high dimensional data (Bickel and Levina, 2004; Ahn and Marron, 2010). Various regularized versions of LDA have been proposed over the recent years, most of which are intended for high dimensional applications. Some regularizations penalize the size of the coefficients in β , while others make β sparse. Sparse vectors have been shown to perform better at making predictions on real high dimensional datasets.

Substantial amount of effort has been made in the area of sparse LDA in recent years. Qiao et al. (2009) took a regression approach by solving a penalized least squares problem with lasso penalty. Clemmensen et al. (2011) considered optimal scoring approach of LDA and enforced sparsity using both lasso penalty and ridge penalty. Witten and Tibshirani (2011) applied lasso and fused lasso (Tibshirani et al., 2005) penalties and recasts the LDA problem to a biconvex one, for which they used minorization-maximization in order to solve. Note that these lasso-based sparse methods can only select up to n variables, which may not be enough when the true structure of data are not too sparse.

Some sparse LDA methods are motivated by the equation (3.1.1) directly, rather than modifying the LDA optimization problem. Cai and Liu (2011) noted that β solves the equation $\Sigma\beta = \mu_1 - \mu_2$ and proposed to directly estimate β , by using a similar approach to the Dantzig selector (Candes and Tao, 2007). Shao et al. (2011) assumed that both the common covariance matrix Σ and the mean difference vector $\mu_1 - \mu_2$ are sparse and suggested hard thresholding diagonals of the sample covariance matrix and entries of the sample mean difference. We note their approach does not necessarily yield a sparse estimate for β .

Unlike some classification methods such as the support vector machine (Hsu and Lin, 2002; Lee et al., 2004), the original LDA can be easily extended to multi-class situations,

by incorporating a between-class scatter in the place of the mean difference. This natural extension is one of the many reasons behind the popularity of LDA. When there are K classes, LDA produces up to $K - 1$ vectors that span the canonical discriminant subspace, which can be used for visual interpretation of the classification problem. Thus when developing its modified versions, it is desirable to take multi-class problems into account. Methods that are proposed by as Clemmensen et al. (2011) and Witten and Tibshirani (2011) are designed to be able to deal with multi-class problems in the same way as the binary case. However, it is not so straightforward for some other methods such as Cai and Liu (2011), Shao et al. (2011), and Mai et al. (2012).

A binary to multi-class (with K classes) extension method based on $\binom{K}{2}$ pairwise comparisons has been suggested by Cai and Liu (2011). Even though it is seemingly reasonable, this method has a few disadvantages. The most obvious one is that there are bound to be cases when a class label cannot be assigned because there is no dominant class in the comparisons. Furthermore, this approach cannot produce a $(K - 1)$ -dimensional discriminant subspace. Also, the method can be computationally intensive when K is large.

In this chapter, we present a general method with which one can generalize a binary LDA method to multi-class, and demonstrate our methodology on the pairwise comparison methods developed by Cai and Liu (2011) and Shao et al. (2011). We compare our method to the existing multi-class methods as well as the pairwise comparison approach with simulated data in Section 3.3.1. Various types of real data examples including RNA-seq data and text classification are considered in Section 3.3.2. We conclude the chapter with a discussion in Section 3.4.

3.2 SPARSE MULTI-CLASS LINEAR DISCRIMINATION

3.2.1 EXISTING METHODS FOR BINARY SPARSE LDA

Suppose that a $p \times n$ data matrix $\mathbf{X}$ consist of p features measured on n observations, each of which belongs to one of two classes. Assume that each column vector $\mathbf{x}_j$ of $\mathbf{X}$ is from multivariate normal distribution $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2$, $j = 1, \dots, n$. The theoretically optimal classifier is to use the discriminant vector $\boldsymbol{\beta}_{\text{Bayes}} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Note that $\boldsymbol{\beta}_{\text{Bayes}}$ is also the solution to the optimization problem that finds the direction vector such that projected data have the largest possible between-class separation while the within-class variance is as small as possible.

The sample version of LDA is found as follows. Let $\mathbf{S}$ be the pooled sample covariance matrix from the two class samples and let $\hat{\boldsymbol{\delta}}$ be the sample mean difference vector that are given by

$$\mathbf{S} = \sum_{k=1}^2 \sum_{j=1}^n (\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_k)^T, \quad \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2,$$

where $\hat{\boldsymbol{\mu}}_k$ is the sample mean vector for Class k , $\hat{\boldsymbol{\mu}}_k = (1/n_k) \sum_{j=1}^{n_k} \mathbf{x}_j$, with n_k being the number of samples in Class k . Then when $\mathbf{S}$ is invertible, the binary LDA solution is

$$\hat{\boldsymbol{\beta}} \propto \mathbf{S}^{-1}\hat{\boldsymbol{\delta}}. \quad (3.2.1)$$

The corresponding classification rule is to classify a new observation $\mathbf{z} \in \mathbb{R}^p$ to Class 1 if and only if

$$(\mathbf{z} - \hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{\beta}} \geq 0, \quad (3.2.2)$$

where $\hat{\boldsymbol{\mu}}_0$ is the mean vector of the whole data.

For HDLSS data with $p \gg n$ where $\mathbf{S}$ is singular, the direct estimation approach proposed by Cai and Liu (2011) targets to find $\boldsymbol{\beta}$ that satisfies the equation $\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\delta}$. In the sample version, a small ridge correction is applied to $\mathbf{S}$ for stable estimation, $\mathbf{S}_\rho = \mathbf{S} + \rho\mathbf{I}$, where $\rho = \sqrt{\log(p)/n}$. An optimization problem is suggested to achieve sparsity of $\boldsymbol{\beta}$ in an analogous

way to the sparse regression in Candes and Tao (2007):

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \|\mathbf{S}_\rho \boldsymbol{\beta} - \hat{\boldsymbol{\delta}}\|_\infty \leq \lambda, \quad (3.2.3)$$

where λ is a tuning parameter that controls the sparsity of $\boldsymbol{\beta}$. Note that the objective and constraint functions are linear, hence the optimization problem (3.2.3) can be solved by linear programming.

Shao et al. (2011) assumed that the common covariance matrix $\boldsymbol{\Sigma}$ is sparse as well as the mean difference vector, $\boldsymbol{\delta}$, and suggested hard-thresholding $\mathbf{S}$ and $\hat{\boldsymbol{\delta}}$ separately. Let $\tau_\gamma(y) = yI(|y| > \gamma)$ be a hard-thresholding function. As for the threshold, $\gamma = M_1 \sqrt{\log p/n}$ and $\gamma = M_2 (\log p/n)^\alpha$ are suggested for off-diagonal elements of $\mathbf{S}$ and entries of $\hat{\boldsymbol{\delta}}$, respectively. Here, M_1 , M_2 and $\alpha \in (0, 1/2)$ are tuning parameters. Let $\tilde{\mathbf{S}}$ and $\tilde{\boldsymbol{\delta}}$ respectively denote the thresholded $\mathbf{S}$ and $\hat{\boldsymbol{\delta}}$. It is worth noting that the resulting discriminant vector $\hat{\boldsymbol{\beta}} \propto \tilde{\mathbf{S}}^{-1} \tilde{\boldsymbol{\delta}}$ is not necessarily sparse. Note also that $\tilde{\mathbf{S}}$ may still be singular, for which case a generalized inverse is used.

In the two aforementioned papers, a multi-class discrimination problem with $K > 2$ classes is also discussed. Both considered all possible pairwise combinations of classes and proposed to classify $\mathbf{z}$ to class k if and only if

$$f_{k\ell}(\mathbf{z}) = (\mathbf{z} - \hat{\boldsymbol{\mu}}_0^{k\ell})^\top \hat{\boldsymbol{\beta}}^{k\ell} \geq 0 \quad \text{for all } \ell \neq k, \quad (3.2.4)$$

where $\hat{\boldsymbol{\mu}}_0^{k\ell}$ is the overall mean of classes k and ℓ and $\hat{\boldsymbol{\beta}}^{k\ell}$ is the solution of the binary problem as above. This pairwise approach potentially has some drawbacks. First, this method suffers from computational burden if the number of classes K is large, since we need to solve $\binom{K}{2}$ binary classification problems. Second, there are likely to be instances where the method cannot assign a class label for an observation, specifically when there is no dominant class label in the comparisons. For example when $K = 3$, one may have $f_{12}(\mathbf{z}) > 0$, $f_{13}(\mathbf{z}) < 0$, and $f_{23}(\mathbf{z}) > 0$, in which case, a class label cannot be assigned to $\mathbf{z}$. Third, even though the

method seems intuitive, it does not produce a $K - 1$ dimensional canonical subspace unlike the original multi-class LDA that will be explained below.

3.2.2 GENERALIZATION TO MULTI-CLASS DISCRIMINATION

Let us first review the original multi-class LDA. Assume that k th Class samples are from $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, \dots, K$ and $K < \min(p, n)$. Let $\mathbf{S}$ be the pooled sample covariance matrix from all K Class samples and let $\mathbf{M}$ be the between-class scatter matrix defined in section (1.2.3). Recall that Fisher's LDA finds a direction vector $\boldsymbol{\beta}$ that maximizes the within-class variation while minimizing the between-class variation, i.e.,

$$\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta}},$$

whose solution is given as an eigenvector from the following generalized eigenvalue problem $\mathbf{M}\boldsymbol{\beta} = \alpha \mathbf{S}\boldsymbol{\beta}$, whose common alternative for high dimensional data is

$$\mathbf{M}\boldsymbol{\beta} = \alpha \mathbf{S}_\rho \boldsymbol{\beta}, \tag{3.2.5}$$

where $\alpha > 0$. Let $\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_{K-1}$ be the generalized eigenvalues, satisfying the orthogonality condition $\tilde{\boldsymbol{\beta}}_i^T \mathbf{S}_\rho \tilde{\boldsymbol{\beta}}_j = 0$. Then the data are projected to the subspace spanned by a few $\tilde{\boldsymbol{\beta}}_i$ s where the classification is usually carried out using (1.2.7). Also projection to the first two $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ provides an optimal two-dimensional visual separation between classes.

We observe that in the original formulation of multi-class LDA, solving the generalized eigenvalue problem to find up to $K - 1$ generalized eigenvalues is equivalent to find the $K - 1$ discriminant vectors from binary LDA. To be more specific, $\hat{\boldsymbol{\delta}}$ in (3.2.1) can be replaced with a basis vector $\mathbf{m}_j$ of $\mathbf{M}$, and the solutions $\mathbf{S}^{-1}\mathbf{u}_1, \dots, \mathbf{S}^{-1}\mathbf{u}_{K-1}$ collectively span the column space of $\mathbf{S}^{-1}\mathbf{M}$.

Theorem 1. *Assume that $\mathbf{S}$ is positive definite. Let $\tilde{\boldsymbol{\beta}}_1, \dots, \tilde{\boldsymbol{\beta}}_{K-1}$ be the generalized eigenvectors that solve (3.2.5), and they consist of basis of $(K - 1)$ -dimensional canonical space,*

denoted by $\mathcal{B}$. Let the vectors $\mathbf{u}_1, \dots, \mathbf{u}_{K-1}$ span the column space of $\mathbf{M}$. Also let $\mathbf{v}_i$ be $\mathbf{S}^{-1}\mathbf{u}_i$, $i = 1, \dots, K-1$, which is a binary LDA solution except for the mean difference vector replaced by $\mathbf{u}_k$. Then $\mathbf{v}_k$ span the same $(K-1)$ -dimensional discriminant subspace as $\tilde{\boldsymbol{\beta}}_k$, which is $\mathcal{B}$.

Proof. Let $\mathcal{C}(\mathbf{A})$ denote the column space of a matrix $\mathbf{A}$, also let $\kappa = K-1$. Note that $\mathcal{C}(\mathbf{S}^{-1}\mathbf{M}) = \mathcal{B}$. Let $\tilde{\mathbf{M}} = [\mathbf{u}_1, \dots, \mathbf{u}_\kappa]$, a horizontal concatenation of basis vectors of $\mathbf{M}$. Since $\mathbf{u}_k$ is in $\mathcal{C}(\mathbf{M})$, we can write $\mathbf{u}_k = \mathbf{M}\mathbf{e}_k$ for some $\mathbf{e}_k$, $k = 1, \dots, \kappa$, i.e., $\tilde{\mathbf{M}} = \mathbf{M}\mathbf{E}$, where $\mathbf{E}$ is a concatenation of $\mathbf{e}_1, \dots, \mathbf{e}_\kappa$. Then we have $\mathcal{C}(\mathbf{S}^{-1}\tilde{\mathbf{M}}) \subset \mathcal{C}(\mathbf{S}^{-1}\mathbf{M})$, since for any $\mathbf{w} \in \mathcal{C}(\mathbf{S}^{-1}\tilde{\mathbf{M}})$, $\mathbf{w} = \mathbf{S}^{-1}\tilde{\mathbf{M}}\mathbf{z} = \mathbf{S}^{-1}\mathbf{M}\mathbf{E}\mathbf{z} = \mathbf{S}^{-1}\mathbf{M}\mathbf{z}^*$ for some $\mathbf{z}$, where $\mathbf{z}^* = \mathbf{E}\mathbf{z}$. Also, since each column of $\mathbf{M}$ is a linear combination of basis vectors $\mathbf{m}_k$, $k = 1, \dots, \kappa$, we can write $\mathbf{M} = \tilde{\mathbf{M}}\mathbf{F}$ for some matrix $\mathbf{F}$. Then we can show that $\mathcal{C}(\mathbf{S}^{-1}\mathbf{M}) \subset \mathcal{C}(\mathbf{S}^{-1}\tilde{\mathbf{M}})$ in a similar way to the earlier argument. □

Motivated by Theorem 1, we propose to use the basis vectors of $\mathbf{M}$ in the place of the mean difference vector $\boldsymbol{\delta}$ in the binary problem, which we will call the *basis* approach. It is clear the basis approach when $k = 2$ is equivalent to the original binary LDA problem since $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 = n_2 n^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$, the between scatter matrix is

$$\begin{aligned} \mathbf{M} &= n_1(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}})^T + n_2(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}})^T \\ &= \frac{n_1 n_2}{n}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T, \end{aligned}$$

whose column space has dimension one and the basis vector is proportional to $\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$.

For general $k > 2$, a natural choice of a basis is using eigenvectors of $\mathbf{M}$, which we use for all our empirical studies. Applying the basis approach to Cai and Liu (2011), we have the following optimization problem for obtaining sparse LDA vector $\boldsymbol{\beta}_k$, $k = 1, \dots, K-1$.

$$\hat{\boldsymbol{\beta}}_k = \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|\mathbf{S}_\rho \boldsymbol{\beta} - \mathbf{u}_k\|_\infty \leq \lambda_k \text{ and } \tilde{\boldsymbol{\beta}}_l^T \mathbf{S}_\rho \boldsymbol{\beta} = 0, \quad l = 1, \dots, k-1, \quad (3.2.6)$$

where $\tilde{\beta}_l$ are the solutions to the generalized eigenvalue problem (3.2.5). Note that we enforce the orthogonality constraint with nonsparse solutions $\tilde{\beta}$ s. Replacing the constraint with $\beta_l^T \mathbf{S}_\rho \beta = 0$ will often yield infeasibility problem, which is suggested as a future investigation.

Note that the proposed approach incorporates the information from all the classes simultaneously. It is also straightforward to visualize the data by projecting onto the discriminant space for interpretation and inspection of the data. As for the choice of the tuning parameters, one can use a common choice $\lambda = \lambda_1 = \dots = \lambda_{K-1}$ which is chosen via cross validation. When the number of classes K is large, one could obtain up to $q \leq K - 1$ discriminant direction vectors where q is regarded as a tuning parameter.

An application of the basis approach to Shao et al. (2011) produces the k th discriminant vector $\hat{\beta}_k = \tilde{\mathbf{S}}^{-1} \tilde{\mathbf{u}}_k$, where $\tilde{\mathbf{u}}_k$ is the thresholded version of $\mathbf{u}_k$. In the next section we demonstrate that the basis method improves the corresponding method for multi-class problems, compared to the pairwise comparison approach.

3.3 EMPIRICAL STUDIES

In this section, we compare the proposed methods with existing sparse multi-class discrimination methods. We evaluate the performances with respect to classification accuracy as well as variable selectivity. We implemented the proposed basis approach to both pairwise comparison methods- linear programming discrimination (LPD) (Cai and Liu, 2011) and thresholded discriminant analysis (TDA) (Shao et al., 2011). Henceforth, we shall refer to the methods by Cai and Liu (2011) and Shao et al. (2011) as LPD-P and TDA-P respectively, and their respective basis counterparts as LPD-B and TDA-B. We compare these methods with the substitution (SUB) method proposed in Chapter 4, sparse linear discriminant analysis (SLDA) (Clemmensen et al., 2011) and penalized linear discriminant analysis (PLDA) (Witten and Tibshirani, 2011).

3.3.1 SIMULATED EXAMPLES

In this section we consider various multi-class classification simulation settings with $K = 3$. For each setting, we consider both balanced ($n_1 = n_2 = n_3 = 30$) and unbalanced ($n_1 = 15, n_2 = 25, n_3 = 50$) training data. The tuning parameters for all the methods are chosen using an independent tuning set of the same size as the training set. We evaluate the methods with independent test data with the size 50 times that of the training set, with 50 repetitions. The datasets for three classes are generated from $p = 500$ dimensional multivariate normal distributions with mean vectors as the following. The first class mean is all zero, $\boldsymbol{\mu}_1 = (0, \dots, 0)^T$, the second class mean has 2 in the first ten entries $\boldsymbol{\mu}_2 = (2, \dots, 2, 0, \dots, 0)^T$, and the third class mean has -2 in the next ten entries $\boldsymbol{\mu}_3 = (0, \dots, 0, -2, \dots, -2, 0, \dots, 0)^T$. The common covariance $\boldsymbol{\Sigma}$ for each setting is as follows:

- **Setting 1 - Auto Regressive (AR(1))** with $\rho = .9$: $\boldsymbol{\Sigma}_{ij} = 0.9^{|i-j|}$, $i, j = 1, \dots, p$.
- **Setting 2 - Inverse AR(1)** with $\rho = .9$: $\boldsymbol{\Sigma}_{ij}^{-1} = 0.9^{|i-j|}$, $i, j = 1, \dots, p$.
- **Setting 3 - AR(1) block**

$$\boldsymbol{\Sigma} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{30 \times 30} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-30} \end{bmatrix},$$

where $\tilde{\boldsymbol{\Sigma}}$ is a block diagonal matrix with three blocks of a 10-dimensional AR(1) matrix with $\rho = .7$. The remaining 470 variables are uncorrelated.

- **Setting 4 - Block Compound Symmetry (CS)**: $\boldsymbol{\Sigma} = \mathbf{I}_5 \otimes \tilde{\boldsymbol{\Sigma}}$,

$$\text{where } \tilde{\boldsymbol{\Sigma}}_{ij} = \begin{cases} 1 & \text{when } i = j, \\ 0.6 & \text{when } i \neq j. \end{cases}$$

The covariance structure for the variables is block diagonal compound symmetric with 5 blocks of size 100, within-block correlation 0.6, and between-block correlation 0.

The covariance matrix in Setting 1 is approximately sparse with each column having at most 130 nonzero entries but the precision matrix Σ^{-1} is highly sparse. Each column has at most 3 nonsparse features and the rest are sparse. This highly sparse precision matrix is relevant in cases where one can either assume that variables i and j that are not close are uncorrelated or the underlying precision matrix is truly sparse. The first true discriminant direction β_1 is 11-sparse, and the second true discriminant direction β_2 is 12-sparse. This means that there are respectively 11 and 12 signals in the first and second true discriminant directions, with the union giving us the total number of signal variables. In this setting, there are 21 signal variables with 479 noise variables. In Setting 2, the covariance matrix is more sparse in comparison to Setting 1. However, the precision matrix is less sparse in this setting. The true discriminant vector, β_1 , has 82 nonzero entries loaded on the first 82 variables. The first 92 variables in β_2 are signal variables. This setting has 92 signal variables and 408 noise variables in totality. The covariance structure for Settings III and IV are set to mimic gene expressions data where within each block the variables are correlated with positive correlation and where the variables are uncorrelated between blocks. In setting 3, both the covariance and the precision matrices are sparse but the latter is more sparse with at most 3 nonzero entries in each row or column. This setting is more sparse than Setting 4 where the covariance and precision matrices are 100-sparse, with 100 nonzero entries in each row and columns. Here, the true discriminant directions β_1 and β_2 are 10-sparse each with a combined signal variables of 20. In Setting 4, the true discriminant vectors are 100-sparse each, with the nonzero loadings found on the first 100 variables.

The average classification errors evaluated with test samples based on 50 replications are shown in Figure 3.3.1. In order to facilitate visual comparison, we used the same line type for the pair of methods based on the same approach. For example, solid lines with different width are used for LPD-P and LPD-B. To show the variability, each average is displayed with error bars of twice standard errors. In the balanced sample cases shown in the left panel, the

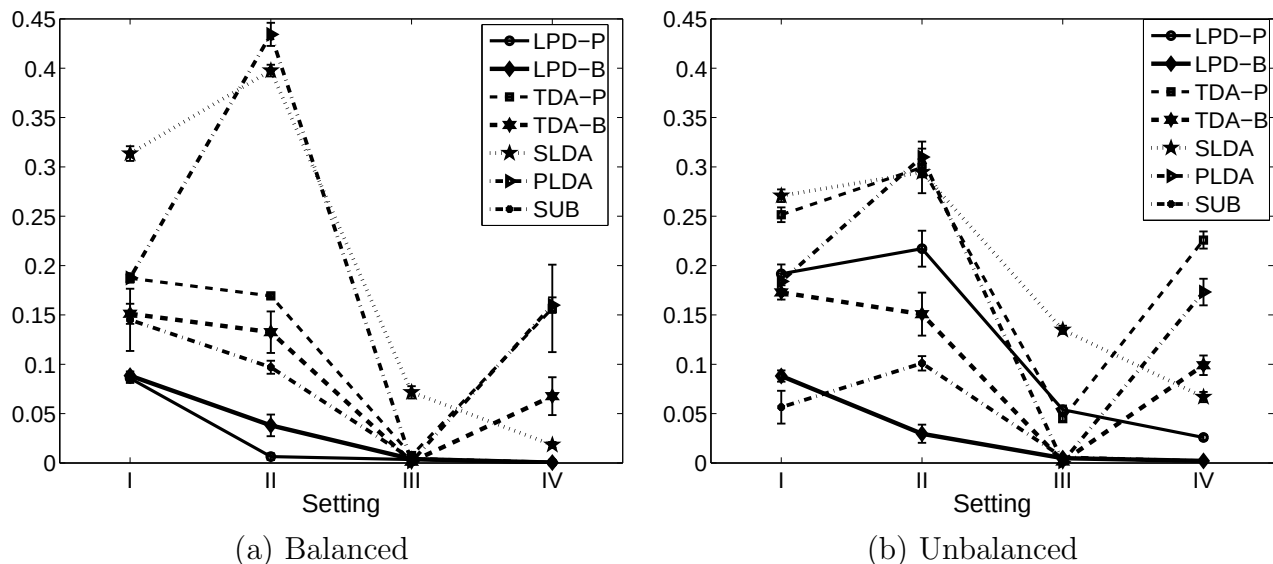
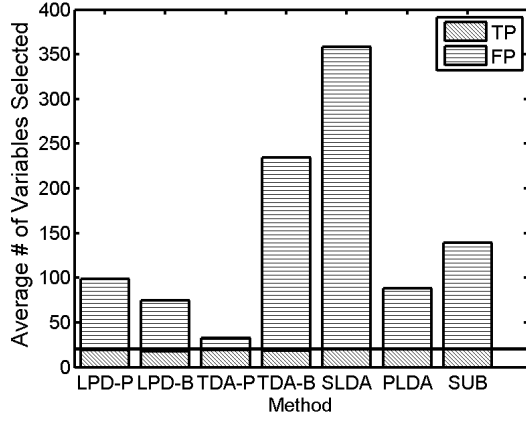


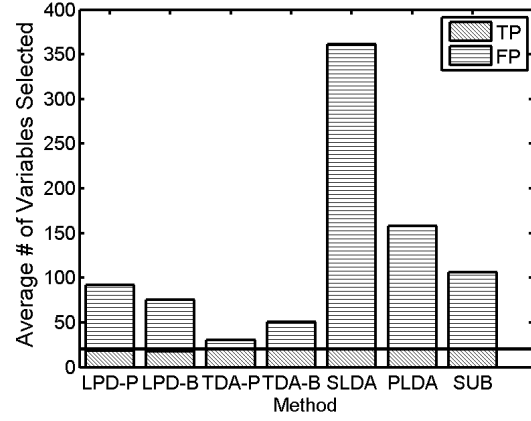
Figure 3.3.1: Average test errors shown with their standard errors based on 50 repetitions. The LPD-B and TDA-B, the basis-based multi-class extensions of LPD and TDA, improve the respective pairwise approaches, LPD-P and TDA-B.

two LPD methods, LPD-P and LPD-B, show the strongest performance, especially LPD-P in Setting II. However, the TDA-B improves TDA-P significantly in Settings I, II, and IV, even though their performances are weaker than the two LPD methods. As expected unequal sample size settings hurt the performances by pairwise comparison approaches such as LPD-P and TDA-P, as seen in the right panel of Figure 3.3.1. The improvement by the proposed basis approach is dramatic for both LPD and TDA methods, while the best performance is made by LPD-B, across Settings I-III.

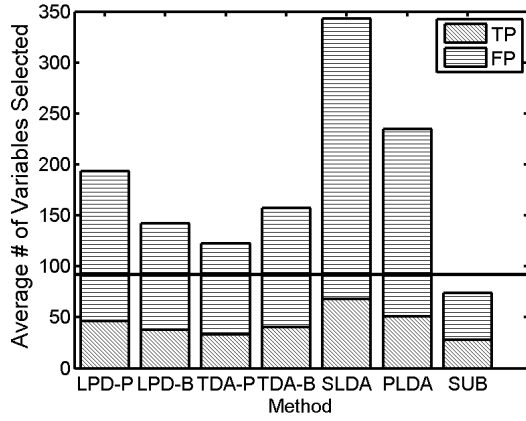
When comparing sparse methods, it is helpful to see the selection performance as well as classification performance. In the bar graphs in Figures 3.3.2 and 3.3.3, average number of selected variables by each method is shown. The height of each bar represents the average number of selected variables, which can be decomposed as the sum of the number of true signal variables selected, also referred to as True Positives (TP), and the number of noise



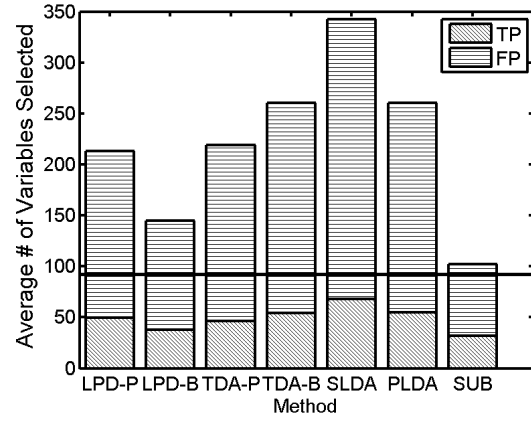
(a) Setting I - Balanced



(b) Setting I - Unbalanced

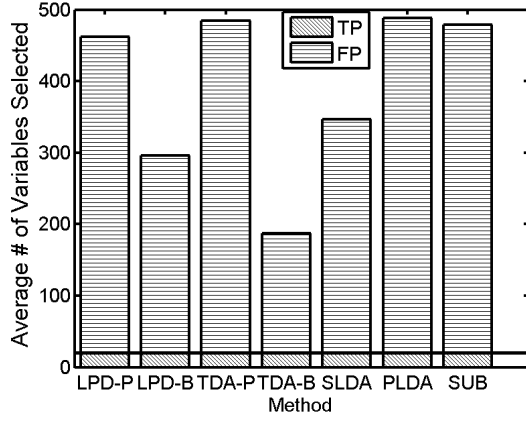


(c) Setting II - Balanced

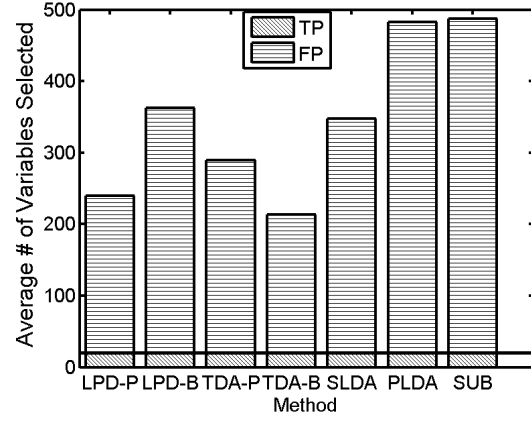


(d) Setting II - Unbalanced

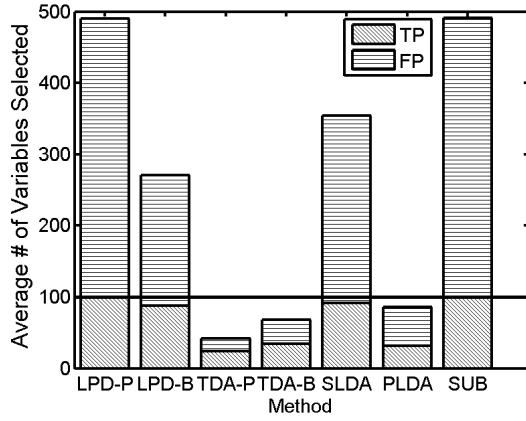
Figure 3.3.2: Variable selection properties of each method for each setting. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables, which is referred to as True Positives (TP) in the bottom and selected noise variables, also defined as False Positives (FP) in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.



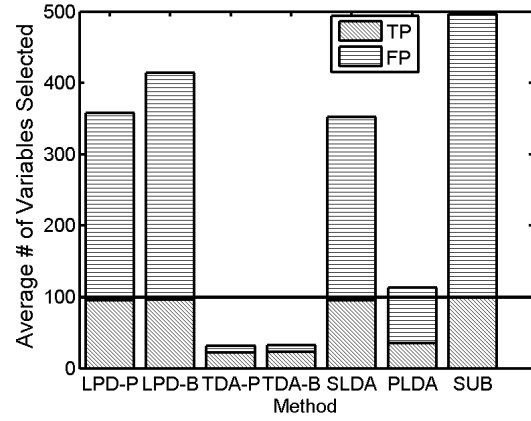
(a) Setting III - Balanced



(b) Setting III - Unbalanced



(c) Setting IV - Balanced



(d) Setting IV - Unbalanced

Figure 3.3.3: Variable selection properties of each method for each setting. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables, also known as True Positives (TP) in the bottom and selected noise variables, also defined as False Positives (FP) in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.

variables selected and also referred to as False Positives (FP), shown in bottom and top of the bar, respectively. A horizontal across each panel represent the number of the true signal variables. For LPD-B, TDA-B, SUB, SLDA and PLDA methods, a variable is selected if it has at least one nonzero loadings in the combined discriminant vectors. For instance, if the first variable in $\hat{\beta}_1$ has zero coefficient but nonzero coefficient in $\hat{\beta}_2$ for any of these methods, then the first variable is considered selected. Similarly, for the pairwise approaches LPD-P and TDA-P, the union of variables with nonzero coefficients in all pairwise combinations results in selectivity.

Comparing the test errors in Figure 3.3.1 and the corresponding selection performance in Figures 3.3.2 and 3.3.3 can reveal interesting relationships between classification performance and variable selectivity. For example, in balanced Setting I, SLDA selected most FP, which is reflected in the test errors. In unbalanced Setting II, even though SLDA has most TP, it also has most FP along with PLDA, which makes their classification performance inferior, where LPD-B has the least FP, which possibly explains the lowest test error. The Setting III is the most challenging setting in terms of the variable selection, however all methods show relatively low test errors. In this setting, it seems that a good selectivity is not required to ensure a good performance, which implies that the accuracy in estimation in l_2 sense makes a difference between the methods. In Setting IV, we note that the methods with higher TP generally yield a lower test errors.

3.3.2 REAL DATA EXAMPLES

In addition to the simulation settings, we also apply the basis method to the analysis of three microarray datasets-lymphoma cancer data, SRBCT data and Brain cancer data (Dettling, 2004) and one RNA-seq dataset (Graveley et al., 2011a) to further assess the performance of our method.

3.3.2.1 MICROARRAY DATASETS

We analyzed the microarray datasets from Dettling (2004) using the proposed methods to identify linear combinations of genes that results in minimum classification error. Tables 3.3.1 and 3.3.2 display the dimensions, sample sizes, and the number of classes for each microarray dataset.

In the analysis, we randomly split each data set in three parts where two-thirds are used as training data and one-third as testing data. A stratified sampling approach is applied to divide the data in order to preserve the original proportions of samples in each class. To reduce computational cost, we first performed ANOVA test on the training data to compare the means for the classes for each gene and selected the 1000 most significant genes (i.e., with the smallest p-values). The corresponding features in the testing data were then used in the analysis to examine the misclassification rates. This avoids using the testing data twice. The optimal tuning parameter was chosen via 5-fold cross validation on the training data and then applied to the testing set. We repeat the foregoing estimation scheme 10 times.

Table 3.3.3 shows the classification performance of the basis methods using the testing data. In general, when compared to the pairwise and multi-class approaches, the basis methods show a competitive performance in terms of classification accuracy. TDA-B does better than its pairwise counterpart, TDA-P, on all three microarray datasets. LPD-B however, shows competitive performance only on the lymphoma dataset when compared to LPD-P. Table 3.3.4 gives the number of variables selected by the methods. One can notice that in general, the number of variables selected by the methods in all the datasets are quite high. The basis methods, especially LPD-B, select less variables mostly than the other methods. It may be observed that for most of the methods, the number of variables selected is positively related to the number of classes - for large number of classes, more variables are selected. Figure 3.3.4 are graphs of the projection of the lymphoma testing data onto an

informative two-dimensional subspace produced by the multi-class methods with their class boundaries using one random sample of the data.

Dataset	n	p	K	Responses
Lymphoma	62	4026	3	subtypes
SRBCT	63	2308	4	tumor types
Brain	42	5597	5	tumor types

Table 3.3.1: Summary statistics of microarray Data

Samples per class	Lymphoma	SRBCT	Brain	Fly
n_1	42	23	10	60
n_2	9	20	10	28
n_3	11	12	10	29
n_4	N/A	8	4	30
n_5	N/A	NA	8	N/A

Table 3.3.2: Number of samples per class for microarray and RNA-seq datasets.

3.3.2.2 RNA-SEQ DATASET

Recently, continuing improvement in technology and decreasing cost of next-generation sequencing have made RNA sequencing (RNA-seq) a widely used method for gene expression studies (Dillies et al., 2013). RNA-seq data after being processed may be in the form of counts (e.g from the Myrna algorithm), but is more often in the form of continuous values when normalized using the suggested techniques in Dillies et al. (2013). Therefore, linear models with continuous high dimensional predictors are reasonable to use for RNA-seq data. However, it is important to check that the processed data is reasonably Gaussian and, if not, to transform the data.

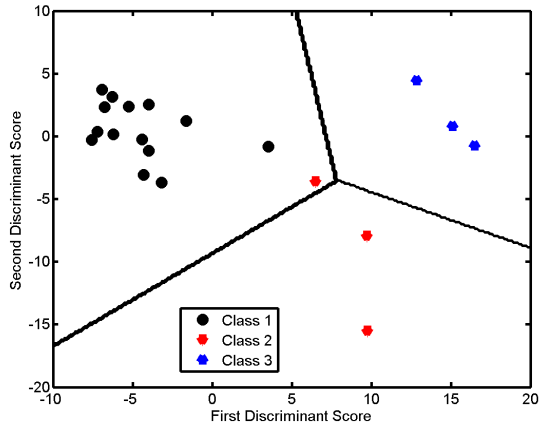
We applied our proposed method to the *Drosophila melanogaster* (Fly) data of Graveley et al. (2011b). The processed data were downloaded from ReCount database (Frazer et al., 2011). Features with more than half their values being zero were filtered out. The remaining features with zero values were truncated at 0.5 and the data were log-transformed. We filtered

	Lymphoma	SRBCT	Brain	Fly
LPD-P	10.00 (1.489)	2.63 (1.176)	6.67 (2.078)	0.00 (0.000)
LPD-B	1.00 (0.667)	4.74 (2.656)	9.17 (2.308)	2.92 (1.527)
TDA-P	11.50 (1.303)	9.47 (1.312)	15.00 (2.992)	7.29 (0.835)
TDA-B	0.00 (0.000)	2.63 (1.177)	9.17 (1.945)	0.63 (0.626)
SLDA	0.00 (0.000)	2.11 (0.860)	9.17 (2.308)	0.00 (0.000)
PLDA	2.50 (1.344)	1.58 (0.803)	10.83 (1.777)	0.00 (0.000)
SUB	0.50 (0.500)	1.58 (0.803)	14.97 (3.45)	0.00 (0.000)

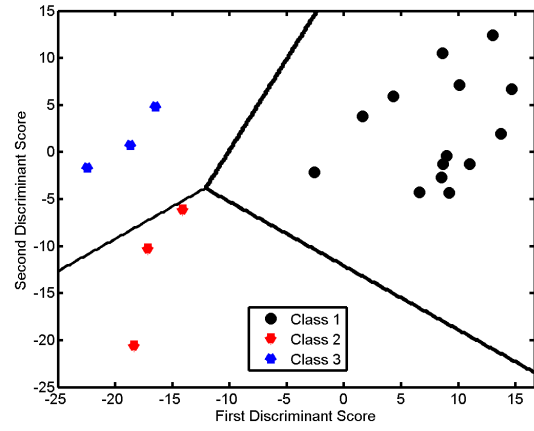
Table 3.3.3: Comparison of the misclassification rates (standard errors) of basis methods to other linear discriminant analysis methods. Error rates and standard errors are in percentage. It is noticeable that the basis methods LPD-B and TDA-B mostly improve on their respective counterparts LPD-P and TDA-P.

	Lymphoma	SRBCT	Brain	Fly
LPD-P	508.40	712.70	997.70	1000.00
LPD-B	514.80	257.90	836.60	257.90
TDA-P	1000.00	1000.00	1000.00	949.90
TDA-B	819.10	653.20	883.60	537.00
SLDA	642.20	814.40	914.30	734.50
PLDA	990.40	801.30	1000.00	992.70
SUB	780.20	261.90	519.90	394.70

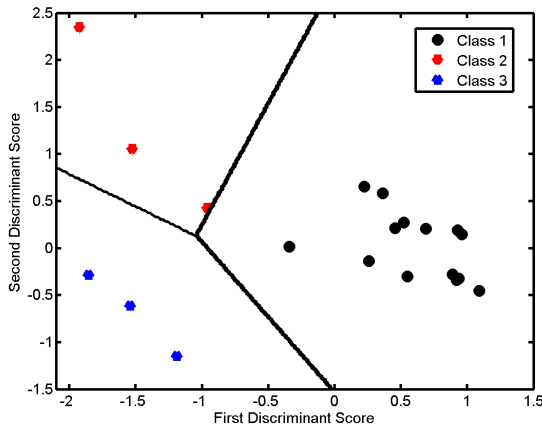
Table 3.3.4: Comparison of number of variables selected by basis methods to other linear discriminant analysis methods. The basis methods LDP-B and TDA-B are more sparse; LPD-B selects fewer variables.



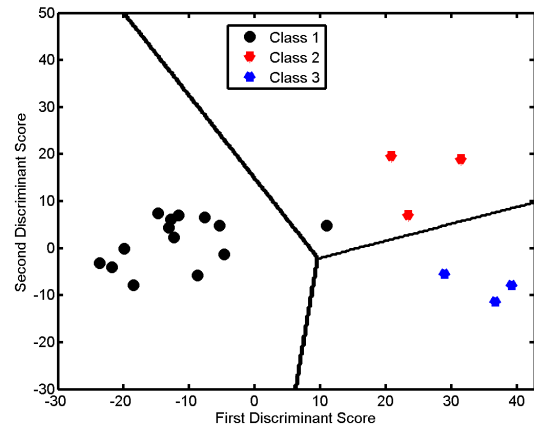
(a) LPD-B



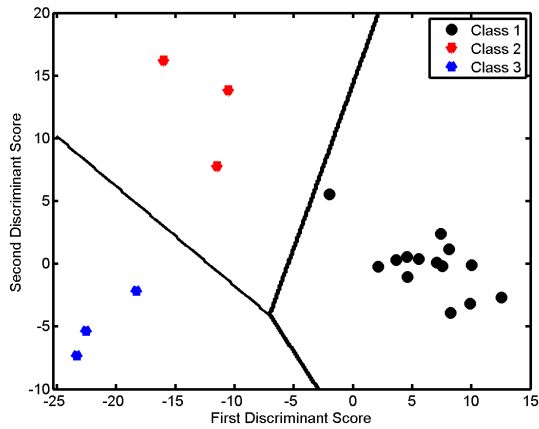
(b) TDA-B



(c) SLDA



(d) PLDA



(e) SUB

Figure 3.3.4: A two dimensional plot of the lymphoma dataset. This is the projection of one random testing data onto an informative two-dimensional subspace obtained by the multi-class methods, with their class boundaries.

out features with low variances, resulting in $p = 1,000$ dimensions. Finally, the data were normalized to have equal medians for each sample, and mean zero and unit variance for each feature. There were four fly classes: Class 1 consisted of all embryos; Class 2 consisted of all larvae; Class 3 consisted of all white prepupae; Class 4 consisted of all adult flies. The dataset consisted of a total of $n = 147$ samples. The analysis was carried out similarly to the microarray analysis.

Table 3.3.3 shows the classification performance of the basis methods using the testing data. LPD-B's performance when compared to LPD-P is suboptimal. On the other hand, the classification performance of TDA-B when compared to TDA-P is superior. LPD-P, SLDA, PLDA and SUB achieve zero error rate while TDA-P does worse in this relatively easy classification problem. The perfect classification accuracy by these three methods may be because of huge differences in fly stages in gene expression. Observe from Table 3.3.4 that the basis method LPD-B and the substitution method SUB are more sparse.

3.4 DISCUSSION

In this chapter, we have proposed a simple and effective multi-class framework of obtaining sparse discriminant vectors. The methodology can be applied to any LDA-based approach that solves Fisher's LDA problem. Our basis method is based on the observation that the binary LDA solution using the orthonormal basis vectors of the between class variability, $\mathbf{M}$, collectively span the same canonical discriminant space as the binary LDA problem that solves Fisher's original LDA problem. The method was shown to perform better (especially under unequal class prevalences) than both the pairwise approach of solving multi-class sparse LDA problems and some existing multi-class approaches, on both simulated datasets and real data applications. The method was applied to microarray and RNA-seq data. Our simulations revealed that the basis method works very well, especially in the case where one

can assume that the true direction vectors are highly sparse and depend on only few relevant variables.

In this work, our simulations focused on cases where the data are drawn from the normal distribution. Since Fisher’s LDA uses only the between-class variance and within-class variance without any distributional assumption on the feature vector, we expect our method to work well in nonnormal cases. Our work focused on obtaining all $K - 1$ sparse LDA vectors in a K class problem. It would be interesting to determine the performance of the basis method for $q < (K - 1)$ direction vectors, especially for very large number of classes. One way of selecting q is to choose the q direction vectors that results in smaller or similar cross validation errors, when choosing the tuning parameter.

3.5 REFERENCES

- Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.
- Bickel, P. J. and Levina, E. (2004). Some theory for fisher’s linear discriminant function, naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.
- Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.
- Clemmensen, L., Hastie, T., Witten, D., and Ersbll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.
- Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593.
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Lalo, D., Gall, C. L., Schaeffer, B., Crom, S. L., Guedj, M., and Jaffrzic, F. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

- Frazee, A., Langmead, B., and Leek, J. (2011). Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 12(449).
- Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011a). The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479.
- Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011b). The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479.
- Hsu, C. W. and Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.
- Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81.

- Lu, J., Plataniotis, K. N., and Venetsanopoulos, A. N. (2003). Face recognition using LDA-based algorithms. *IEEE Transactions on Neural Networks*, 14(1):195–200.
- Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, pages 29–42.
- Qiao, Z., Zhou, L., and Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data. *IAENG International Journal of Applied Mathematics*, 39(1):48–60.
- Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics.*, 39:1241–1265.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.
- Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

CHAPTER 4

SPARSE ANALYSIS FOR HIGH DIMENSIONAL DATA¹

¹Sandra Safo, Jeongyoun Ahn (2014+) Sparse Analysis for High Dimensional Data. To be submitted.

ABSTRACT

A core idea of most multivariate data analysis methods is to project higher dimensional data vectors on to a lower dimensional subspace spanned by a few meaningful directions. Many multivariate methods, such as canonical correlation analysis (CCA), multivariate analysis of variance (MANOVA), and linear discriminant analysis (LDA), solve a generalized eigenvalue problem. We propose a general framework, called substitution method, with which one can easily obtain a sparse estimate for a solution vector of a generalized eigenvalue problem. We employ the idea of direct estimation in high dimensional data analysis and suggests a flexible framework for sparse estimation in all statistical methods that use generalized eigenvectors to find interesting low-dimensional projections in high dimensional space. We illustrate the framework with sparse CCA and LDA to demonstrate its effectiveness.

KEYWORDS: High Dimension, Low Sample Size; Linear Discriminant Analysis; Generalized Eigenvalue; MANOVA; Canonical Correlation Analysis; Sparsity

4.1 INTRODUCTION

A key idea in most traditional multivariate methods is finding a lower dimensional subspace spanned by a few linear combinations of all available variables for dimensionality reduction and to aid in exploratory data analysis. These linear combinations are often times eigenvectors of $\mathbf{S}^{-1}\mathbf{M}$ from the GEV problem (1.2.16), assuming that $\mathbf{S}$ is invertible. In HDLSS where $\mathbf{S}$ is singular, the GEV problem becomes

$$\mathbf{M}\mathbf{v} = \alpha\mathbf{S}_\eta\mathbf{v}, \tag{4.1.1}$$

with $\mathbf{S}_\eta$ being a positive definite matrix of $\mathbf{S}$, and is usually obtained by adding a small multiple η of the identity matrix to $\mathbf{S}$. This is usually chosen to be $(\log(p)/n)^{1/2}$ in HDLSS studies (Bickel and Levina, 2008; Cai and Liu, 2011). Here, p is the number of variables

and n is the number of observations. From section (1.2.3), the solution to the GEV problem (4.1.1) are the eigenvalue-eigenvector pair of $\mathbf{S}_\eta^{-1}\mathbf{M}$. These eigenvectors, $\tilde{\mathbf{v}}$, use all available variables making it difficult to interpret results in HDLSS problems. In the next section, we discuss our proposition to making $\tilde{\mathbf{v}}$ sparse.

4.2 THE SUBSTITUTION METHOD

Candes and Tao (2007) proposed the Dantzig selector (DS) for sparse estimation of regression coefficients in multiple regression analysis. Specifically, they used l_1 bound on the regression coefficients and imposed l_∞ constraint on the size of the residual vector. They theoretically showed that the DS selector satisfies the oracle property of variable selection consistency and can be used as an effective variable selection method. An advantage of the DS over other sparse regularization methods is that it solves a simple convex optimization problem which can easily be recast and solved conveniently by linear programming. Cai and Liu (2011) also used the DS for sparse representations of linear discriminant vectors in a binary classification problem. Following the success and easy implementation of the DS, we make the solution vector $\mathbf{v}$ of the GEV problem sparse by imposing l_∞ constraint and minimizing l_1 objective function. A direct application of the DS estimator (1.2.10) to the GEV problem (4.1.1) will be to naively solve the optimization problem

$$\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to} \quad \|(\mathbf{M} - \alpha\mathbf{S}_\eta)\mathbf{v}\|_\infty \leq \tau, \quad (4.2.1)$$

where $\tau > 0$ is a tuning parameter that controls how many coefficients in $\mathbf{v}$ are set to zero. This a naive approach because from (4.2.1), $\hat{\mathbf{v}} = \mathbf{0}$ always satisfies the constraint regardless of the tuning parameter, i.e., the zero vector is always the solution to (4.2.1). We have the following lemma.

Lemma 1. *Consider the optimization problem*

$$\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to} \quad \|(\mathbf{M} - \alpha\mathbf{S}_\eta)\mathbf{v}\|_\infty \leq \tau. \quad (4.2.2)$$

The vector $\mathbf{v} = \mathbf{0}$ is always a solution regardless of the value of τ .

Proof. First, we rewrite the criterion (4.2.2) using Lagrange multipliers:

$$\begin{aligned} & \|\mathbf{v}\|_1 + \tau \|(\mathbf{M} - \alpha \mathbf{S}_\eta) \mathbf{v}\|_\infty \\ & \leq \|\mathbf{v}\|_1 + \tau \|(\mathbf{M} - \alpha \mathbf{S}_\eta)\|_\infty \|\mathbf{v}\|_1, \end{aligned} \quad (4.2.3)$$

where we use the inequality $\|\mathbf{A}\mathbf{x}\|_\infty \leq \|\mathbf{A}\|_\infty \|\mathbf{x}\|_1$ for matrix $\mathbf{A}$ and vector $\mathbf{x}$, and where $\|\mathbf{A}\|_\infty$ is the elementwise l_∞ norm of an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ with entries a_{ij} and is defined as $\max_{1 \leq i, j \leq p} |a_{ij}|$. Now, we differentiate with respect to $\mathbf{v}$, set the derivative to 0, and solve for $\mathbf{v}$:

$$\Gamma + \tau \max_{1 \leq i, j \leq p} |(\mathbf{M} - \alpha \mathbf{S}_\eta)_{ij}| \Gamma = 0 \quad (4.2.4)$$

where $\Gamma = \{\mathbf{v} = (v_1, \dots, v_p)^T : v_i = \text{sign}(v_i) \text{ if } v_i \neq 0 \text{ and } v_i \in [-1, 1] \text{ if } v_i = 0, i = 1, \dots, p\}$.

Thus, there is some $\mathbf{v} \in \Gamma$ such that

$$\mathbf{v} + \tau \max_{1 \leq i, j \leq p} |(\mathbf{M} - \alpha \mathbf{S}_\eta)_{ij}| \mathbf{v} = \mathbf{0} \quad (4.2.5)$$

The Karush-Kuhn-Tucker conditions for optimality consist of (4.2.4) and $\tau \|(\mathbf{M} - \alpha \mathbf{S}_\eta) \mathbf{v}\|_\infty = 0$. Now for any $\tau > 0$, we have from (4.2.5) that

$$\mathbf{v}(1 + \tau \max_{1 \leq i, j \leq p} |(\mathbf{M} - \alpha \mathbf{S}_\eta)_{ij}|) = \mathbf{0}, \quad (4.2.6)$$

which implies that either $\mathbf{v} = \mathbf{0}$ or $1 + \tau \max_{1 \leq i, j \leq p} |(\mathbf{M} - \alpha \mathbf{S}_\eta)_{ij}| = 0$ or both are true. Now, if $\mathbf{v} \neq \mathbf{0}$, then $1 + \tau \max_{1 \leq i, j \leq p} |(\mathbf{M} - \alpha \mathbf{S}_\eta)_{ij}| = 0$ must hold, which is clearly not true since τ is a positive constant. Hence $\mathbf{v} = \mathbf{0}$. This completes the proof, showing that regardless of the value of τ , the solution to the optimization problem is always the zero vector.

□

We are interested in a solution vector, $\mathbf{v}$, that has at least one nonzero coefficient. Hence, we substitute the left term in $\|\cdot\|_\infty$ of equation (4.2.1) with the nonsparse eigenvector, $\tilde{\mathbf{v}}$ of

$\mathbf{S}_\eta^{-1}\mathbf{M}$, that corresponds to the largest eigenvalue, $\tilde{\alpha}$, and call this approach the *substitution method*. In other words, to obtain sparse solution vector $\mathbf{v}$, we solve the revised optimization problem

$$\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to} \quad \|\mathbf{M}\tilde{\mathbf{v}} - \tilde{\alpha}\mathbf{S}_\eta\mathbf{v}\|_\infty \leq \tau, \quad (4.2.7)$$

where $\tau > 0$. The tuning parameter $\tau > 0$ may be selected from a grid of finite values via cross validation. If $\tau = 0$, the nonsparse solution vector $\tilde{\mathbf{v}}$ is recovered. We note that we substitute the left term of $\|\cdot\|_\infty$ with $\tilde{\mathbf{v}}$ in (4.2.7) instead of the right term as in

$$\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to} \quad \|\mathbf{M}\mathbf{v} - \tilde{\alpha}\mathbf{S}_\eta\tilde{\mathbf{v}}\|_\infty \leq \tau, \quad (4.2.8)$$

for mathematical and computational reasons. Mathematically, from convex optimization theory, if $\mathbf{M}$ is singular, the solution obtained using (4.2.8) is not stable. In some problems such as LDA, $\mathbf{M}$ is indeed singular, and we obtain stable estimates when we use (4.2.7) because of the LDA optimization problem. In other problems where $\mathbf{M}$ is nonsingular, we achieve great computational savings in our simulations when we use the optimization problem in (4.2.7).

In order to solve the optimization problem (4.2.7) using linear programming, the objective and constraint functions must be linear, i.e., satisfy $f(\theta\mathbf{v} + (1 - \theta)\mathbf{y}) = \theta f(\mathbf{v}) + (1 - \theta)f(\mathbf{y})$ for $0 \leq \theta \leq 1$ and $\mathbf{v}, \mathbf{y} \in \mathbb{R}^p$. The objective function $\|\mathbf{v}\|_1$ is convex since for $0 \leq \theta \leq 1$ and $\mathbf{v}, \mathbf{y} \in \mathbb{R}^p$,

$$\begin{aligned} f(\theta\mathbf{v} + (1 - \theta)\mathbf{y}) = \|\theta\mathbf{v} + (1 - \theta)\mathbf{y}\|_1 &= \sum_{i=1}^p |\theta v_i + (1 - \theta)y_i| \\ &\leq \sum_{i=1}^p |\theta v_i| + \sum_{i=1}^p |(1 - \theta)y_i| \\ &= \theta f(\mathbf{v}) + (1 - \theta)f(\mathbf{y}). \end{aligned} \quad (4.2.9)$$

Following a similar reasoning, the constraint function can be shown to be convex. Observe that convexity is more general than linearity and that any linear program is a convex optimization problem. Thus, the optimization problem (4.2.7) may be solved via linear programming:

$$\begin{aligned}
& \min \sum_{i=1}^p r_i \quad \text{subject to} \\
& -v_i \leq r_i \quad \text{for all } 1 \leq i \leq p \\
& +v_i \leq r_i \quad \text{for all } 1 \leq i \leq p \\
& -\tilde{m}_i - \tilde{\alpha} \hat{\boldsymbol{\sigma}}^T \leq \tau \quad \text{for all } 1 \leq i \leq p \\
& +\tilde{m}_i - \tilde{\alpha} \hat{\boldsymbol{\sigma}}^T \leq \tau \quad \text{for all } 1 \leq i \leq p
\end{aligned}$$

where $\mathbf{r} = (r_1, \dots, r_p)^T$ and $\mathbf{v}$ are the optimization variables, $\tilde{m}_i$ are the elements of $\mathbf{M}\tilde{\mathbf{v}}$ and where $\hat{\boldsymbol{\sigma}}_i$ are the columns of $\mathbf{S}_\eta$.

In some problems, one may be interested in obtaining $K > 2$ sparse solution vectors. For instance, in linear discriminant analysis with more than two classes, it may be true that the first discriminant vector is not enough in discriminating well between the classes. Subsequent sparse solution vectors are obtained by assuming that they are uncorrelated with previously estimated directions. Hence, we impose orthogonality constraints on the previous and subsequent sparse solution vectors, and solve the optimization problem for $\mathbf{v}_j, j = 2, \dots, K$:

$$\min_{\mathbf{v}} \|\mathbf{v}\|_1 \quad \text{subject to} \quad \|\mathbf{M}\tilde{\mathbf{v}}_j - \tilde{\alpha}_j \mathbf{S}_\eta \mathbf{v}\|_\infty \leq \tau_j \quad \text{and} \quad \mathbf{B}_{j-1} \mathbf{S}_\eta \mathbf{v} = 0 \quad (4.2.10)$$

where K is the rank of $\mathbf{M}$ and $\mathbf{B}_{j-1} = [\mathbf{v}_1, \dots, \mathbf{v}_{j-1}]$ are the previous sparse solution vectors. The tuning parameters $\tau_j > 0$ can be chosen via cross validation. One can use a common tuning parameter $\tau = \dots = \tau_j$. Under these constraints, some coefficients in $\mathbf{v}_j$ from (4.2.10) will be exactly zero. When τ_j gets larger, the coefficients are more sparse. We note that the extreme case of sparsity, where all coefficients are zero, is achieved when $\tau_j = \|\mathbf{M}\tilde{\mathbf{v}}_j\|_\infty$, giving us an upper bound, say $\tau_{\max}$, on τ . The tuning parameter τ_j is chosen to be less

than $\tau_{\max}$, to ensure at least one nonzero coefficient. What follows are demonstrations of the substitution method on LDA and CCA.

4.3 SUBSTITUTION FOR SPARSE LINEAR DISCRIMINANT ANALYSIS

The substitution method can be directly applied to LDA by observing that $\mathbf{M}$ and $\mathbf{S}$ in (4.2.7) are the between class scatter and pooled covariance matrices in Fisher's LDA problem (1.2.4). In Chapter 1, Section 2.3, we saw that the problem of finding direction vectors that results in maximal separation between classes and minimal variation within classes reduces to the GEV (4.1.1) with solutions being eigenvalue-eigenvector pair of $\mathbf{S}_\eta^{-1}\mathbf{M}$, where $\mathbf{S}_\eta$ is the HDLSS nonsingular pooled covariance matrix. Hence, for first sparse linear discriminant vectors $\hat{\boldsymbol{\beta}}$, the optimization problem in (4.2.7) must be solved. Subsequent discriminant vectors $\hat{\boldsymbol{\beta}}_k, k = 1, \dots, K - 1$ can be obtained from the optimization problem in (4.2.10). Once the sparse discriminant vectors have been obtained, one can classify by assigning a new entity, $\mathbf{z}$ to the closest population using nearest centroid rule (1.2.7).

In Chapter 3, Section 3.3, the performance of the substitution method is compared to the basis methods and other existing sparse LDA methods in simulated processes and real data analyses. From the simulations and under equal and unequal class prevalences, we observed that the classification accuracy of the substitution method was competitive and comparable to the basis method LPD-B especially under Settings III and IV. The performance of the substitution method in these settings suggest that our method will not only perform well when the underlying precision matrix is sparse but also in situations where this matrix is less sparse. In the real data analyses, we also observed a competitive performance of the substitution method in terms of classification accuracy. We noticed that this method is, in general, more sparse than other existing LDA methods.

4.4 SUBSTITUTION FOR SPARSE CANONICAL CORRELATION ANALYSIS

4.4.1 INTRODUCTION

The goal of sparse canonical correlation analysis (sparse CCA) is to find linear combinations of two sets of variables using a fraction of the variables so that these linear combinations have maximum correlation. Traditionally, CCA methods use all available variables which makes interpretations of HDLSS problems daunting.

Recently, sparse CCA methods have gained popularity in the literature (Waaijenborg et al., 2008; Parkhomenko et al., 2009; Witten et al., 2009; Chalise and Fridley, 2012). Most of these works achieve sparsity via l_1 regularization or a variant of it as discussed in Chapter 1, Section 2.4. Chalise and Fridley (2012) used the CCA algorithm of Parkhomenko et al. (2009) and compared several sparsity penalty functions such as lasso (Tibshirani, 1994), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001) and hard-thresholding. They concluded that elastic net and SCAD, in particular SCAD, achieves maximum correlation between the canonical covariates and is more sparse.

We propose a sparse CCA method that is based on the substitution method described in Section 4.2. Our method differs from Waaijenborg et al. (2008) in that we consider optimizing the CCA problem directly instead of using regression approach. Similar to Parkhomenko et al. (2009) and Witten et al. (2009), we decompose the covariance matrix between the sets of variables using SVD. However, instead of soft-thresholding and applying penalty functions directly to the left and right singular vectors, we only initialize our algorithm with these vectors but solve the eigenvalue problem arising from the CCA optimization problem. Our sparse CCA method also differs from the above methods in the selection of the sparseness parameter that controls the number of variables in the canonical covariates. Unlike the method proposed by Parkhomenko et al. (2009) which only implements the first canonical

covariate, and therefore becomes problematic where additional pairs need to be obtained, our substitution method can easily be used for subsequent canonical correlation pairs.

4.4.2 CANONICAL CORRELATION ANALYSIS

Suppose that we have two data matrices, a $n \times p$ matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$, and a $n \times q$ matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_q]$. Without loss of generality, assume the variables have been centered. The goal of CCA (Hotelling, 1936) is to find linear combinations of all the variables in $\mathbf{X}$, say $\mathbf{X}\boldsymbol{\alpha}$ and the linear combinations of all the variables in $\mathbf{Y}$, say $\mathbf{Y}\boldsymbol{\beta}$ such that the correlation between these linear combinations is maximized. Let $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ be the population covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ respectively, and let $\boldsymbol{\Sigma}_{xy}$ be the $p \times q$ covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. Let $\rho = \text{corr}(\mathbf{X}\boldsymbol{\alpha}, \mathbf{Y}\boldsymbol{\beta})$ be the correlation between the canonical covariates. Mathematically, the goal of CCA is to find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that solves

$$\rho = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \text{corr}(\mathbf{X}\boldsymbol{\alpha}, \mathbf{Y}\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha}} \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta}}}. \quad (4.4.1)$$

The correlation coefficient in (4.4.1) is not affected by scaling of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, hence one can choose the denominator to be equal to one and solve the equivalent problem: find $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that solves the optimization problem

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} = 1 \quad \text{and} \quad \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} = 1. \quad (4.4.2)$$

Subsequent directions are obtained by imposing additional orthogonality constraints

$$\boldsymbol{\alpha}_i^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha}_{i'} = \boldsymbol{\beta}_i^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta}_{i'} = \boldsymbol{\alpha}_i^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta}_{i'} = 0, \quad i \neq i', \quad i, i' = 1, \dots, \min(p, q).$$

Using Lagrangian multipliers λ and μ , for the first canonical coefficients we have

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda, \mu) = \boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\beta} - (\lambda/2)(\boldsymbol{\alpha}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha} - 1) - (\mu/2)(\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta} - 1), \quad (4.4.3)$$

where we have divided the multipliers by 2 for convenience. Differentiating (4.4.3) with respect to α and β and setting the derivatives to zero yields

$$\frac{\partial L}{\partial \alpha} = \Sigma_{xy}\beta - (2\lambda/2)\Sigma_{xx}\alpha = 0; \quad (4.4.4)$$

$$\frac{\partial L}{\partial \beta} = \Sigma_{yx}\alpha - (2\mu/2)\Sigma_{yy}\beta = 0. \quad (4.4.5)$$

Pre-multiplying equations (4.4.4) and (4.4.5) by α^T and β^T respectively and subtracting the two, while using the constraints $\alpha^T \Sigma_{xx} \alpha = \beta^T \Sigma_{yy} \beta = 1$ results in $\lambda = \mu = \rho$, a common constant. Next, by letting

$$\Sigma_O = \begin{bmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{bmatrix}, \quad \Sigma_D = \begin{bmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{bmatrix} \quad \text{and} \quad \mathbf{w} = [\alpha, \beta]^T, \quad (4.4.6)$$

equations (4.4.4) and (4.4.5) may be jointly re-written in the generalized eigenvalue form of (4.2.1)

$$\Sigma_O \mathbf{w} = \rho \Sigma_D \mathbf{w}.$$

The above generalized eigenvalue problem can be solved by applying singular value decomposition (SVD) on the matrix

$$K = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2}, \quad (4.4.7)$$

from which the first canonical coefficient variate of $\mathbf{X}$ can be obtained as $\alpha = \Sigma_{xx}^{-1/2} \mathbf{e}_1$, where $\mathbf{e}_1$ is the first left singular vector of K . Similarly, the first canonical coefficient variate of $\mathbf{Y}$ is $\beta = \Sigma_{yy}^{-1/2} \mathbf{f}_1$, and $\mathbf{f}_1$ is the first right singular vector of K . The maximum canonical correlation is $\rho = \lambda_1^{1/2}$, where $\lambda_1^{1/2}$ is the first singular value of K . In general, the i th canonical coefficient variates of $\mathbf{X}$ and $\mathbf{Y}$ are $\alpha_i = \Sigma^{-1/2} \mathbf{e}_i$ and $\beta_i = \Sigma^{-1/2} \mathbf{f}_i$ with $\mathbf{e}_i$ and $\mathbf{f}_i$, $i = 1, \dots, r$ being the i th left and right singular vectors of K respectively where $r = \text{rank}(\Sigma_{xy})$, and $\rho_i = \lambda_i^{1/2}$ is the i th canonical correlation coefficient. In practice, Hotelling (1936) proposed to replace $\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2}$ by the sample versions $\mathbf{S}_{xx}^{-1/2} \mathbf{S}_{xy} \mathbf{S}_{yy}^{-1/2}$, which results in consistent estimators of α and β for fixed dimensions p, q and large sample size n .

4.4.3 SPARSE CANONICAL CORRELATION ANALYSIS

The classical CCA method fails for two reasons when applied to HDLSS problems. Firstly, the population covariance matrices Σ_{xx}^{-1} and Σ_{yy}^{-1} cannot be estimated properly because they are not full rank. A solution to the singularity problem may be to use their respective generalized inverses Σ_{xx}^{-} and Σ_{yy}^{-} . However, using generalized inverses produce unstable estimates that cannot be generalized on data drawn from the same population. A common and stable refinement that fixes the singularity of the sample covariances adds a ridge-type regularization (Hoerl and Kennard, 1970) as done in Vinod (1970). One can also assume that Σ_{xx} and Σ_{yy} are diagonal matrices, which for standardized data are identity matrices, and maximize the covariance instead. This approach showed good results in diagonal LDA (DLDA) proposed by Dudoit et al. (2002) where it was reported that for microarray data, ignoring correlations between genes led to better classification results. Bickel and Levina (2004) also showed better classification performance for naive Bayes (which is equivalent to DLDA for standardized data) than Fisher's LDA under correlated variables. Secondly, in the classical CCA solution, all available variables from both $\mathbf{X}$ and $\mathbf{Y}$ are included in the canonical vectors $\hat{\alpha}$ and $\hat{\beta}$. However, in HDLSS where the number of variables far outnumbers the number of samples, interpreting the canonical vectors is next to impossible. We demonstrate next how the substitution method may be used to obtain sparse canonical covariates.

Assume there exists two sets of sample data $\mathbf{X}$ and $\mathbf{Y}$ with the columns of each set standardized to have zero mean and unit variance. Let $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ be either ridge-corrected sample covariance or identity matrices. To obtain the first canonical covariates, one finds $\hat{\alpha}$ and $\hat{\beta}$ that solves the optimization problem (4.4.2). The zeros of the derivatives of the Langrangian problem (4.4.3) given in equations (4.4.4) and (4.4.5) is recast as the following generalized eigenvalue problem:

$$\begin{bmatrix} 0 & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \rho \begin{bmatrix} \mathbf{S}_{xx} & 0 \\ 0 & \mathbf{S}_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}. \quad (4.4.8)$$

Let $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ be the solution for (4.4.8) and let $\tilde{\rho}_1$ be the first canonical correlation based on $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$. Observe that equation (4.4.8), the solution to the CCA optimization problem (4.4.2) is of form (4.2.1), and hence the substitution method may be applied to obtain the first sparse canonical vectors $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\beta}}_1$ by solving iteratively until convergence

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_1 - \tilde{\rho}_1\mathbf{S}_{xx}\boldsymbol{\alpha}\|_\infty \leq \tau_x \\ \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_1 - \tilde{\rho}_1\mathbf{S}_{yy}\boldsymbol{\beta}\|_\infty \leq \tau_y, \end{aligned} \quad (4.4.9)$$

where $\tau_x > 0$ and $\tau_y > 0$ are tuning parameters controlling how many of the coefficients of the direction vectors will be exactly zero. For the rest of the sparse canonical directions $\hat{\boldsymbol{\alpha}}_k$ and $\hat{\boldsymbol{\beta}}_k$, $k = 2, \dots, r$, $r = \text{rank}(\mathbf{S}_{xy})$, solve iteratively until convergence

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_k - \tilde{\rho}_k\mathbf{S}_{xx}\boldsymbol{\alpha}\|_\infty \leq \tau_{xk} \text{ and } \hat{\boldsymbol{\alpha}}_j^\top \mathbf{S}_{xx}\boldsymbol{\alpha} = \hat{\boldsymbol{\beta}}_j^\top \mathbf{S}_{yx}\boldsymbol{\alpha} = 0, \\ & j = 1, \dots, r-1 \end{aligned}$$

and

$$\begin{aligned} \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_k - \tilde{\rho}_k\mathbf{S}_{yy}\boldsymbol{\beta}\|_\infty \leq \tau_{yk}, \text{ and } \hat{\boldsymbol{\beta}}_j^\top \mathbf{S}_{yy}\boldsymbol{\beta} = \hat{\boldsymbol{\alpha}}_j^\top \mathbf{S}_{xy}\boldsymbol{\beta} = 0, \\ & j = 1, \dots, r-1 \end{aligned}$$

where $\tilde{\rho}_k$ is the k th canonical correlation coefficient between $\mathbf{X}\tilde{\boldsymbol{\alpha}}_k$ and $\mathbf{Y}\tilde{\boldsymbol{\beta}}_k$.

Selection of Tuning Parameters

Our sparse CCA algorithm has two tuning parameters in the optimization set-up that control how many variables of each canonical correlation coefficient vector will be exactly zero, and may be selected from a grid of finite values using V -fold cross validation (Waaijenborg et al., 2008; Parkhomenko et al., 2009) or via permutation tests (Witten and Tibshirani, 2011). In

the cross validation approach, the data is divided into V folds. For each fold except the j th fold, $j = 1, 2, \dots, V$, one solves the CCA problem to obtain $\hat{\alpha}$ and $\hat{\beta}$ respectively. These canonical coefficients are then applied to the testing set, which is the j th fold dataset and the correlation between the canonical pairs are obtained. The optimal tuning parameter pair may then be chosen to maximize the average correlation vectors in the testing set after cycling through V times (Parkhomenko et al., 2009) or may be chosen to minimize the average difference between the canonical correlation of the training and testing sets (Waaijenborg et al., 2008). When the latter is used, the average correlation is

$$Avg_{corr} = \frac{1}{V} \sum_{j=1}^V \left| |\text{corr}(\mathbf{X}^{-j} \hat{\alpha}^{-j}, \mathbf{Y}^{-j} \hat{\beta}^{-j})| - |\text{corr}(\mathbf{X}^j \hat{\alpha}^{-j}, \mathbf{Y}^j \hat{\beta}^{-j})| \right|, \quad (4.4.10)$$

where V is the number of times the cross-validation is performed, $\hat{\alpha}^{-j}$ and $\hat{\beta}^{-j}$ are the canonical coefficients in the training sets $\mathbf{X}^{-j}$ and $\mathbf{Y}^{-j}$ respectively, in which the j th subset was removed. Also, $\mathbf{X}^j$ and $\mathbf{Y}^j$ are the respective test sets. The estimate (4.4.10) is computed for each tuning parameter in the finite set of values and the pair that results in the minimum estimate is selected as the optimal sparseness tuning parameters. A potential drawback with this approach is that there may be a lot of variability in the V correlation estimates since the correlations from the training set is mostly higher than the correlations from the testing sets.

Alternatively, the optimal tuning parameter pair may be chosen via permutation tests (Witten and Tibshirani, 2009). The permutation test approach avoids splitting the set of samples into training and testing as done in cross validation. Instead, the rows of $\mathbf{X}$ are permuted several times, and the sparse canonical coefficient vectors $\hat{\alpha}$ and $\hat{\beta}$ are obtained using the interchanged rows of the data matrix $\mathbf{X}$ and the original data matrix $\mathbf{Y}$. The correlation coefficients are then computed and compared to the correlation coefficient using the original datasets $\mathbf{X}$ and $\mathbf{Y}$. The optimal tuning parameter pair is either chosen to maximize the standardized difference or minimize the p-value of the correlation coefficients. An advantage

of the permutation test is that one can determine whether the canonical pairs result in large correlation only by chance, or are statistically significant. This limitation in the cross validation approach may be overcome by testing for the statistical significance of the canonical pairs after they are obtained. A potential drawback of the permutation test approach is that there is no clear cut line on the number of permutation sets to use as the method is heavily dependent on that; a small number of permutation sets yields highly variable results and a large number of permutation sets increases computational costs.

We use V -fold cross validation but instead of maximizing the average correlation using criterion (4.4.10), we adopt a more natural measure that leverage's the variability in the average correlation by minimizing over the differences between the average canonical correlations from the training and testing sets;

$$Avg_{corr} = \left\| \frac{1}{V} \sum_{j=1}^V \text{corr}(\mathbf{X}^{-j} \hat{\boldsymbol{\alpha}}^{-j}, \mathbf{Y}^{-j} \hat{\boldsymbol{\beta}}^{-j}) - \frac{1}{V} \sum_{j=1}^V \text{corr}(\mathbf{X}^j \hat{\boldsymbol{\alpha}}^{-j}, \mathbf{Y}^j \hat{\boldsymbol{\beta}}^{-j}) \right\|. \quad (4.4.11)$$

The tuning parameter pair (τ_x, τ_y) that yields the criterion (4.4.11) is best selected by performing a grid search over the entire pre-specified set of parameter values. However, this approach is computationally expensive for searching over many grid values. Hence, we select these tuning parameter pair by performing a cross search over the pre-specified set of parameter values. For a fixed value in the τ_y set of values, we search over the entire space of τ_x values and select $\tau_{x_{opt}}$ that minimizes criterion (4.4.11) given τ_y . Using $\tau_{x_{opt}}$, we search the entire τ_y space and choose $\tau_{y_{opt}}$ which also minimizes the criterion (4.4.11). The optimal tuning parameter pair $(\tau_{x_{opt}}, \tau_{y_{opt}})$ is then used to obtain the canonical coefficients $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$.

As stated earlier on, the population covariance matrices $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$, which are not full rank in HDLSS may be estimated by their respective sample generalized inverses or the identity matrices. We choose the latter in the implementation of our sparse CCA algorithm as it has been shown to have good performance in other problems such as naive Bayes.

Our sparse CCA algorithm iteratively solves for $\hat{\alpha}$ and $\hat{\beta}$ until convergence. At the current iteration, previous optimal tuning parameter may be used to obtain current canonical vectors. However, there may be instances where this tuning parameter is too large at the current iteration to produce at least one nonzero coefficient in the canonical vectors. As a result, at each iteration, we choose a different optimal tuning parameter using criterion (4.4.11). Regarding the number of iterations until convergence, our simulations and real data analyses converged mostly at the third iteration, hence we set the maximum iteration in our algorithm to 5.

Algorithm for sparse canonical correlation analysis

1. Let the columns of the data matrices be standardized to have zero mean and unit variance.
2. Select range of tuning parameter values for τ_x and τ_y .
3. For first iteration $i = 1$, obtain first sparse solution vectors $\hat{\alpha}_{11}$ and $\hat{\beta}_{11}$, where the first subscript denotes sparse solution vector and the second subscript the iteration number, by performing V -fold cross validation as follows.
 - (a) For all except the j th fold, $j = 1, 2, \dots, V$, and for each tuning parameter in range of τ_x , fix τ_y and perform a cross search to obtain optimal τ_x
 - i. Initialize by obtaining first nonsparse vectors $\tilde{\alpha}_{11}^{-j}$ and $\tilde{\beta}_{11}^{-j}$, which are respectively the first left and right singular vectors of $\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}$ using all except the j th dataset. Let $\tilde{\rho}_1^{-j}$ be the first singular value.
 - ii. Obtain the first sparse canonical coefficients $\hat{\alpha}_1^{-j}$ and $\hat{\beta}_1^{-j}$ by

$$\begin{aligned} \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\tilde{\beta}_{11}^{-j} - \tilde{\rho}_{1j}^{-j}\mathbf{S}_{xx}\alpha\|_{\infty} \leq \tau_{x_1} \\ \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\tilde{\alpha}_{11}^{-j} - \tilde{\rho}_{1j}^{-j}\mathbf{S}_{yy}\beta\|_{\infty} \leq \tau_{y_1}. \end{aligned}$$

- iii. Normalize $\hat{\alpha}_1^{-j}$ and $\hat{\beta}_1^{-j}$.
- iv. Obtain training correlation coefficient $\hat{\rho}_1^{-j}$ using $\hat{\alpha}_1^{-j}$ and $\hat{\beta}_1^{-j}$ and all except the j th-fold dataset.
- (b) Obtain testing correlation coefficient $\hat{\rho}_1^j$ using the j th-fold dataset and $\hat{\alpha}_1^{-j}$ and $\hat{\beta}_1^{-j}$.
- (c) For the given τ_y , cycle through V times and obtain $\tau_{x_{opt_1}}$ by minimizing criterion (4.4.11).
- (d) For each tuning parameter τ_y and $\tau_{x_{opt_1}}$ repeat steps (a)-(c) to obtain $\tau_{y_{opt_1}}$.
- (e) Using $\tau_{x_{opt_1}}$ and $\tau_{y_{opt_1}}$ and the whole training set, obtain $\hat{\alpha}_{11}$ and $\hat{\beta}_{11}$ and normalize

$$\begin{aligned} \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\tilde{\beta}_{11} - \tilde{\rho}_{11}\mathbf{S}_{xx}\alpha\|_{\infty} \leq \tau_{x_{opt_1}} \\ \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\tilde{\alpha}_{11} - \tilde{\rho}_{11}\mathbf{S}_{yy}\beta\|_{\infty} \leq \tau_{y_{opt_1}}. \end{aligned} \quad (4.4.12)$$

4. For $i = 2$ until convergence, repeat step (3) with updated canonical correlations and coefficients to obtain $\hat{\alpha}_{1i}$, $\hat{\beta}_{1i}$ and hence $\hat{\rho}_{1i}$ by solving

$$\begin{aligned} \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\hat{\beta}_{1_{i-1}} - \hat{\rho}_{1_{i-1}}\mathbf{S}_{xx}\alpha\|_{\infty} \leq \tau_{x_{opt_{1_{i-1}}}} \\ \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\hat{\alpha}_{1_{i-1}} - \hat{\rho}_{1_{i-1}}\mathbf{S}_{yy}\beta\|_{\infty} \leq \tau_{y_{opt_{1_{i-1}}}}. \end{aligned} \quad (4.4.13)$$

5. Update to obtain first sparse solution vectors $\hat{\alpha}_1$, $\hat{\beta}_1$ and hence $\hat{\rho}_1$.
6. For the rest of the sparse canonical directions $\hat{\alpha}_k$ and $\hat{\beta}_k$, $k = 2, \dots, r = \text{rank}(\mathbf{S}_{xy})$, repeat steps (3)-(5) by adding additional constraints appropriately and solving

$$\begin{aligned} \min_{\alpha} \|\alpha\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{xy}\tilde{\beta}_{k_{i-1}} - \tilde{\rho}_{k_{i-1}}\mathbf{S}_{xx}\alpha\|_{\infty} \leq \tau_{x_{k_{opt_{i-1}}}} \\ & \text{and} \quad \hat{\alpha}_{k_{i-1}}^T \mathbf{S}_{xx}\alpha = \hat{\beta}_{k_{i-1}}^T \mathbf{S}_{yx}\alpha = 0, k = 1, \dots, r-1. \\ \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad & \|\mathbf{S}_{yx}\tilde{\alpha}_{k_{i-1}} - \tilde{\rho}_{k_{i-1}}\mathbf{S}_{yy}\beta\|_{\infty} \leq \tau_{y_{k_{opt_{i-1}}}} \\ & \text{and} \quad \hat{\beta}_{k_{i-1}}^T \mathbf{S}_{yy}\beta = \hat{\alpha}_{k_{i-1}}^T \mathbf{S}_{xy}\beta = 0, k = 1, \dots, r-1 \end{aligned}$$

Some Remarks on Algorithm

1. Prior to applying the sparse CCA algorithm, the data is standardized so that all variables have zero means and unit variances by subtracting column means and dividing by column standard deviations. Then the sample variance-covariance matrices $\mathbf{S}_{xx}$, $\mathbf{S}_{yy}$ and $\mathbf{S}_{xy}$ become correlation matrices. The canonical correlations are invariant to scaling so that the correlations using the original variables are the same as using the standardized variables. Also, the canonical coefficient vectors for the standardized variables are related to the canonical coefficient vectors in the original variables. Particularly, if $\boldsymbol{\alpha}_k$ is the k th canonical coefficient vector when the original variables are used, then $\mathbf{V}_{xx}^{1/2} \boldsymbol{\alpha}_k$ is the k th canonical coefficient vector using the standardized variables, where $\mathbf{V}_{xx}^{1/2}$ is a diagonal matrix with i th diagonal element $\sqrt{\sigma_{ii}}$. This is also true for $\boldsymbol{\beta}_k$. Therefore, to convert the coefficients back to the original space, one needs just to multiply each coefficient by $\mathbf{V}_{xx}^{-1/2}$ or $\mathbf{V}_{yy}^{-1/2}$.
2. Normalizing the coefficient vectors by dividing by l_2 norm ensures that they lie in the interval $[-1, 1]$. Constraining the coefficient vectors in this interval usually facilitates a visual comparison of the coefficients, and is recommended if the variables have been standardized.
3. For convergence, we set current iteration to be within ϵ of previous iteration or if the algorithm reaches some number of maximum iterations. That is
 - (a) If either $\|\hat{\boldsymbol{\alpha}}_{k(i)} - \hat{\boldsymbol{\alpha}}_{k(i-1)}\|_2 / \|\hat{\boldsymbol{\alpha}}_{k(i)}\|_2 \leq \epsilon$ or $\|\hat{\boldsymbol{\beta}}_{k(i)} - \hat{\boldsymbol{\beta}}_{k(i-1)}\|_2 / \|\hat{\boldsymbol{\beta}}_{k(i)}\|_2 \leq \epsilon$, stop; else we update old values to current values.
 - (b) If at current iteration, $\hat{\boldsymbol{\alpha}}_{k(i)} = \mathbf{0}$ or $\hat{\boldsymbol{\beta}}_{k(i)} = \mathbf{0}$, stop; set $\hat{\boldsymbol{\alpha}}_{k(i)} = \hat{\boldsymbol{\alpha}}_{k(i-1)}$ or $\hat{\boldsymbol{\beta}}_{k(i)} = \hat{\boldsymbol{\beta}}_{k(i-1)}$.
 - (c) If maximum iteration is reached, stop and set to current solution vectors.

4.4.4 SIMULATION STUDIES

In this section, we compare the proposed method, which we denote SUB, with existing sparse CCA methods. We evaluate the performances with the estimated canonical correlation coefficients, variable selectivity, and Matthew's correlation coefficient. We compare the proposed method with sparse CCA (SCCA) by Parkhomenko et al. (2009), penalized matrix decomposition, (PMD) by Witten and Tibshirani (2009) and sparse CCA via SCAD thresholding (SCCA-SCAD) by Chalise and Fridley (2012).

Let the first dataset $\mathbf{X}$ have $p = 200$ variables and the second dataset $\mathbf{Y}$ have $q = 150$ variables, all drawn on the same sample size $n = 80$. We consider two different sparsity scenarios. In the first scenario, the true canonical vector $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both sparse with 20 and 15 signal variables in each vector respectively. In the second scenario, $\boldsymbol{\alpha}$ is sparse with 20 signal variables, and $\boldsymbol{\beta}$ is nonsparse with all $q = 150$ being signal variables. The true $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the first left and right singular vectors from the SVD of equation (4.4.7). We simulate the data such that the signal variables in $\mathbf{X}$ are correlated with the signal variables in $\mathbf{Y}$ with correlation 0.6. Also, the data $(\mathbf{X}, \mathbf{Y})$ are simulated with joint probability distribution from $\text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix}$$

is the joint covariance matrix, and $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$, $\boldsymbol{\Sigma}_{xy}$, are covariance matrices within $\mathbf{X}$ and $\mathbf{Y}$ respectively, and between them.

Let $\mathbf{X}_s$ and $\mathbf{Y}_s$ be the signal variables and $\mathbf{X}_n$ and $\mathbf{Y}_n$ be the noise variables in each dataset. Also let $\rho(\mathbf{X}_s, \mathbf{X}_s)$ and $\rho(\mathbf{Y}_s, \mathbf{Y}_s)$ be the correlations between signal variables in each dataset. Similarly, let $\rho(\mathbf{X}_n, \mathbf{X}_n)$ and $\rho(\mathbf{Y}_n, \mathbf{Y}_n)$ be the correlations between noise variables in each dataset. Denote the cross-correlation of signal variables between the datasets as $\rho(\mathbf{X}_s, \mathbf{Y}_s)$. Four settings that differ in the strength of association are considered. We first consider the settings for scenario one where both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are sparse.

- Setting I - High $\rho(\mathbf{X}_s, \mathbf{X}_s)$, $\rho(\mathbf{Y}_s, \mathbf{Y}_s)$ and low $\rho(\mathbf{X}_n, \mathbf{X}_n)$, $\rho(\mathbf{Y}_n, \mathbf{Y}_n)$

$$\Sigma_{xx} = \begin{bmatrix} \tilde{\Sigma}_{20 \times 20} & \mathbf{0} \\ \mathbf{0} & \bar{\Sigma}_{p-20} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} \tilde{\Sigma}_{15 \times 15} & \mathbf{0} \\ \mathbf{0} & \bar{\Sigma}_{q-15} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.6\mathbf{J}_{20 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\tilde{\Sigma} = 0.7\mathbf{J} + (1 - 0.7)\mathbf{I}$, $\bar{\Sigma} = 0.1\mathbf{J} + (1 - 0.1)\mathbf{I}$

- Setting II - High $\rho(\mathbf{X}_s, \mathbf{X}_s)$, $\rho(\mathbf{Y}_s, \mathbf{Y}_s)$ and zero $\rho(\mathbf{X}_n, \mathbf{X}_n)$, $\rho(\mathbf{Y}_n, \mathbf{Y}_n)$

$$\Sigma_{xx} = \begin{bmatrix} \tilde{\Sigma}_{20 \times 20} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-20} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} \tilde{\Sigma}_{15 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q-15} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.6\mathbf{J}_{20 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\tilde{\Sigma} = 0.7\mathbf{J} + (1 - 0.7)\mathbf{I}$

- Setting III - Moderately low $\rho(\mathbf{X}_s, \mathbf{X}_s)$, $\rho(\mathbf{Y}_s, \mathbf{Y}_s)$ and low $\rho(\mathbf{X}_n, \mathbf{X}_n)$, $\rho(\mathbf{Y}_n, \mathbf{Y}_n)$

$$\Sigma_{xx} = \begin{bmatrix} \tilde{\Sigma}_{xx20 \times 20} & \mathbf{0} \\ \mathbf{0} & \bar{\Sigma}_{p-20} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} \tilde{\Sigma}_{yy15 \times 15} & \mathbf{0} \\ \mathbf{0} & \bar{\Sigma}_{q-15} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.35\mathbf{J}_{20 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\tilde{\Sigma}_{xx15 \times 15} = 0.5\mathbf{J} + (1 - 0.5)\mathbf{I}$, $\bar{\Sigma} = 0.3\mathbf{J} + (1 - 0.3)\mathbf{I}$, $\tilde{\Sigma}_{yy15 \times 15} = 0.3\mathbf{J} + (1 - 0.3)\mathbf{I}$

- Setting IV - Moderately low $\rho(\mathbf{X}_s, \mathbf{X}_s)$, $\rho(\mathbf{Y}_s, \mathbf{Y}_s)$ and zero $\rho(\mathbf{X}_n, \mathbf{X}_n)$, $\rho(\mathbf{Y}_n, \mathbf{Y}_n)$

$$\Sigma_{xx} = \begin{bmatrix} \tilde{\Sigma}_{xx20 \times 20} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-20} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} \tilde{\Sigma}_{yy15 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q-15} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.35\mathbf{J}_{20 \times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\tilde{\Sigma}_{xx} = 0.5\mathbf{J} + (1 - 0.5)\mathbf{I}$, $\tilde{\Sigma}_{yy} = 0.3\mathbf{J} + (1 - 0.3)\mathbf{I}$

In scenario two where $\boldsymbol{\alpha}$ is sparse and $\boldsymbol{\beta}$ is nonsparse, we consider similar settings but make the following changes to the covariance matrix Σ_{yy} :

$$\Sigma_{yy} = \begin{cases} \tilde{\Sigma}_{150 \times 150} & \text{for Settings I and II} \\ \bar{\Sigma}_{yy200 \times 200} & \text{for Settings III and IV} \end{cases},$$

where $\tilde{\Sigma} = 0.7\mathbf{J} + (1 - 0.7)\mathbf{I}$ and $\bar{\Sigma}_{yy} = 0.3\mathbf{J} + (1 - 0.3)\mathbf{I}$.

In the analysis, we generate 20 realizations of data for each setting. We use 5-fold cross validation to select the optimal tuning parameters from criterion (4.4.11), and then obtain $\hat{\alpha}$ and $\hat{\beta}$ using the whole training set. What follows is a discussion of the results.

Figure 4.4.1 is a plot of the average estimated canonical correlation coefficient from canonical covariates $(\mathbf{X}\hat{\alpha}, \mathbf{Y}\hat{\beta})$ for the four methods. We also show the true maximum canonical correlation for reference. Recall that $\hat{\alpha}$ and $\hat{\beta}$ produced by the methods yield maximum correlation between the two datasets. From the plot, one can notice that when both α and β are sparse, the four methods have correlation coefficients that approximate the truth well in settings I, II, and IV. In particular, the substitution method has smaller correlation bias in all three settings. However, the correlation coefficient is underestimated in setting III by the methods, much more so by the substitution method. In setting III, the correlation between noise variables within each dataset may clutter the relevant variables selected by the methods. The poor performance in setting III is also noticeable from the variable selection plots in Figures 4.4.2 - 4.4.5. These figures show the number of variables selected by the methods, which is the height of each bar, decomposed into the number of signal variables selected (True Positives, TP), and the number of noise variables selected (False Positives, FP). From Figures 4.4.3, and 4.4.5, we observe that all methods, especially the substitution method, select more false positives in setting III when compared to the other settings. This may be because of the moderately high correlation among the noise variables, causing the methods to read those as signals and hence select them. Even though they are selected, there is not much information in them to contribute to the correlation between the datasets. This is seen from the relatively low canonical correlation value in setting III shown in the left panel of figure 4.4.1. It is interesting to note the superior performance of the substitution method in all the other settings. Our method selects only signal variables for the canonical variate $\hat{\alpha}$, and very small noise variables in $\hat{\beta}$ as seen in Figures 4.4.2 - 4.4.5.

When α is sparse and β is nonsparse, indicating that all $q = 150$ variables in β are important, we notice that in general, the methods select less false positives in all settings when compared to both α and β sparse. The poor performance by the methods in setting III, with the exception of PMD, is non-existent in this case. Across all four settings, the substitution method selects all 20 and 150 signal variables in $\hat{\alpha}$ and $\hat{\beta}$ respectively. However, it selects some few noise variables in $\hat{\alpha}$ which, when compared to the other methods is small. PMD is very sparse; the method selects less signals, and erroneously assigns zero weights to most of the signal variables.

It would be interesting to consider the variables selected and the level of sparsity in tandem. To this end, we use Matthew correlation coefficient (MCC), which ties together variable selectivity and sparsity. The MCC formula is

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}, \quad (4.4.14)$$

where TN is the number of noise variables with zero weights, FN is the number of signals considered as noise and therefore not selected, and TP and FP are as defined before. MCC lies in the interval $[-1, 1]$. A value of 1 corresponds to selection of all signal variables and no noise variables, a perfect estimation. A value of -1 indicates total disagreement between what are signal and noise variables, and a value of 0 indicates random guessing. If any of the four sums in the denominator is zero, which is the case for β when α is sparse and β is nonsparse (since TN and FN are zero), the denominator may be set to 1, giving a MCC of 0. We do not discuss the MCC for this case.

Figure 4.4.6 is a plot depicting the estimated MCC values by the methods. Again, we observe a superior performance by the substitution method in all but setting III when α and β are sparse. In particular, we achieve a perfect estimation for $\hat{\alpha}$ in setting I and close to perfect in settings II and IV. For $\hat{\beta}$, our MCC values are very high and close to the true value than the other methods. Again, in setting III, we have a worse performance which, as

discussed before, may be because of the correlation between the noise variables within each data set. When α is sparse and β is nonsparse, our substitution method is again superior.

The performance of the substitution method is overwhelming. We can say that we are better in cases where one can assume that some variables in each high dimensional dataset are noise variables not contributing to the correlation between the datasets, and where these noise variables are themselves uncorrelated. This may be the case in many microarray studies where gene expression data are correlated within a pathway and uncorrelated between pathways.

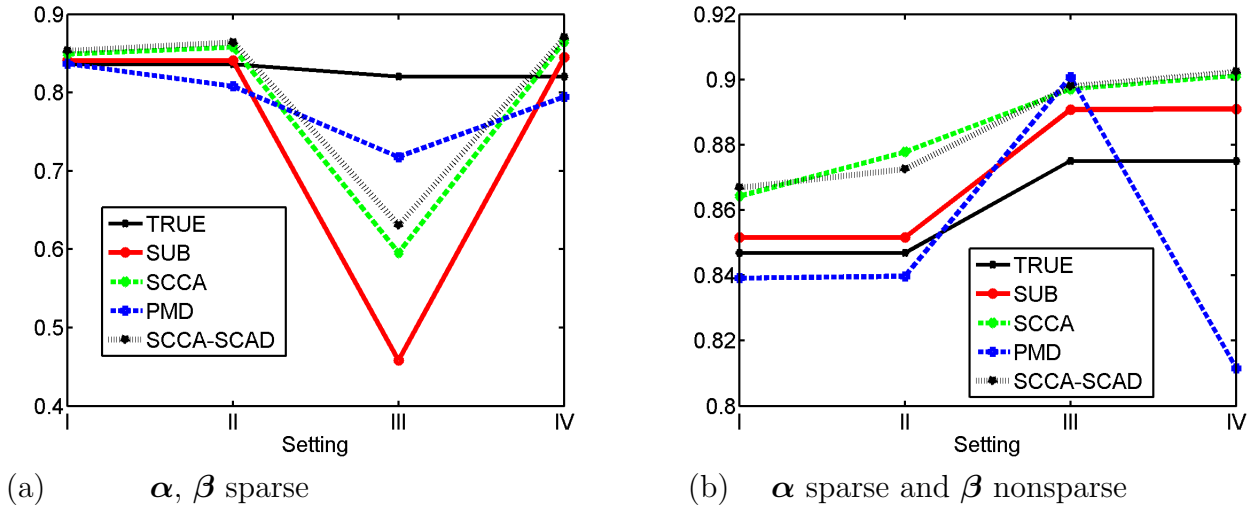
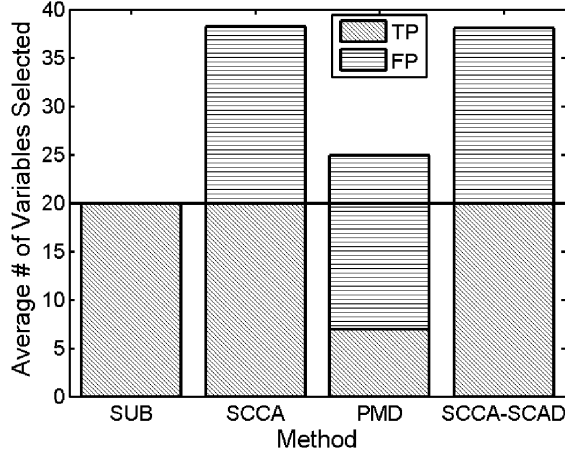
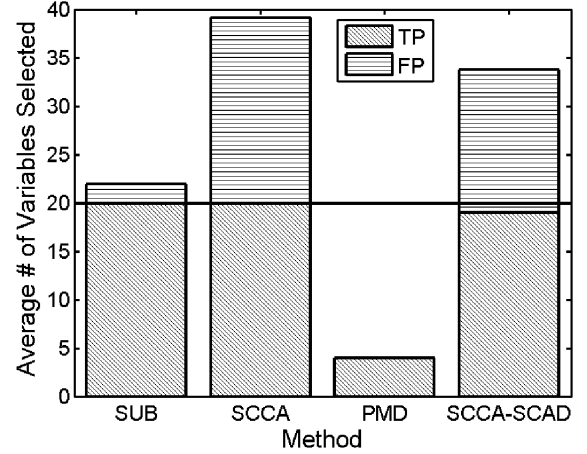


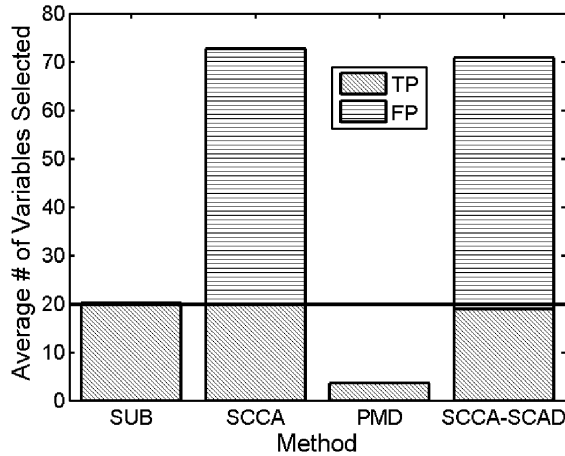
Figure 4.4.1: Average maximum canonical correlation coefficients based on 20 repetitions for α and β sparse (left panel), and when β is nonsparse (right panel). Compared to the true canonical correlations, the substitution method has smaller bias, with the exception of Setting III when both canonical variates are sparse. The poor performance in this setting may be attributed to the correlation between noise variables in each dataset. This relatively high correlation, when compared to other settings causes the substitution method to select these noise variables as signals. However when β is nonsparse, all variables are important. This has the effect of overshadowing the noise effects in α , thus resulting in better correlation estimates in comparison to when both are sparse.



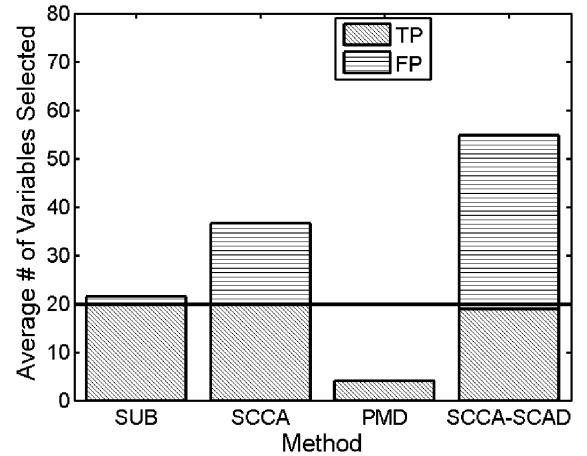
(a) Setting I - α, β Sparse



(b) Setting I - α Sparse, β Non-Sparse

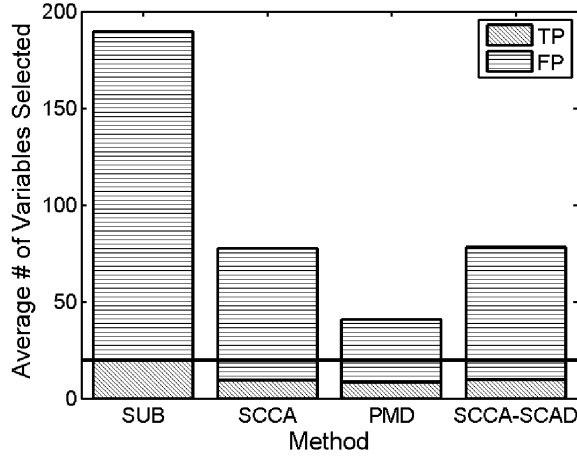


(c) Setting II - α, β Sparse

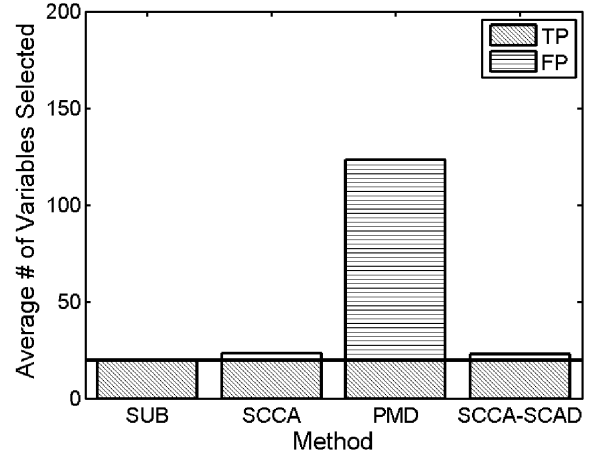


(d) Setting II - α Sparse, β Non-Sparse

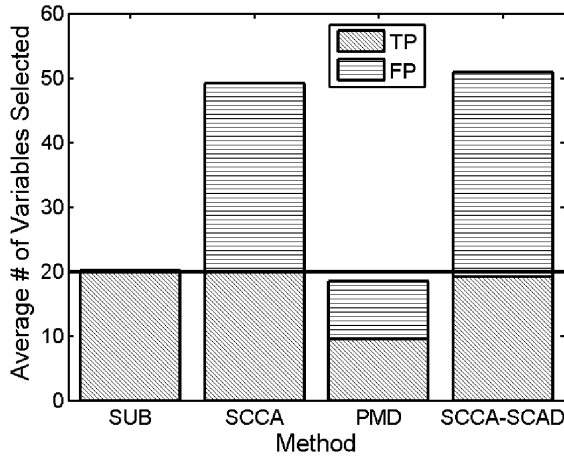
Figure 4.4.2: Variable selection properties of $\hat{\alpha}$ for each method under Settings I and II. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables in the bottom and selected noise variables in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.



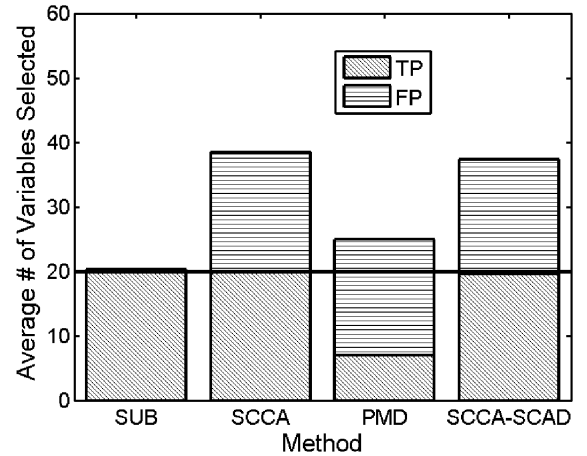
(a) Setting III - α, β Sparse



(b) Setting III - α Sparse, β Non-Sparse

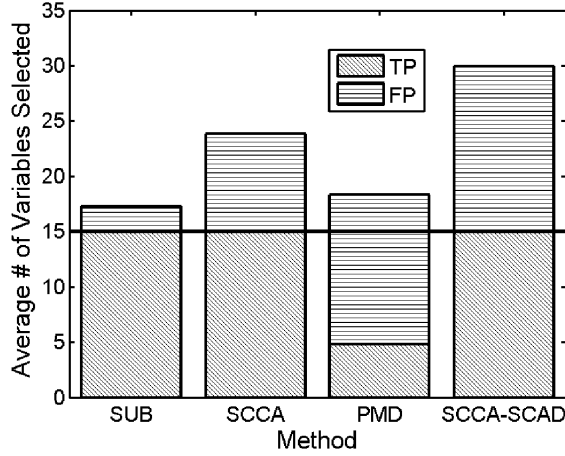


(c) Setting IV - α, β Sparse

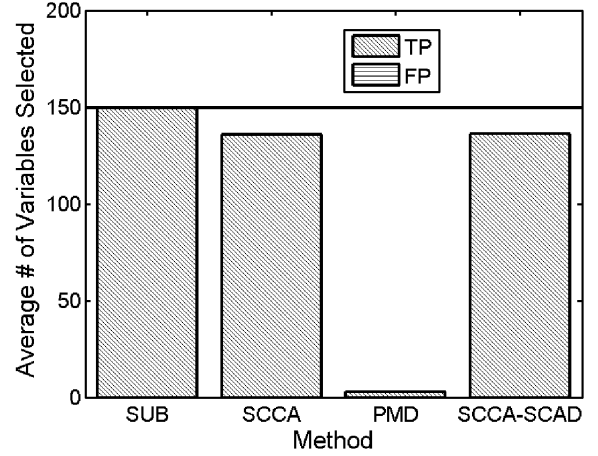


(d) Setting IV - α Sparse, β Non-Sparse

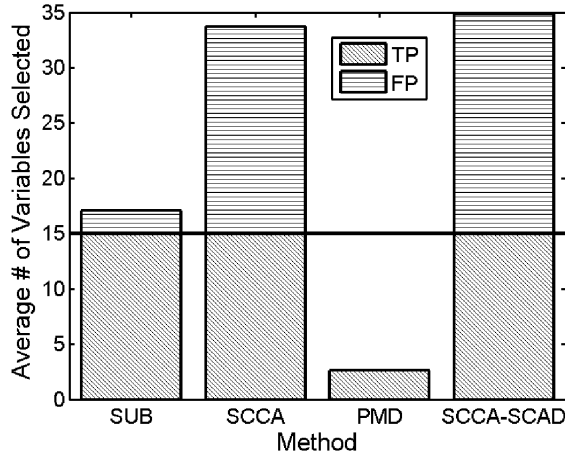
Figure 4.4.3: Variable selection properties of $\hat{\alpha}$ for each method under Settings III and IV. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables in the bottom and selected noise variables in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.



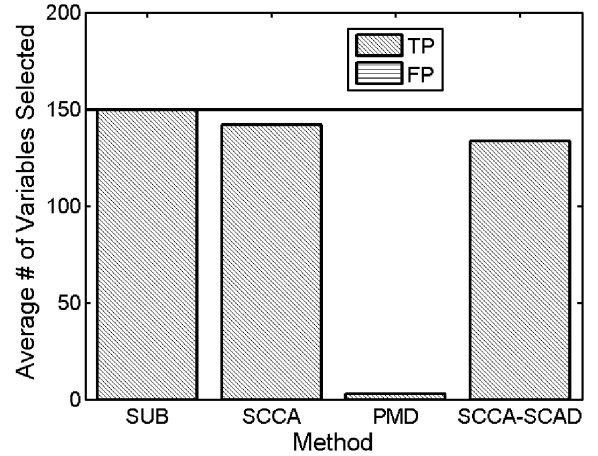
(a) Setting I - α, β Sparse



(b) Setting I - α Sparse, β Non-Sparse

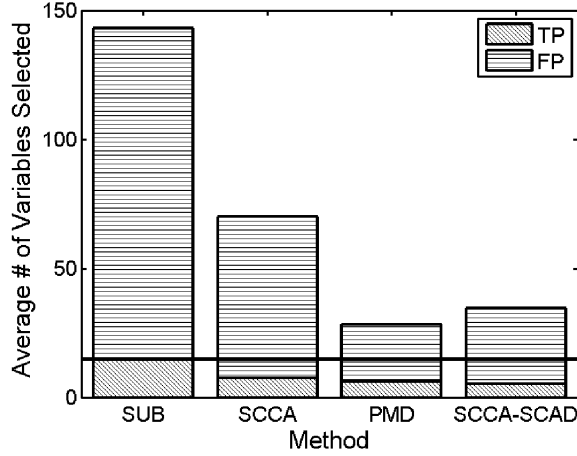


(c) Setting II - α, β Sparse

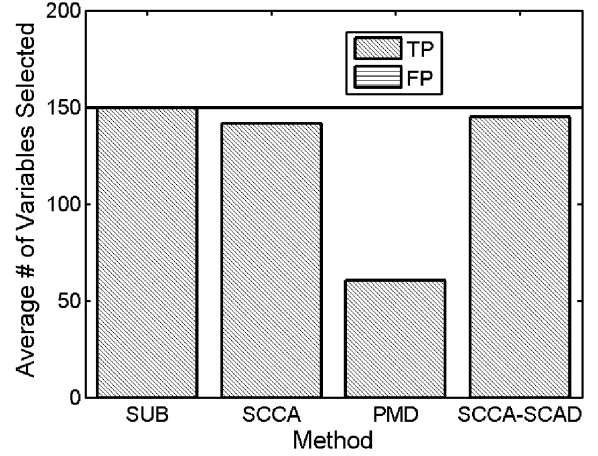


(d) Setting II - α Sparse, β Non-Sparse

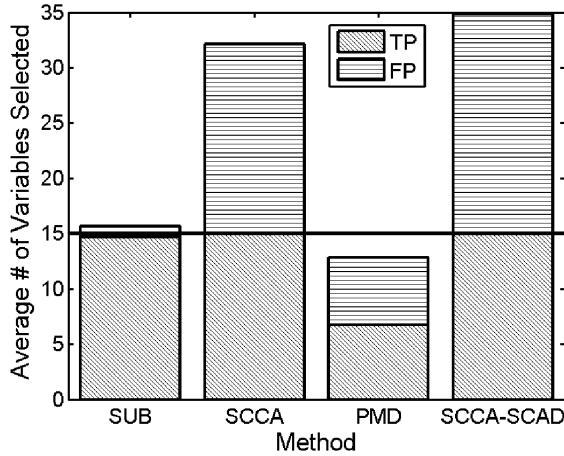
Figure 4.4.4: Variable selection properties of $\hat{\beta}$ for each method under Settings I and II. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables in the bottom and selected noise variables in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.



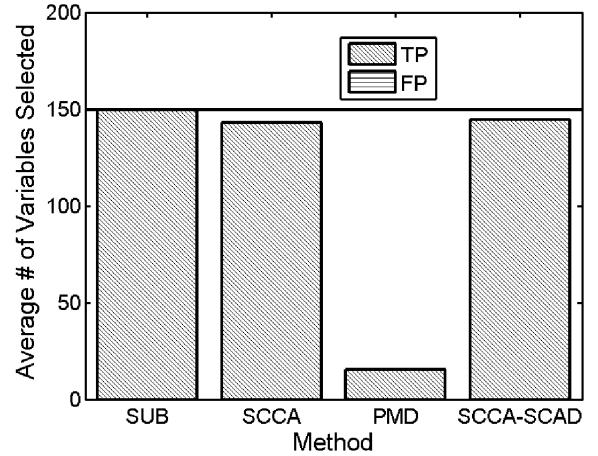
(a) Setting III - α, β Sparse



(b) Setting III - α Sparse, β Non-Sparse

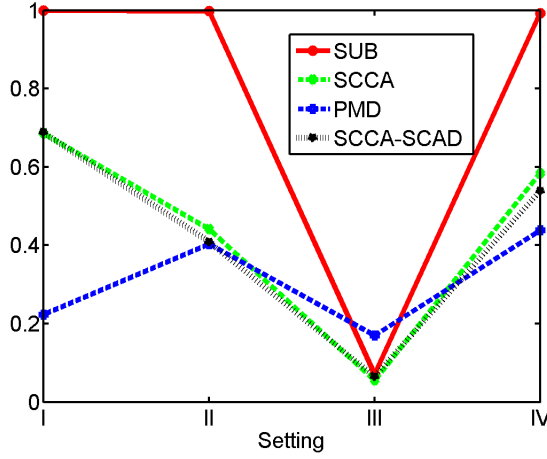


(c) Setting IV - α, β Sparse

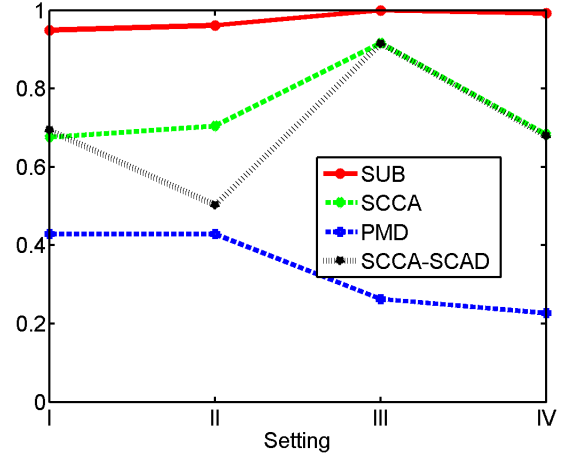


(d) Setting IV - α Sparse, β Non-Sparse

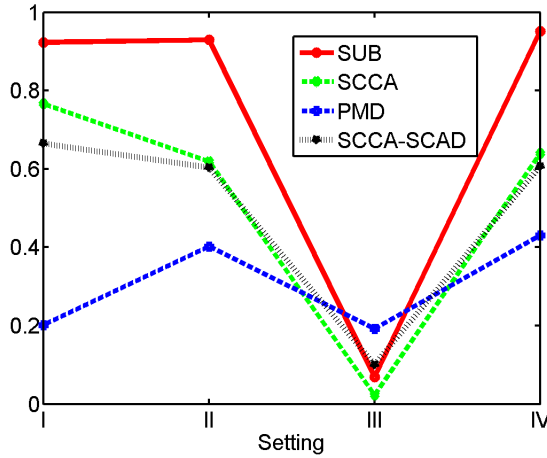
Figure 4.4.5: Variable selection properties of $\hat{\beta}$ for each method under Settings III and IV. The total height of each bar represents the average number of selected variables, which is a sum of the number of selected signal variables in the bottom and selected noise variables in the top of the bar. The horizontal line representing the number of true signal variables is added for a reference.



(a) $\hat{\alpha}$ MCC - α, β Sparse



(b) $\hat{\alpha}$ MCC - α Sparse, β Non-Sparse



(c) $\hat{\beta}$ MCC - α, β Sparse

Figure 4.4.6: Matthew's correlation coefficients for $\hat{\alpha}, \hat{\beta}$ when both α, β are sparse and when only α is sparse. The substitution method yields better MCC value in most of the settings. This suggests that our method does not only select the right variables, but also has less false negatives. Note that the MCC when β is nonsparse is assumed zero and not discussed.

4.4.5 APPLICATION OF PROPOSED METHOD ON GENOMIC DATASETS

It is becoming more common in genomic research to use multiple measurements to characterize the same set of patients. For instance, DNA copy number variations (CNV) and gene expression data might be available on the same sets of patients. CNV, a form of structural variation in the genome, are changes in the DNA of a genome that results in the cell having an abnormal regions of the DNA. These structural variations include insertions, deletions and duplications of the genome. CNV can affect phenotypes by altering the level of genes and gene products, and may lead to the development of complex diseases in the presence of genetic or environmental factors.

There has been extensive research to analyze these data sets separately, but few research to jointly combine these datasets to study interrelations between them. If DNA copy number and gene expression data are available on the same set of patients, then it would be interesting to identify set of genes that have expression levels that is correlated with chromosomal gains or losses. CCA may be used to integrate these genomic datasets to study the idea that changes in DNA copy could have an effect on the expression of genes and also changes in gene expression levels may be caused by changes in DNA copy. This has been demonstrated in (Waaijenborg et al., 2008; Parkhomenko et al., 2009; Witten et al., 2009; Witten and Tibshirani, 2009).

We investigate the performance of the substitution method on a breast cancer data set publicly available at Witten et al. (2013). The dataset is made up of $n = 89$ samples for which both DNA copy and gene expression measurements are available. Without loss of generality, let $\mathbf{X}$ be gene expression measurements and $\mathbf{Y}$ be copy number changes. There are $p = 19,672$ gene expression measurements and $q = 2,149$ DNA copy number measurements. There were 23 different chromosomes making up the DNA copy data, with their associated genes in the gene expression dataset. In the analysis, we removed variables in $\mathbf{X}$ that did not have chromosomes in $\mathbf{Y}$. We also filtered out genes that had low profile variance resulting in a

final dataset with 17,333 genes in $\mathbf{X}$ and 1,934 genes in $\mathbf{Y}$. The data were normalized to have mean zero and unit variance for each gene.

We first analyze the breast cancer data using all the gene expression measurements and chromosome one. Figure 4.4.7 gives the distribution of the number of genes located on each chromosome for $\mathbf{X}$ and $\mathbf{Y}$. The goal of the analysis is to find regions of copy number variation on chromosome one that are correlated with gene expression measurements anywhere on the genome. Since we use only chromosome one in the DNA copy dataset, most of all the variables that are selected to have nonzero weights in the gene expression by the methods should be located on chromosome one.

Table 4.4.1 shows the number of variables that had both zero and nonzero weights in each canonical vector. From columns 2-5, one may observe that substitution method results in more sparse canonical vectors. Substitution method identified 273 gene expressions and 58 DNA variables with correlation 0.8036. Also, out of 273 variables selected by substitution method, 245 of them were located on chromosome one. This is noteworthy because the analysis was performed with CNV variables on chromosome one and hence for cis interactions, CNV measurements on chromosome one should be correlated with the gene expressions on chromosome one. Figure 4.4.8 gives a distribution of locations of chromosomes in the gene expressions sets found to be correlated with CNV measurements on chromosome one. Notice that the substitution method did not identify expression genes on some chromosomes as being correlated with CNV measurements on chromosome one. We also observed that all variables in α selected by substitution method was a subset of the variables selected by $\text{PMD}(L_1, FL)$. For $\hat{\beta}$, all but one genes selected by substitution is a subset of $\text{PMD}(L_1, FL)$. We note that $\text{PMD}(L_1, FL)$ uses the default settings found in Witten et al. (2013). If our proposed method selects fewer variables with maximum correlation similar to $\text{PMD}(L_1, FL)$, then it could be suggested that we do not need that many variables to determine chromosomal loss or gains in the gene expression measurements. Figures 4.4.9 and 4.4.10 are graphical displays of the

strength of association between the canonical covariates and the variable selection properties of CNV canonical vectors.

For further analysis of the canonical covariates, we determine whether the substitution method is capturing real structure in the breast cancer data. Due to the large dimensionality of the variables, there is a large probability that the estimated canonical variates have high correlation by chance. To assess whether the canonical correlation, and hence the selected variables in each canonical correlation vector is not random, we perform statistical significance test of the estimated correlation, $\hat{\rho}$, from the canonical variates pair $(\mathbf{X}\hat{\boldsymbol{\alpha}}, \mathbf{Y}\hat{\boldsymbol{\beta}})$ via permutation tests. Permutation test is a nonparametric approach for determining statistical significance based on rearrangement of samples of a dataset. The estimated correlation from the original dataset would be compared with the distribution of correlation estimates from the permuted dataset. The null hypothesis is that the data are exchangeable. Differently put, we have

H_0 : average correlation from permuted data equals estimated correlation from original data

H_A : correlation from original data is greater than average correlation from permuted data.

If H_0 is true, then the high correlation obtained from the estimated canonical vectors $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and the original dataset is just by chance, implying that the canonical vectors may not be showing any real structure in the breast cancer dataset. If H_0 is false, the estimated canonical correlation may not be random, and hence the canonical vectors may be capturing real effect in the breast cancer dataset. We obtain the p-value as follows (Witten et al., 2009):

- Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ be the canonical vectors selected by the substitution method via 5-fold CV. Denote the correlation between the canonical covariates using the original dataset as c . That is $c = \text{Corr}(\mathbf{X}\hat{\boldsymbol{\alpha}}, \mathbf{Y}\hat{\boldsymbol{\beta}})$.

- For $i = 1, \dots, P$, P large, permute the samples in $\mathbf{X}$ to obtain $\mathbf{X}^*$; then compute $\hat{\boldsymbol{\alpha}}^*$ and $\hat{\boldsymbol{\beta}}^*$ using the substitution method and based on data $(\mathbf{X}^*, \mathbf{Y})$. Obtain the canonical correlation between $\hat{\boldsymbol{\alpha}}^*$ and $\hat{\boldsymbol{\beta}}^*$ and the permuted dataset as $c_i^* = \text{Corr}(\mathbf{X}^* \hat{\boldsymbol{\alpha}}^*, \mathbf{Y} \hat{\boldsymbol{\beta}}^*)$.
- The p-value is obtained as (Phipson and Smyth, 2010)

$$\frac{1 + \sum_{i=1}^P I_{|c_i| \geq |c|}}{1 + P}. \quad (4.4.15)$$

From our analyses with $P = 100$, we obtain a p-value of 0.0099 which is statistically significant at a level of 0.05. This suggests that the estimated canonical covariates have high correlation not by chance but may be depicting real associations between the gene expressions and CNV datasets.

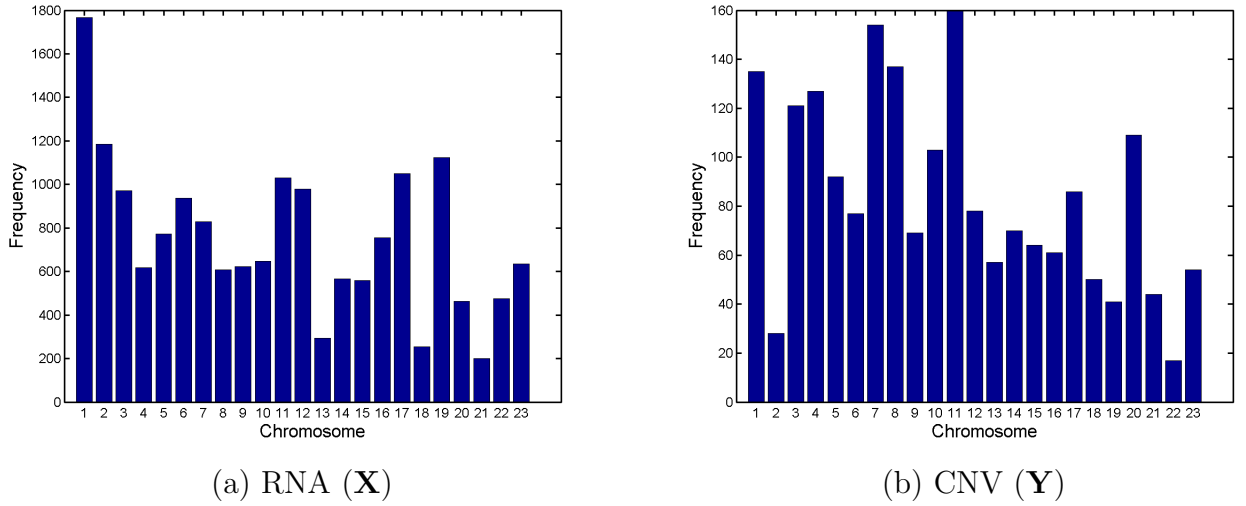
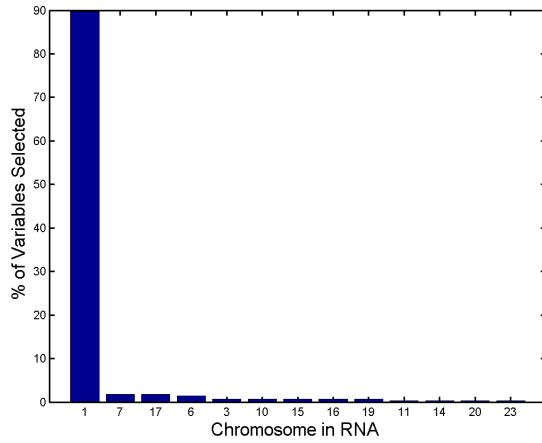


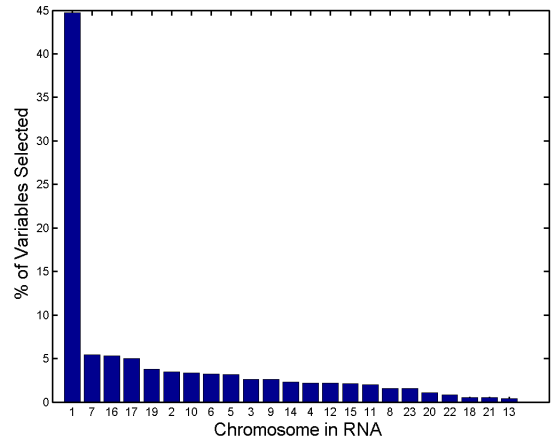
Figure 4.4.7: Distribution of genes and copy number variations for breast cancer data.

4.4.6 CONCLUSION

Canonical correlation analysis finds weighted combinations of variables within each dataset that have maximal correlation between the different sets of data. However, CCA uses all available variables in the weighted combinations. The application of CCA on studies with



(a) Substitution



(b) PMD(L_1, FL)

Figure 4.4.8: Distribution of gene expressions selected on chromosome one. A higher percentage of CNV measurements on chromosome one were found to be correlated with gene expressions on chromosome one by the substitution method, depicting a higher cis interactions. Notice that the substitution method did not identify gene expressions located on some chromosomes as being correlated with CNV measurements on chromosome one.

Method	#nonzero $\hat{\alpha}$	#zeros $\hat{\alpha}$	#nonzero $\hat{\beta}$	#zeros $\hat{\beta}$	$\hat{\rho}$	% nonzeros on chromosome one
SUB	273	17060	58	77	0.8036	89.74
PMD(L_1, FL)	955	16378	60	75	0.8050	44.71

Table 4.4.1: Comparison of the Substitution method and PMD(L_1, FL) using all gene expression measurements and chromosome one in the DNA copy dataset. It may be observed that Substitution method results in more sparse loadings and shows higher cis interactions. Also, both methods have comparable canonical correlation coefficient, but the substitution method selected fewer variables. We note that all the variables selected in $\hat{\alpha}$ by the substitution method are subset of PMD, and all but a gene in $\hat{\beta}$ are subset of PMD.

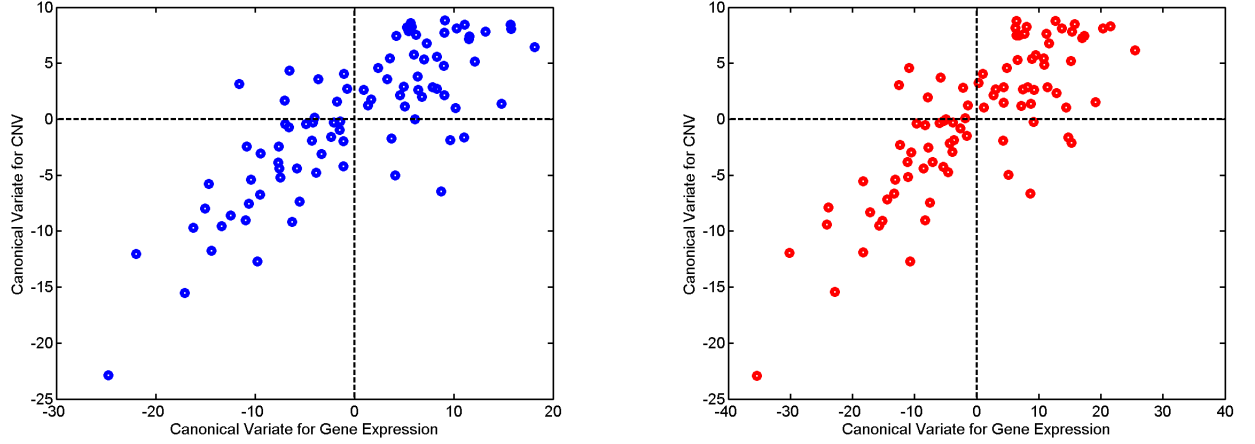


Figure 4.4.9: Estimated gene expression variates and CNV variates. The left panel shows the projection of data onto $(\hat{\alpha}, \hat{\beta})$ from substitution method. The right panel is a projection of data onto $(\hat{\alpha}, \hat{\beta})$ from $\text{PMD}(L_1, FL)$. The strength of association between the gene expressions and CNV is comparable in the two methods. However, substitution method uses fewer variables compared to $\text{PMD}(L_1, FL)$ to achieve similar correlation value (refer to Table 4.4.1). The CCA variates from substitution yield maximum canonical correlation coefficient of 0.8036, and the CCA variates from $\text{PMD}(L_1, FL)$ results in a maximum canonical correlation coefficient of 0.8050.

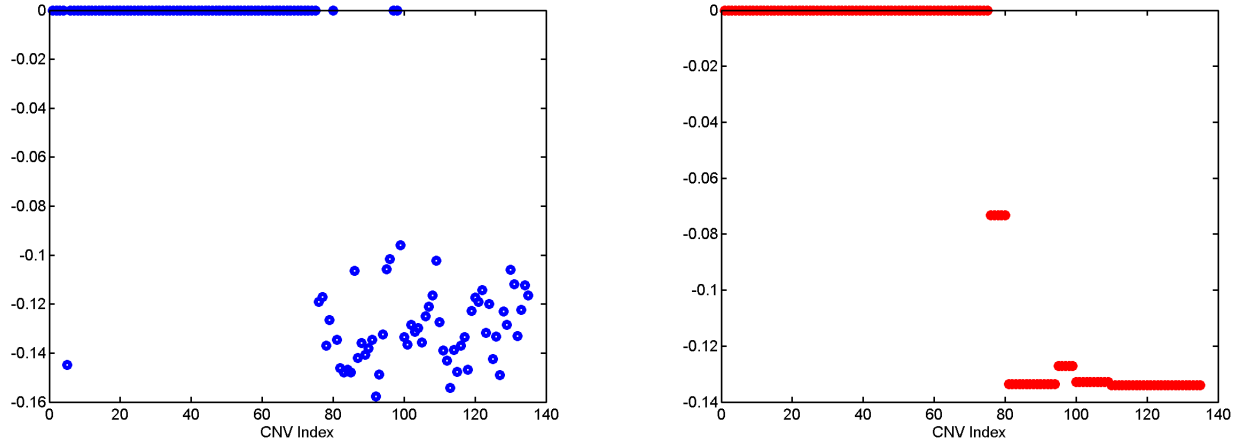


Figure 4.4.10: CNV canonical correlation vectors compared. The left panel shows CNV variable selection for substitution method. The right panel shows CNV variable selection property for $\text{PMD}(L_1, FL)$. Both methods select some common variables. Of the 58 variables selected by the substitution method, 57 is a subset of the 60 variables selected by $\text{PMD}(L_1, FL)$.

very large number of variables compared to the sample size may not be practical due to the high dimensional nature, and it may lack interpretability. In this section, we have introduced a new method of obtaining linear combinations of variables that use only a fraction of the variables through the substitution method. These sparse CCA vectors are easier to interpret and may serve as inputs to other statistical analysis.

Our simulation studies revealed an overwhelming performance of the substitution method compared to some existing methodologies in terms of estimated CCA correlation, variables selected and estimated Matthew’s correlation coefficient on the training sample. The difference is especially large in cases where one can assume that some variables in each dataset are noise variables that do not contribute to the correlation between the datasets, and where these variables are themselves not correlated. This is the case in many microarray studies where gene expression data are correlated within a pathway and are independent between pathways.

We also applied the substitution method on a publicly available breast cancer data which had both gene expressions and copy number variations measurements on the same set of patients. We demonstrated the application of our method on chromosome one in the CNV data, with the goal of finding chromosomal locations in the expression sets that are correlated with CNV on chromosome one. We observed a high cis interactions compared to PMD (Witten and Tibshirani, 2009) meaning that most of the nonzero weights assigned by the substitution method in the expression sets that shows strong correlation with chromosome one of CNV where located on chromosome one in the gene expression data. We also observed that our method selected fewer variables than PMD but both had similar canonical correlation coefficient, indicating that we may not need that many variables in studying associations of gene expression sets and CNV.

In our simulations and real data analysis, we focused on obtaining and explaining the first canonical correlation coefficient and variates. This is in no way a reflection of a limitation

of our methodology, as our algorithm can be used to obtain subsequent canonical variates and coefficients. On the other hand, our methodology focused on two sets of variables. When there are more than two high dimensional measurements on the same sets of samples, our proposed method cannot be used directly. An extension of the substitution method to deal with multiple datasets is underway.

A general limitation of canonical correlation analysis is that it only studies linear associations between sets of variables. This may not be useful in complex situations where the association is nonlinear. Kernel CCA has been proposed in the literature to study nonlinear associations between sets of variables. Kernel CCA finds functions $f(X)$ and $g(Y)$ in a RKHS such that these functions have maximal correlation. For future work, the substitution method will be extended to Kernel CCA to obtain sparse solution vectors.

4.5 DISCUSSION

We have developed a framework for obtaining sparse solution vectors for high dimension, low sample size problems. Our methodology capitalized on the idea of generalized eigenvalue problem, which many multivariate statistical problems can be recast into, to develop solution vectors that have zero weights on some of the variables. The solution vectors from the traditional generalized eigenvalue problem uses all available variables, which in HDLSS makes interpretation of results practically impossible. Hence, for sparse solution vectors, we imposed l_∞ constraint on the generalized eigenvalue problem. We showed that naively bounding with l_∞ results in a trivial solution vector, meaning that no variable is selected. Thus, to ensure that at least one variable is selected, we substituted the left term in the generalized eigenvalue equation with the nonsparse solution vector before imposing l_∞ constraint. This approach was termed the substitution method.

We demonstrated the use of the substitution method in linear discriminant analysis and canonical correlation analysis. The substitution method applied to LDA showed a compet-

itive performance in terms of test error rates and variable selectivity. In particular, it was demonstrated that LDA via substitution method tends to select fewer yet signal variables. For canonical correlation analysis, we observed a superior performance of the substitution method in terms of canonical correlation coefficients, variable selectivity and Matthew's correlation coefficient in a simulation study. The simulations revealed that the substitution method works well and is superior in cases where there is low or no correlation between noise variables in each dataset. It is worth mentioning that the substitution method can be applied to obtain sparse solution vectors in several multivariate statistical problems such as principal component analysis, multiple linear regression and multiple analysis of variance, to mention but a few.

4.6 REFERENCES

- Bickel, P. J. and Levina, E. (2004). Some theory for fisher’s linear discriminant function, naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.
- Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, (1):199–227.
- Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.
- Chalise, P. and Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics and Data Analysis*, 56:245–254.
- Dudoit, S., Fridlyand, J., and Terence, P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression. *Technometrics*, pages 152–177.
- Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, pages 312–377.

- Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8.
- Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1544–6115.
- Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Vinod, H. D. (1970). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, pages 147–166.
- Waaijenborg, S., de Wit Hamar, P. C. V., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7.
- Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). Package ‘pma’. <http://cran.r-project.org/web/packages/PMA/PMA.pdf>. Version 1.0.9.
- Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.
- Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8.

- Witten, D. M., Tibshirani, R. J., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

CHAPTER 5

CONCLUSION

In this chapter, we provide a few closing remarks regarding the performance of our proposed methods and some suggestions for future work.

In Chapter 2, a new sample size method for training regularized logistic regression was developed. A sample size n is adequate if the developed predictor's expected performance is close to the optimal performance (either logistic regression slope or misclassification error) as $n \rightarrow \infty$. The sample size method proposed exploits structural similarity between regularized logistic regression prediction and errors-in-variables models. In particular, errors-in-variables models were used to recover asymptotic slope of the logistic regression. The sample size method was shown to perform well on a pilot dataset. If no pilot dataset exists, the method can be used with samples from Monte Carlo simulations. The method was shown to provide better sample size estimates on simulated data, and seemed to provide more reasonable estimate on real data in comparison to existing methodology.

The reader may have noted that classification scores W , and bootstrap variance, σ^2 , of the errors-in-variables model were estimated by resampling from the pilot dataset using cross validation and bootstrap techniques respectively. Feature selection procedure was embedded in these resampling techniques. Since the bootstrap dataset has overlap and contains about $0.632n$ independent samples, there is potential bias in the bootstrap estimation procedure. As a future work, the bootstrap procedure may be modified to reduce bias in estimation. One way of modifying the bootstrap procedure is to ensure that the stringency parameter

selected in the feature selection procedure in the bootstrap algorithm matches that of the cross validation algorithm. If the stringency parameter in the logistic regression is selected using deviance, then the multiple occurrences of samples in the bootstrap data contribute more than once in estimating the logistic coefficients and in predicting the log-likelihood using the testing data. As a result the deviance from the bootstrap is likely going to underestimate the deviance from cross validation. This results in tuning parameters that clutters the signal variables selected from the bootstrap with too many noise variables. As an improvement, the weighted deviance below is suggested in the bootstrap algorithm:

$$L(\alpha, \delta, \gamma) = - \sum_{i=1}^n w_i \{y_i \ln[\pi(g_i, z_i)] + (1 - y_i) \ln[1 - \pi(g_i, z_i)]\}.$$

where

$$w_i = \frac{1}{\sqrt{\pi_i(1 - \pi_i)}}.$$

The above normalized log-likelihood has the effect of assigning lower weights to samples with large variances, thus limiting its effect on model fit. Preliminary simulations using the above resulted in bootstrap variance estimate which has smaller bias. More simulations need to be carried out to study the effect of the above on sample size estimates. We also note that if this weighted deviance correction works, it would not only be applicable to the limited sample size setting, but could be useful to anyone bootstrapping a high dimensional dataset for whatever reason.

Also, for future consideration, to determine how sensitive our sample size method is to feature selection criteria, sample size estimates using both deviance and misclassification error should be compared. Another future direction of the sample size method would be to develop methods for regularized discriminant analysis. Corresponding errors-in-variables methods for regularized discriminant analysis would be needed as a pre-requisites to such

an extension.

In Chapter 3, a sparse linear discriminant function for classifying new entities into more than two classes was considered. A new method, known as the basis method, that generalizes binary linear discriminant analysis problem to multi-class was proposed. The methodology exploited the relationship between linear discriminant functions using basis vectors of the between class scatter and Fisher's linear discriminant solution. In particular, it was shown that the solution space spanned by binary linear discriminant analysis problem using orthonormal basis vector of the between class scatter matrix is the same as the solution space spanned by the original LDA discriminant vectors. We showed that the proposed method performed better and overcame the limitations of two existing multi-class linear discriminant functions which are pairwise combinations of binary linear discriminant problems. Simulation processes and real data analyses showed superior performance of our method.

One limitation of the basis method is that it obtains all $K - 1$ sparse linear discriminant vectors in a K class problem, which can be computationally disadvantageous for large K . As a future direction, it would be interesting to study the performance of the basis method for $q < (K - 1)$ direction vectors.

In Chapter 4, a framework for obtaining sparse solution vectors for many multivariate statistical problems in high dimension, low sample size was considered. The methodology capitalized on direction vectors spanning a lower subspace in many multivariate statistical problems and their connections with the generalized eigenvalue problem. In particular, it was observed that many multivariate statistical problems had a core idea of finding linear combinations of available variables to study some structure in the data and these weight vectors often times happened to be the generalized eigenvectors of the generalized eigenvalue problem. The traditional generalized eigenvector, which uses all available variables was made

sparse by imposing l_∞ constraint on the generalized eigenvalue problem. We showed that naively bounding with l_∞ results in a trivial solution vector, meaning that no variable is selected. Thus, to ensure that at least one variable is selected, we substituted the left term in the generalized eigenvalue equation with the nonsparse solution vector before imposing l_∞ constraint. This approach was termed the substitution method.

A demonstration of the use of the substitution method in linear discriminant analysis and canonical correlation analysis was given. The substitution method applied to these problems showed superior performance. While the method was demonstrated on these two problems, it is worth mentioning that it can be applied to obtain sparse solution vectors in several multivariate statistical problems with some slight modifications. In the future, other multivariate problems such as multivariate multiple regression, multivariate analysis of variance and principal component analysis would be studied.

CHAPTER 6

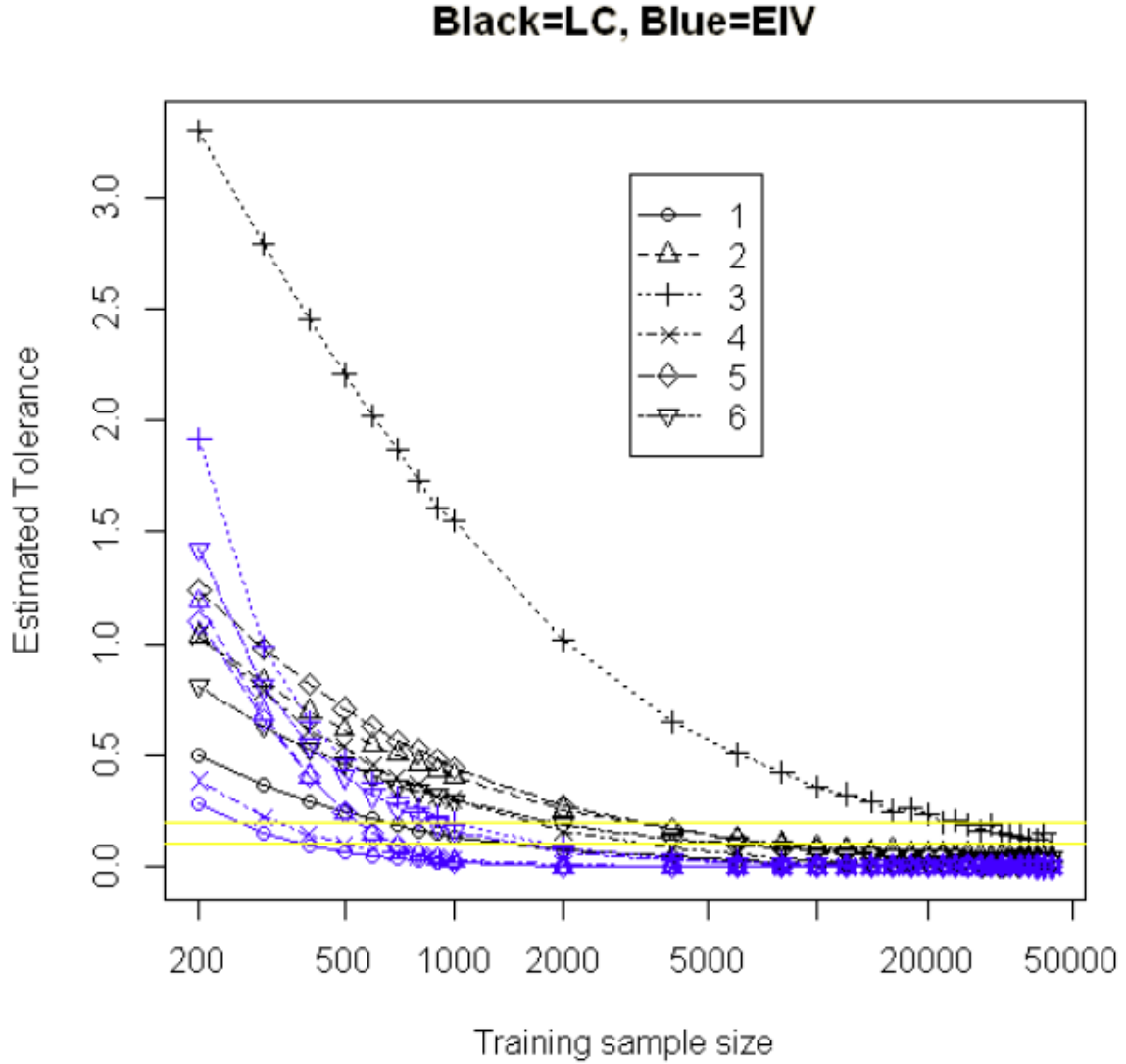
SUPPLEMENT

SUPPLEMENTAL MATERIAL TO SAMPLE SIZE METHOD FOR REGULARIZED HIGH DIMENSIONAL CLASSIFICATION

Section	Page Number
Supplementary Figures	128
Supplementary Tables	132
Algorithms	135
Simulation Details	138
Theoretical Discussions	143

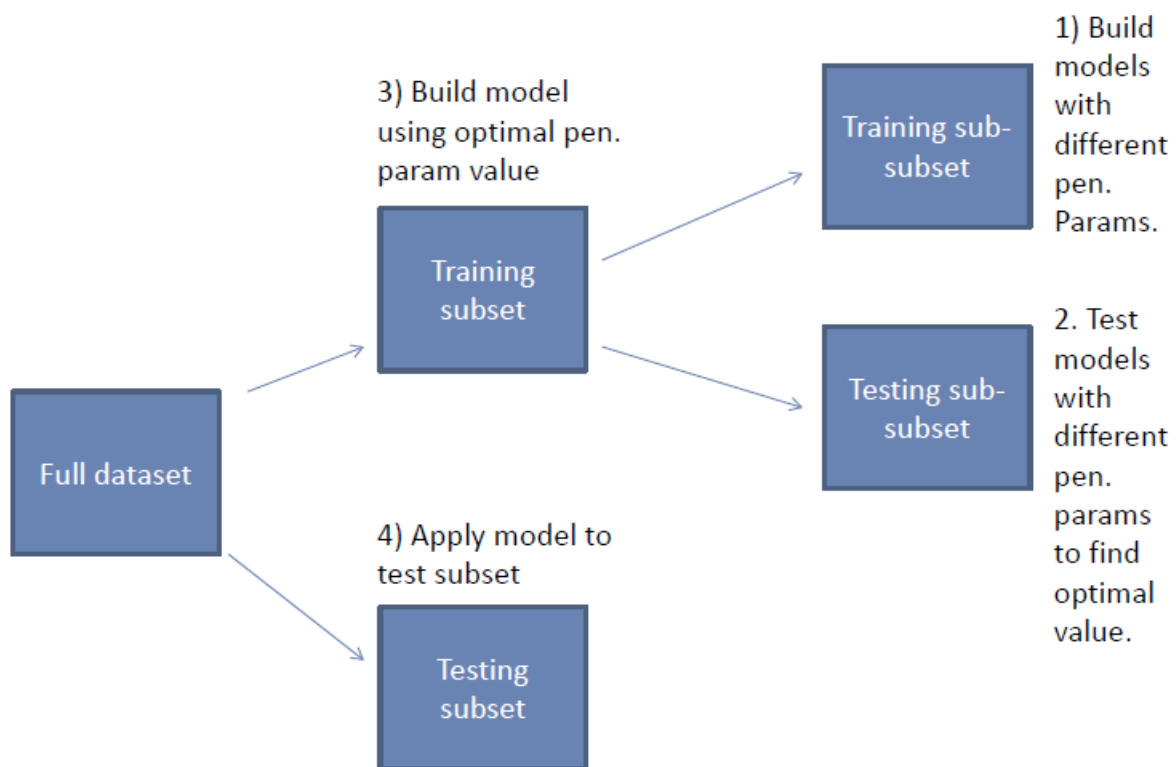
6.0.1 SUPPLEMENTARY FIGURES

Figure 6.0.1: Simulated datasets results. Dataset 1 is Identity covariance with slope $\beta_\infty = 3$. Dataset 2 is AR1 covariance with $\beta_\infty = 3$. Dataset 3 is AR1 covariance with $\beta_\infty = 4$. Dataset 4 is Identity covariance with $\beta_\infty = 4$. Dataset 5 is Compound Symmetric covariance with $\beta_\infty = 3$. Dataset 6 is Compound symmetric with $\beta_\infty = 4$. Identity covariance has one informative feature. CS and AR1 covariance cases have 9 informative features in 3 blocks of 3, with correlation parameter 0.7.



For the six simulated dataset, above are the estimated plots of $Y=\text{Tolerance}$ versus $x=\text{sample size}$ (log scale) for our method (EIV) and the Mukherjee method (LC). Yellow horizontal lines indicate the estimated tolerances of 0.10 and 0.20 in the table.

Figure 6.0.2: Nested cross-validation procedure



5) Standardize test subset scores to have mean 0 variance 1

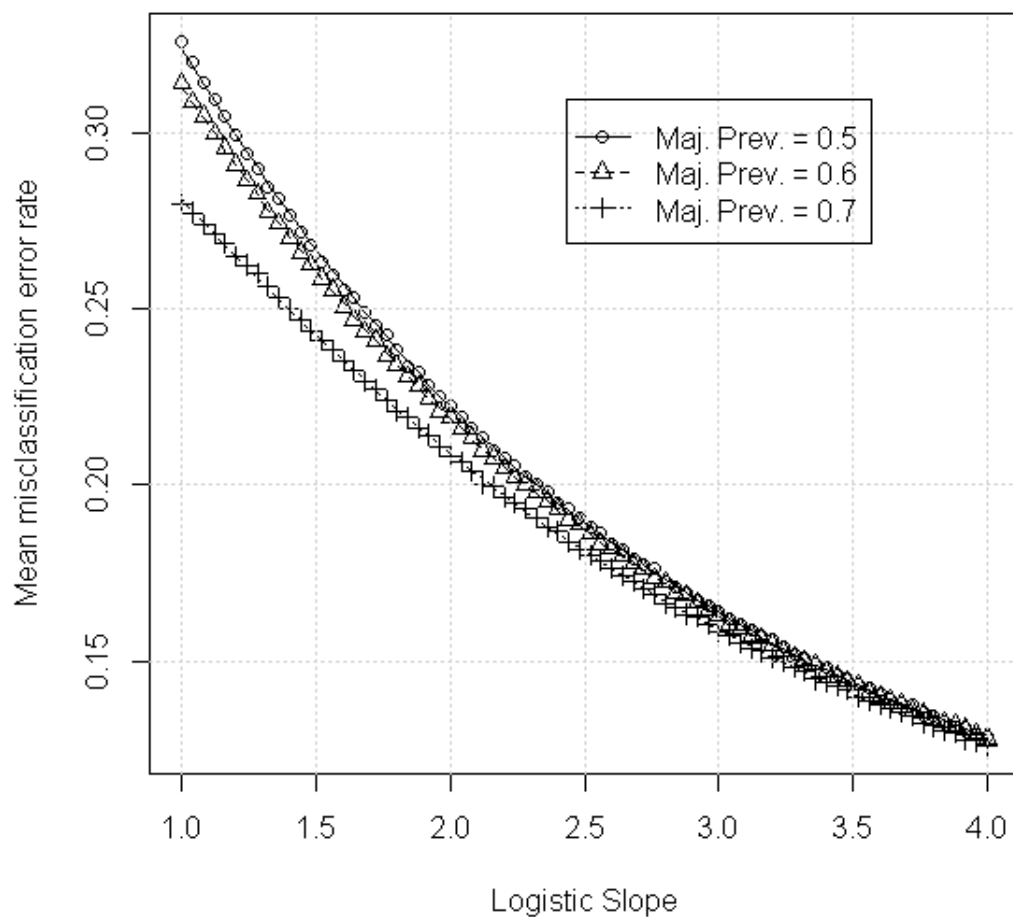
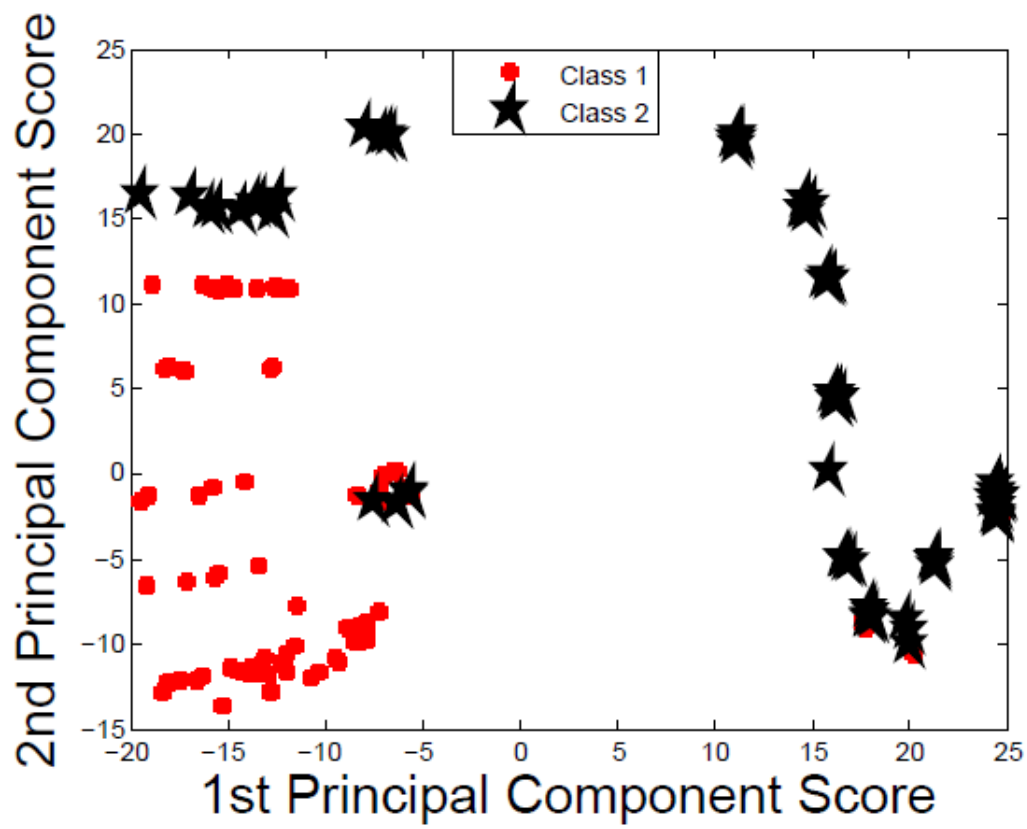


Figure 6.0.3: Converting between a logistic slope and the classification error rate. Plot of $x = \beta_{\infty}$ versus $y = \widehat{acc}_{\infty}$, for different population prevalence setting for the majority class. Assumes no clinical covariates present in model.

Figure 6.0.4: Principal component plot of the fly data, labeled by class.



6.0.2 SUPPLEMENTARY TABLES

Table 6.0.1: Mixture normal simulations results. Same as normal simulations but with the high dimensional data generated from a homoscedastic multivariate normal model.

n_{pilot}	Cov.	β_∞	mean $\hat{\beta}_\infty$	mean $\hat{\sigma}_n^2$	mean $\hat{\beta}_n$
300	AR1	2	1.96	0.1416	1.57
400	AR1	2	2.00	0.0965	1.71
300	AR1	3	2.87	0.0677	2.51
400	AR1	3	2.97	0.0479	2.69
300	AR1	4	3.78	0.0444	3.40
400	AR1	4	3.94	0.0317	3.63
300	AR1	5	4.70	0.0334	4.26
400	AR1	5	4.95	0.0238	4.58
300	Identity	2	2.00	0.0486	1.85
400	Identity	2	1.99	0.0302	1.90
300	Identity	3	2.94	0.0242	2.80
400	Identity	3	2.95	0.0172	2.85
300	Identity	4	3.92	0.0178	3.74
400	Identity	4	3.93	0.0134	3.80
300	Identity	5	4.89	0.0141	4.68
400	Identity	5	4.90	0.0107	4.74

Table 6.0.2: Comparison of LC and EIV asymptotic slope estimates from six Monte Carlo simulations. LC is the learning-curve based method of Mukherjee et al. (2003). EIV is the errors-in-variables method presented in our paper. β_∞ is the true value of the slope used to generate the simulated data, and Err_∞ is the corresponding misclassification rate. $\hat{\beta}_\infty$ are the estimated asymptotic slopes. “LC %” and “EIV %” are the errors of the estimated slopes (or error rates) as a proportion of the true slopes. Numbers in parentheses on the first row correspond to numbers on Figure 1.

Covariance	AR1(2)	CS(5)	Identity(1)	AR1(3)	CS(6)	Identity(4)
β_∞	3	3	3	4	4	4
LC $\hat{\beta}_\infty$	3.04	3.32	3.24	6.24	3.36	5.00
EIV $\hat{\beta}_\infty$	2.93	2.95	2.82	4.30	3.86	3.54
LC %	1%	11%	8%	56%	-16%	25%
EIV %	-2%	-2%	-6%	7.5%	-3.5%	-11%
Err_∞	0.164	0.164	0.164	0.129	0.129	0.129
LC $\widehat{Err}_\infty$	0.162	0.151	0.154	0.086	0.150	0.105
EIV $\widehat{Err}_\infty$	0.167	0.165	0.172	0.121	0.132	0.143
LC %	-1%	-8%	-6%	-33%	16%	-19%
EIV %	2%	1%	5%	-6%	2%	11%

Table 6.0.3: Evaluation of the sample size estimates from identity covariance. The number in the pilot dataset is 300. $\beta_\infty = 3$ with one informative feature, an identity covariance matrix, and $p = 500$ total features. Estimates evaluated using 400 Monte Carlo simulations with the estimated sample size. The mean tolerance from the 400 simulations, and the proportion of the 400 within the specified tolerance are given in the rightmost two columns.

t_{target}	$\hat{n}$	n for MC	Mean MC tol	% of MC within tol
0.10	312	312	0.07	80%
0.20	236	236	0.12	86%
0.30	195	195	0.14	90%
0.40	167	167	0.19	90%
0.50	145	145	0.21	93%
0.60	128	128	0.25	93%
0.70	113	113	0.30	94%

Table 6.0.4: Evaluation of the sample size estimates from CS covariance. The number in the pilot dataset is 400. $\beta_\infty = 4$ with nine informative features, 3 blocks of size 3 in a compound symmetric covariance matrix with parameter 0.7. Estimates evaluated using 400 Monte Carlo simulations with the estimated sample size. The mean tolerance from the 400 simulations, and the proportion of the 400 within the specified tolerance are given in the rightmost two columns.

t_{target}	$\hat{n}$	n for MC	Mean MC tol	% of MC within tol
0.10	1490	1490	0.11	49%
0.20	872	872	0.20	58%
0.30	640	640	0.29	60%
0.40	514	514	0.39	59%
0.50	435	435	0.46	68%
0.60	380	380	0.55	67%
0.70	339	339	0.62	69%

Table 6.0.5: Resampling studies. Dataset is the dataset used for resampling. Rep is the replication number of 5 independent random subsamples (without replacement) of size n_{Pilot} . nFull is the size of the full dataset. Classes for the Shedden dataset were Alive versus Dead. Classes for Rosenwald dataset were Germinal-Center B-Cell lymphoma type versus all others. $err(nFull)$ is estimated from 200 (50 for Shedden) random cross-validation estimations on the full dataset using different partitions each time, and this serves as the gold standard error rate for $nFull$. $\widehat{err}(nFull)$ is the estimated error rate for the full dataset based on Mukherjee et al.’s method or our method. Similarly, $\widehat{err}(\infty)$ is the asymptotic error rate based on the Mukherjee et al. method or our method. For the Shedden dataset, used conditional score EIV and SIMEX (format is cond. score/SIMEX); for Rosenwald dataset, we used quadratic SIMEX EIV because the variance estimates greatly exceeded our bound formula (Section 2.5).

Dataset	Rep	nPilot	nFull	$err(nFull)$	LC method		Our Method		nFull err %	
					$\widehat{err}(nFull)$	$\widehat{err}(\infty)$	$\widehat{err}(nFull)$	$\widehat{err}(\infty)$	Mukh.	Our
Rosenwald	1	100	240	0.1129	0.0855	0.0729	0.1344	0.1135	-25%	19%
Rosenwald	2	100	240	0.1129	0.0611	0.0435	0.1078	0.0933	-46%	-5%
Rosenwald	3	100	240	0.1129	0.0298	0.0089	0.0771	0.0691	-74%	-32%
Rosenwald	4	100	240	0.1129	0.1443	0.1270	0.1396	0.1379	28%	24%
Rosenwald	5	100	240	0.1129	0.0682	0.0480	0.0864	0.0783	-40%	-23%
Average				0.1129	0.0778	0.0601	0.1091	0.0984	-31%	-3%
Shedden	1	200	443	0.4207	0.4638	0.4634	0.4347/0.4548	0.4347/0.4548	10%	3%/8%
Shedden	2	200	443	0.4207	0.4496	0.4481	0.4154/0.4473	0.4151/0.4473	7%	-2%/6%
Shedden	3	200	443	0.4207	0.4300	0.4258	0.2778/0.3885	0.2778/0.3885	9%	-34%/-8%
Shedden	4	200	443	0.4207	0.4166	0.4126	0.3550/0.3900	0.3550/0.3900	-1%	-16%/-7%
Shedden	5	200	443	0.4207	0.4159	0.4117	0.2907/0.3447	0.2894/0.3447	-1%	-31%/-18%
Average				0.4207	0.4406	0.4385	0.3548/0.4051	0.3544/0.4051	5%	-16%/-4%

6.0.3 ALGORITHMS

6.0.3.1 NESTED, SCALED, CROSS-VALIDATION ALGORITHM

For k -fold cross-validation.

1. Randomly sort the pilot set and partition the n samples into k subsets, each of size $\text{floor}(n/k)$ or $\text{floor}(n/k) + 1$, so that they are equal or almost equal in size.
2. Leave out the first subset.
 - (a) Partition the remaining $n(k-1)/k$ samples again into k sub-subsets that are again equal or almost equal in size.
 - (b) Leave out the first sub-subset.
 - i. Develop lasso predictors on the remaining samples (roughly $n(k-1)^2/k^2$) using all values of the tuning parameter (see `cv.glmnet` documentation).
 - ii. For each lasso predictor, calculate a model selection criterion. We recommend the mean error rate applied to the left-out sub-subset, that is (making the obvious adjustments as needed if the left out subsample sizes are not exactly n/k^2 in size):

$$C_i(\lambda) = \frac{1}{n/k^2} \sum_{j=1}^{n/k^2} 1_{\{\hat{y}_j = y_j\}}$$

- (c) Cycle through, leaving out each sub-subset in turn and agglomerate the model selection criteria for each possible value of the penalty parameter λ , for example,
$$C_{\text{agglomerated}}(\lambda) = \sum_{i=1}^k C_i(\lambda).$$
 - (d) Pick the λ that optimizes the model selection criterion, $\hat{\lambda}_k = \min_{\lambda} C_{\text{agglomerated}}(\lambda)$.
3. Using the $\hat{\lambda}_k$ chosen from the inner cross-validation loop, develop a lasso predictor on the $n(k-1)/k$ samples in the outer training set.

4. Apply the developed predictor to the left-out samples in the outer cross-validation, i.e., the outer validation set, creating continuous estimated scores W_{ij} .
5. Cycle through until each sample has a left-out prediction score from the outer cross-validation loop.
6. Center and scale each cross-validated batch to have mean zero and variance 1. For 5-fold cross-validation, there would be 5 batches.
7. Perform logistic regression of Y_i on W_{ij} to obtain $\hat{\beta}_{j,cv}$.
8. Finally, estimate the prediction error variance $\hat{\sigma}_n^2$ as described elsewhere in the paper, and obtain the final estimate $\hat{\beta}_n = \hat{\beta}_{j,cv} / \sqrt{1 + \hat{\sigma}_n^2}$.

6.0.3.2 NESTED, CASE-CROSS-VALIDATED LEAVE-ONE-OUT BOOTSTRAP ALGORITHM

1. Randomly draw a sample with replacement of size n from the pilot dataset. Call this the bootstrap sample b .
2. Randomly assign numbers to each unique sample in the bootstrap sample, say $j = 1, \dots, k \leq n$.
3. Sort the samples based on the j 's from lowest to highest.
4. As close as possible, divide the samples into k equal-sized subsets of size $n(k-1)/k$, in such a way that no sample $j = j_0$ appears in more than one subset. If a sample appears in more than one subset, then move the dividing line to the closest placement that results in disjoint sets with no overlapping samples.
5. Leave out the first subset.
 - (a) Repeat the process described above to divide the remaining samples into k sub-subsets, each of size approximately n/k^2 , with no overlap.

(b) Leave out the first sub-subset.

- i. Develop lasso predictors to the remaining samples using all values of the tuning parameter (see `cv.glmnet` documentation).
- ii. For each lasso predictor, calculate a model selection criterion. We recommend the error rate applied to the left-out sub-subset:

$$C_{i,lasso}(\lambda) = \frac{1}{n/k^2} \sum_{j=1}^{n/k^2} 1_{\{\hat{y}_j = y_j\}}$$

Since some samples in the left-out sub-subset may be repeated, weighting samples by the inverse of their cardinality is an alternative.

- iii. Cycle through leaving out each sub-subset in turn and agglomerate the model selection criteria for each possible value of the penalty parameter λ .

(c) Pick the λ that optimizes the model selection criterion in the inner case cross-validated loop, call it $\hat{\lambda}_k = \min_{\lambda} \sum_{i=1}^k C_{i,lasso}(\lambda)$.

(d) Apply the lasso with penalty $\hat{\lambda}_k$ to the full bootstrap sample set to obtain estimates of $(\hat{\alpha}, \hat{\delta}, \hat{\gamma})$. Apply $\hat{\gamma}$ to every sample that was omitted from the bootstrap sample b to obtain a preliminary set of predicted classification scores.

(e) Center and scale the scores from the previous step to have mean zero and variance one.

6. Repeat the whole process for $b = 1, \dots, 35$, obtaining multiple prediction scores $W_{ij1}, \dots, W_{ijb_i}$ for each sample, with $b_i \leq 35$.

7. $\hat{\sigma}_i^2 = \frac{1}{b_i-1} \sum_{b=1}^{b_i} (W_{ijb} - \overline{W}_{ij\cdot})^2$.

8. Finally, $\hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_i^2$.

6.0.4 SIMULATION DETAILS

6.0.4.1 ADJUSTMENT OF SLOPE FOR MEASUREMENT ERROR IN SIMULATIONS

For the Monte Carlo simulations, an issue is that we don't a priori know what the prediction error variance should be for the targeted sample size. Yet this variance is needed to unbiasedly estimate the tolerance associated with a training set size, denoted $Tol(n) = \beta_\infty - \beta_n$. We therefore fit the curve $y = f \times n^g$ to a plot with $x=n$ =training sample size versus y = estimated measurement error variance using least squares. Then $\hat{y} = \hat{f} \times n^{\hat{g}}$ provides an estimate of the measurement error variance associated with any sample size. Then $\hat{\beta}_n = \hat{\beta}_n^{naive} / \sqrt{1 + \hat{y}}$. This is used for the calculations in the tables evaluating the adequacy of the sample size estimates.

6.0.4.2 GENERATING LOGISTIC DATA: NORMAL

Generate $g_i \sim MVN(0, \Sigma)$, and define L^* as a vector that will be proportional to L . Calculate the quantity $L^{*\prime} \Sigma L^*$, and define $L = L^* / (L^{*\prime} \Sigma L^*)$. Calculate the asymptotic classification scores $X_i = L' g_i$. Then calculate the probabilities of classification from the logistic regression model, so that Y_i is Bernoulli with success probability $(1 + \text{Exp}[-\alpha - \beta_\infty X_i])^{-1}$.

The result is a set of data (Y_i, g_i) from a logistic regression model with asymptotic variance β_∞ . Marginally, the data are multivariate normal.

6.0.4.3 GENERATING LOGISTIC DATA: MIXTURE NORMAL

Now we consider generating logistic regression data from a mixture normal distribution. The mixture will consist of two normal populations with different means and identical covariance structures. Individuals with $Y_i = 1$ will have mean μ , and individuals with $Y_i = 0$ will have mean $-\mu$. Marginally, the data will look like a high dimensional mixture normal.

In particular, let

$$g_i \sim \begin{cases} Normal(\mu, \Sigma) & y_i = 1 \\ Normal(-\mu, \Sigma) & y_i = 0 \end{cases}$$

Assume marginally $P(Y_i = 1) = \pi$. Then given observed data g_j ,

$$Prob(y_j = 1|g_j) = \frac{Exp[\alpha + \gamma'g_j]}{1 + Exp[\alpha + \gamma'g_j]}$$

where

$$\begin{aligned} \alpha &= \log\left(\frac{\pi}{1-\pi}\right) \\ \gamma' &= 2\mu'\Sigma^{-1}, \gamma = 2\Sigma^{-1}\mu, \end{aligned}$$

and where μ' is $1 \times p$ and Σ^{-1} is a nonsingular $p \times p$ matrix (Efron, 1975). We calculate the marginal variance:

$$Var_{\mathcal{P}}(\gamma'g_j) = \gamma'Var_{\mathcal{P}}(g_j)\gamma.$$

Now

$$\begin{aligned} Var_{\mathcal{P}}(g_j) &= E[Var(g_j|Y_j)] + Var(E[g_j|Y_j]) \\ &= E[\Sigma] + Var(\mu \times (-1 + 2 \times 1_{\{Y_j=1\}})) \\ &= \Sigma + \pi(\mu - \tilde{\mu})(\mu - \tilde{\mu})' + (1 - \pi)(-\mu - \tilde{\mu})(-\mu - \tilde{\mu})', \end{aligned}$$

where $\tilde{\mu} = \pi\mu - (1 - \pi)\mu = \mu(2\pi - 1)$. Noting that $\mu - \tilde{\mu} = 2\mu(1 - \pi)$, $(\mu - \tilde{\mu})(\mu - \tilde{\mu})' = 4(1 - \pi)^2\mu\mu'$, and $\mu + \tilde{\mu} = 2\pi\mu$, yields $(\mu + \tilde{\mu})(\mu + \tilde{\mu})' = 4\pi^2\mu\mu'$. Yielding the simplified version,

$$\begin{aligned} Var_{\mathcal{P}}(g_j) &= \Sigma + (4\pi(1 - \pi)^2 + 4(1 - \pi)\pi^2)\mu\mu' \\ &= \Sigma + 4\pi(1 - \pi)\mu\mu'. \end{aligned}$$

So finally,

$$\begin{aligned} Var_{\mathcal{P}}(\gamma' g_j) &= 4\mu' \Sigma^{-1} \{\Sigma + 4\pi(1 - \pi)\mu\mu'\} \Sigma^{-1} \mu \\ &= 4\mu' \Sigma^{-1} \mu + 16\pi(1 - \pi)(\mu' \Sigma^{-1} \mu)^2. \end{aligned}$$

We want to standardize so that these score variances are equal to β_{∞}^2 . Then, when scores are standardized to have unit variance, the corresponding slope will be β_{∞} . If we let $\pi = 1/2$, then this results in the quadratic equation

$$(\mu' \Sigma^{-1} \mu)^2 + (\mu' \Sigma^{-1} \mu) - \beta_{\infty}^2/4 = 0.$$

The solution is

$$\mu' \Sigma^{-1} \mu = \frac{-1 + \sqrt{1 + \beta_{\infty}^2}}{2}.$$

For example, for $\beta_{\infty} = 2, 3, 4$, we have $\mu' \Sigma^{-1} \mu \approx 0.618, 1.081, 1.562$.

If $\Sigma = I$, then we can calculate the corresponding mean vectors μ . If there is 1 feature related to the classification, so that $\mu = (c, 0, \dots, 0)'$. then $c = \sqrt{\mu' \Sigma^{-1} \mu} = 0.786, 1.040, 1.250$ for $\beta_{\infty} = 2, 3, 4$. If there are k features, each with effect size c , then $c = 0.786/\sqrt{k}, 1.040/\sqrt{k}, 1.250/\sqrt{k}$. If $\Sigma = \Sigma_{sub} \otimes I_3$, where

$$\Sigma_{sub} = \begin{pmatrix} 1.0 & 0.7 & 0.7 \\ 0.7 & 1.0 & 0.7 \\ 0.7 & 0.7 & 1.0 \end{pmatrix}.$$

Then $\Sigma^{-1} = \Sigma^* \otimes I_3$ where

$$\Sigma^* = \begin{pmatrix} 2.36111 & -0.972222 & -0.972222 \\ -0.972222 & 2.36111 & -0.972222 \\ -0.972222 & -0.972222 & 2.36111 \end{pmatrix}.$$

If $\mu = (c, \dots, c)$ has length 9, then $\mu' \Sigma^{-1} \mu = c^2(9 * 2.36111 + 18 * (-0.972222)) = 3.75c^2$. This leads to

$$\begin{aligned} 3.75c^2 &= \frac{-1 + \sqrt{1 + \beta_\infty^2}}{2} \\ c &= \sqrt{\frac{-1 + \sqrt{1 + \beta_\infty^2}}{2 \times 3.75}}. \end{aligned}$$

Similarly, if

$$\Sigma_{sub2} = \begin{pmatrix} 1.0 & 0.7 & 0.49 \\ 0.7 & 1.0 & 0.7 \\ 0.49 & 0.7 & 1.0 \end{pmatrix}$$

then,

$$\Sigma_2^* = \begin{pmatrix} 1.96078 & -1.37255 & 0 \\ -1.37255 & 2.92157 & -1.37255 \\ 0 & -1.37255 & 1.96078 \end{pmatrix}$$

and $\mu' \Sigma^{-1} \mu = c^2(3 * 2.92157 + 6 * 1.96078 - 12 * 1.37255) = 4.05879c^2$ and

$$c = \sqrt{\frac{-1 + \sqrt{1 + \beta_\infty^2}}{2 \times 4.05879}}.$$

Table 6.0.6: Effect size for different covariance structures and number of informative features

β_∞	Σ structure	No. of informative features	Effect size	Σ Diagonal	Σ off diagonal
2	Identity	1	0.786	1	0
3	Identity	1	1.040	1	0
4	Identity	1	1.250	1	0
5	Identity	1	1.432	1	0
2	CS	9	0.4060	1	0.7
3	CS	9	0.5369	1	0.7
4	CS	9	0.6453	1	0.7
5	CS	9	0.7393	1	0.7
2	AR1	9	0.3902	1	$0.7^{ r-c }$
3	AR1	9	0.5161	1	$0.7^{ r-c }$
4	AR1	9	0.6203	1	$0.7^{ r-c }$
5	AR1	9	0.7106	1	$0.7^{ r-c }$

6.0.5 THEORETICAL DISCUSSIONS

6.0.5.1 REGULARITY CONDITIONS FOR $|E_n[\beta_j]| < |\beta_\infty|$ AND $\lim_{n \rightarrow \infty} Tol(n) = 0$

The value $E_n[\beta_j]$ is the mean of the maximum likelihood estimator of the slope parameter β using the incorrect likelihood function (with the W_{ij} as the univariate covariate instead of X_i). By Theorem 5.1 of Stefanski and Carroll (1985), the estimator will be consistent under the following conditions (restated here for reference):

1. The function

$$G_n(\beta) = \frac{1}{n} \sum_{i=1}^n [(1 + \exp[-\beta w_{ij}])^{-1} \times \text{Ln} \{(1 + \exp[-\beta w_{ij}])^{-1}\} + (1 + \exp[\beta w_{ij}])^{-1} \text{Ln} \{(1 + \exp[\beta w_{ij}])^{-1}\}]$$

converges pointwise to a function $G(\beta)$ possessing a unique maximum at β_∞ .

- 2.

$$\sum_{i=1}^n (||w_i||)^2 = o(n^2)$$

- 3.

$$E(U_{ij}^2) < \infty.$$

6.0.5.2 CONDITIONAL SCORES

Conditional scores is a fully consistent method of eliminating bias in measurement error models. In this method, one conditions the response vector on a sufficient statistics to eliminate nuisance parameters arising from the error prone variables, resulting in consistent estimators. Consider the logistic regression model:

$$\log \left(\frac{Pr(y_i = 1|x_i, \mathbf{z}_i)}{Pr(y_i = 0|x_i, \mathbf{z}_i)} \right) = \alpha + \beta_x x_i + \beta_z^T \mathbf{z}_i, \quad i = 1, \dots, n$$

where $\boldsymbol{\beta} = (\beta_x, \boldsymbol{\beta}_z^T)^T$ is the vector of regression parameters of interest. Here $X = [x_1, \dots, x_n]^T$ is $n \times 1$ covariate measured with error and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_q]$ is $n \times q$ error-free covariates. Assume additive measurement error model from equation (1.2.1) and let σ_u^2 be the variance of the error term. The conditional distribution of $W_i = w$ given $X_i = x$ is $\sim N(x, \sigma_u^2)$, and has density

$$f_{W_i}(w|X_i) \propto \exp\left(-\frac{(w-x)^2\sigma^{-2}}{2}\right) = \exp\left(\frac{-(w^2 - wx + x^2)}{2\sigma^2}\right) \quad (6.0.1)$$

and the density of $\mathbf{y}$ in exponential family distribution form is

$$f_Y(y|X = x, \mathbf{Z} = \mathbf{z}) = \exp\left\{\frac{y\eta - D(\eta)}{\phi} + c(y, \phi)\right\} \quad (6.0.2)$$

where $\eta = \alpha + \beta_x x + \boldsymbol{\beta}_z^T \mathbf{z}$, $D(\eta) = -\ln\left(1 - \frac{1}{1+\exp(-\eta)}\right)$, $c(y, \phi) = 0$ and $\phi = 1$. Combining (6.0.1) and (6.0.2), the joint density of W and y is

$$\begin{aligned} f(W, y) &= f_{W_i}(w_i|X_i)f_{Y_i}(y|(X_i, \mathbf{Z}_i)) \\ &\propto \exp\left\{(-1/2\sigma^2)(w_i^2 - 2w_ix_i + x_i^2)\right\} \cdot \exp\left\{\frac{y_i\eta - D(\eta)}{\phi} + c(y_i, \phi)\right\} \\ &\propto \exp\left(-\frac{w_i^2}{2\sigma^2} + \frac{x_i}{\sigma^2}(w_i + y_i\sigma^2\beta_x) + c(y_i, \phi)\right). \end{aligned}$$

Therefore,

$$\Delta_i = W_i + y_i\sigma^2\beta_x \quad (6.0.3)$$

is a sufficient statistics for X (because X is unknown and all other parameters are known so by factorization theorem, this is true). We next obtain the joint density of the response vector and the sufficient statistic Δ and hence obtain the conditional loglikelihood for all the subjects.

Let $\delta = w + y\sigma^2\beta_x \Rightarrow w = \delta - y\sigma^2\beta_x$. Then the transformation from (y, W) into (y, Δ) has Jacobian 1. Therefore, the joint density of (y, Δ) is the product of the density of y and

the density of W which is

$$\begin{aligned}
f_{(Y,\Delta)(y,\delta)} &= f_{Y_i}(y|(\Delta_i, \mathbf{Z}_i)) \cdot f_{\Delta_i}(\delta_i|X_i) \\
&\propto \exp \left\{ \frac{y\eta - D(\eta)}{\phi} + c(y, \phi) \right\} \cdot \exp \left\{ (-1/2\sigma^2)((\delta - y\sigma^2\beta_x)^2 - (\delta - y\sigma^2\beta_x)x) \right\} \\
&\propto \exp \left(y(\alpha + \beta_z^T \mathbf{z} + \beta_x \delta + c(y, \phi)) - D(\eta) - \frac{y^2 \beta_x^2 \sigma^2}{2} + \delta^2 \sigma^{-2} \right) \\
&\propto \exp \left(y(\alpha + \beta_z^T \mathbf{z} + \beta_x \delta) + c(y, \phi) - \frac{y^2 \beta_x^2 \sigma^2}{2} \right).
\end{aligned}$$

The conditional distribution of Y given $\Delta = \delta$ is

$$\begin{aligned}
f_{Y|(\Delta)}(y|\delta) &= \frac{f(y, \delta)}{f(\delta)} \\
&= \frac{f(y, \delta)}{\sum_y f(y, \delta)}, \quad \text{summing out } y \text{ gives the density of } \delta. \\
&= \propto \exp \left(y(\alpha + \beta_z^T \mathbf{z} + \beta_x \delta) - \frac{y^2 \beta_x^2 \sigma^2}{2} - \log \sum_y \left[\exp \left(y(\alpha + \beta_z^T \mathbf{z} + \beta_x \delta) - \frac{y^2 \beta_x^2 \sigma^2}{2} \right) \right] \right) \\
&= \exp (y\eta_* - D_*(\eta_*, \beta_x, \beta_z, \sigma) + c_*(y, \beta_x, \sigma^2))
\end{aligned}$$

where $\eta_* = \alpha + \beta_z^T \mathbf{z} + \beta_x \delta$, $c_*(y, \beta_x, \sigma^2) = -\frac{y^2 \beta_x^2 \sigma^2}{2}$,

$D_*(\eta_*, \beta, \sigma) = \log [1 + \exp (\alpha + \beta_z^T \mathbf{z} + \beta_x \delta - (1/2)\sigma^2 \beta_x^2)]$. Therefore, the conditional log-likelihood contributed from all the subjects is

$$\begin{aligned}
l_c = \sum_{i=1}^n L_{ci} &= \sum_{i=1}^n \log(f_{Y_i|(\Delta)}(y_i|\delta_i)) \\
&= \sum_{i=1}^n \{y_i(\alpha + \beta_z^T \mathbf{z}_i + \beta_x \delta_i)\} \\
&\quad - \sum_{i=1}^n \log [1 + \exp (\alpha + \beta_z^T \mathbf{z}_i + \beta_x \delta_i - (1/2)\sigma^2 \beta_x^2)] \\
&\quad - \sum_{i=1}^n (1/2) \{y_i^2 \sigma^2 \beta_x^2\}. \tag{6.0.4}
\end{aligned}$$

The conditional log-likelihood in (6.0.4) is the likelihood function to be maximized in order to obtain the parameter coefficients associated with the measurement error model. We derived it by conditioning the response vector $\mathbf{y}$ on a sufficient statistic Δ for the unobserved classification scores X . Now, differentiating equation (6.0.4) with respect to the parameters and

setting to zero, we obtain the conditional score function

$$\Psi_{Cond}(Y_i, \mathbf{Z}_i, \Delta_i, \boldsymbol{\theta}) = [Y_i - E(Y_i | \Delta_i, \mathbf{Z}_i)] \begin{pmatrix} 1 \\ t(\delta)_i \\ \mathbf{Z}_i \end{pmatrix}. \quad (6.0.5)$$

where

$$E(Y_i | \Delta_i, \mathbf{Z}_i) = H = \frac{\exp\left(\alpha + \beta_z Z + \beta_x \Delta - \frac{\sigma^2 \beta_x^2}{2}\right)}{1 + \exp\left(\alpha + \beta_z Z + \beta_x \Delta - \frac{\sigma^2 \beta_x^2}{2}\right)}. \quad (6.0.6)$$

Hanfelt and Liang (1997) considered a slightly different estimator $t(\delta)$ from the one in equation (6.0.3) by defining the sufficient statistic

$$\Delta_{iHL} = \Delta_i + (H_i - 1)\sigma^2\beta. \quad (6.0.7)$$

Now, if

$$\begin{aligned} \boldsymbol{\Omega}_1 &= n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \Psi_{cond}(\mathbf{y}_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) \\ \boldsymbol{\Omega}_2 &= Cov(\Psi_{cond}(\mathbf{y}_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) | \Delta). \end{aligned}$$

and $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_x^T, \hat{\boldsymbol{\beta}}_z^T)^T$ are solutions to $\sum_{i=1}^n \Psi_{cond}(\mathbf{y}_i, W_i, \mathbf{Z}_i, \boldsymbol{\theta}) = 0$, it can be shown (Carroll et. al. (2006)) that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow N(0, \boldsymbol{\Omega}_1^{-1} \boldsymbol{\Omega}_2 \boldsymbol{\Omega}_1^{-1T})$.

If we assume $t(\delta) = \delta$, then the diagonal entries of Ω_1 are

$$\begin{aligned}
n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \alpha} \Psi_{cond}(y_i, w_i, z_i, \theta) &= n^{-1} \sum_{i=1}^n \frac{\exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)}{\left(1 + \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)\right)^2} \\
&= n^{-1} \sum_{i=1}^n H_i(1 - H_i) \\
n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \beta_z} \Psi_{cond}(y_i, w_i, z_i, \theta) &= n^{-1} \sum_{i=1}^n \frac{z_i^2 \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)}{\left(1 + \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)\right)^2} \\
&= n^{-1} \sum_{i=1}^n z_i^2 H_i(1 - H_i) \\
n^{-1} \sum_{i=1}^n \frac{\partial}{\partial \beta_x} \Psi_{cond}(y_i, w_i, z_i, \theta) &= n^{-1} \sum_{i=1}^n y_i^2 \sigma^2 - n^{-1} \sum_{i=1}^n \frac{y_i \sigma^2}{1 + \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)} \\
&\quad - n^{-1} \sum_{i=1}^n \frac{\delta_i(\delta_i + \sigma^2 \beta_x(y_i - 1)) \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)}{\left(1 + \exp\left(-\alpha - \beta_z^T z_i - \beta_x^T \delta_i + \frac{\sigma^2 \beta_x^T \beta_x}{2}\right)\right)^2} \\
&= n^{-1} \sum_{i=1}^n \{y_i \sigma^2(1 - H_i) - \delta_i(\delta_i + \sigma^2 \beta_x(y_i - 1))H_i(1 - H_i)\}.
\end{aligned}$$

On the other hand, if we assume $t(\delta) = \delta_{HL}$, then the covariance formula simplifies since

$$\Omega_2 = \Omega_1 + G = \sum_{i=1}^n H_i(1 - H_i)(1, \Delta_{iHL}^T)^T(1, \Delta_{iHL}^T),$$

where G contains 0's except for the lower $p \times p$ submatrix given by

$\sum_{i=1}^n H_i(1 - H_i) \{\sigma^2 + H_i(1 - H_i)\sigma^2 \beta_x \beta_x^T \sigma^2\}$. The covariance estimates, Ω_2 may be estimated empirically via sandwich method. In our simulations, we found $t(\delta) = \delta_{HL}$ to be more stable for small sample sizes than $t(\delta) = \delta$. We used Newton Raphson algorithm to estimate these parameters.

BIBLIOGRAPHY

- Ahn, J. and Marron, J. S. (2010). The maximal data piling direction for discrimination. *Biometrika*, 97(1):254–259.
- Bi, X., Rexer, B., Arteaga, C. L., Guo, M., and Mahadevan-Jansen, A. (2014). Evaluating her2 amplification status and acquired drug resistance in breast cancer cells using raman spectroscopy. *J Biomed Opt*, 19.
- Bickel, P. J. and Levina, E. (2004). Some theory for fisher’s linear discriminant function, naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.
- Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, (1):199–227.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Buhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, New York.
- Cadima, J. and Jolliffe, I. T. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22:203–212.
- Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *The Annals of Statistics*, 35(6):2313–2351.

- Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, 2nd edition.
- Chalise, P. and Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics and Data Analysis*, 56:245–254.
- Clemmensen, L., Hastie, T., Witten, D., and Ersbll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.
- Cook, J. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.
- Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, New York.
- de Valpine, P., Bitter, H., Brown, M., and Heller, J. (2009). A simulation-approximation approach to sample size planning for high-dimensional classification studies. *Biostatistics*, 10:424–435.
- Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593.
- Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Lalo, D., Gall, C. L., Schaeffer, B., Crom, S. L., Guedj, M., and Jaffrzic, F. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683.

- Dobbin, K. K. and Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional dna microarray data. *Biostatistics*, 8(1):101–117.
- Dobbin, K. K., Zhao, Y., and Simon, R. M. (2008). How large a training set is needed to develop a classifier for microarray data. *Clinical Cancer Research*, pages 108–114.
- Donoho, D. L. (2000). Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality.
- Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley, New York.
- Dudoit, S., Fridlyand, J., and Terence, P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.
- Frazee, A., Langmead, B., and Leek, J. (2011). Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 12(449).
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011a). The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479.

Graveley, B., Brooks, A., Carlson, J., Duff, M., Landolin, J., Yang, L., Artieri, C., van Baren, M., Boley, N., Booth, B., Brown, J., Cherbas, L., Davis, C., Dobin, A., Li, R., Lin, W., Malone, J., Mattiuzzo, N., Miller, D., Sturgill, D., Tuch, B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J., Wan, K., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P., Brenner, S., Brent, M., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Oliver, B., and Celniker, S. (2011b). The developmental transcriptome of *drosophila melanogaster*. *Nature*, 471:473–479.

Hanash, S., Baik, C., and Kallioniemi, O. (2011). Emerging molecular biomarkers – blood-based strategies to detect and monitor cancer. *Nature Reviews Clinical Oncology*, 8:142–150.

Hanfelt, J. J. and Liang, K. Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika*, 82(3):pp. 461–477.

Hanfelt, J. J. and Liang, K. Y. (1997). Approximate likelihoods for generalized linear errors-in-variables models. *Journal of the Royal Statistical Society. Series B*, 59:627–637.

- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression. *Technometrics*, pages 152–177.
- Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, pages 312–377.
- Hsu, C. W. and Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.
- Huang, Y. and Wang, C. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, 96:1469–1482.
- Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12:531–547.
- Jung, S., Bang, H., and Young, S. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6:157–169.
- Lachenbruch, P. A. (1968). On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics*, 24(4):823–834.
- Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81.
- Li, S. S., Bigler, J., Lampe, J., Potter, J., and Feng, Z. (2005). Fdr-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine*, 15:2267–2280.
- Liu, P. and Hwang, J. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23:739–746.

- Lu, J., Plataniotis, K. N., and Venetsanopoulos, A. N. (2003). Face recognition using LDA-based algorithms. *IEEE Transactions on Neural Networks*, 14(1):195–200.
- Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, pages 29–42.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (2003). *Multivariate Analysis*. Academic Press.
- Moehler, T., Seckinger, A., Hose, D., Andrulis, M., Moreaux, J., Hielscher, T., Willhauck-Fleckenstein, M., Merling, A., Bertsch, U., Jauch, A., Goldschmidt, H., Klein, B., and Schwartz-Albiez, R. (2013). The glycome of normal and malignant plasma cells. *PLoS One*, 8:e83719.
- Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying dna microarray data. *Journal of Computational Biology*, 10(2):119–142.
- Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1):127–137.
- Novick, S. J. and Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481.
- Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8.
- Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024.

- Pfeffer, U., Romeo, F., Noonan, D. M., and Albini, A. (2009). Prediction of breast cancer metastasis by genomic profiling: where do we stand. *Clinical Exp Metastasis*, 26:547–558.
- Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1544–6115.
- Pounds, S. and Cheng, C. (2005). Sample size determination for the false discovery rate. *Bioinformatics*, 21:4263–4271.
- Qiao, Z., Zhou, L., and Huang, J. Z. (2009). Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data. *IAENG International Journal of Applied Mathematics*, 39(1):48–60.
- Richard, J. A. and Wichern, W. D. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 6th edition.
- Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346:1937–1947.
- Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*, 39:1241–1265.

- Shao, Y. and Tseng, C. H. (2007). Sample size calculation with dependence adjustment for fdr-control in microarray studies. *Statistics in Medicine*, 26:4219–4237.
- Simon, R. (2010). Clinical trials for predictive medicine: new challenges and paradigms. *Clinical Trials*, 7:516–524.
- Simon, R., Radmacher, M., Dobbin, K., and McShane, L. (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95:14–18.
- Sriperumbudur, B., Torres, D. A., and Lanckriet, R. (2011). A majorization-minimization approach to sparse generalized eigenvalue problem. *Journal of Machine Learning*, 85:3–39.
- Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74:703–716.
- Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.
- Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, 7:106.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.
- Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. theory of probability and its applications. *Theory of Probability and its Applications*, 16:264–280.
- Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91.

- Vinod, H. D. (1970). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, pages 147–166.
- Waaijenborg, S., de Witt Hamer, P. C. V., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7.
- Witten, D., Tibshirani, R., Gross, S., and Narasimhan, B. (2013). Package ‘pma’. <http://cran.r-project.org/web/packages/PMA/PMA.pdf>. Version 1.0.9.
- Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher’s linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.
- Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8.
- Witten, D. M., Tibshirani, R. J., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.
- Zhang, J. X., Song, W., Chen, Z. H., Wei, J. H., Liao, Y. J., Lei, J., Hu, M., Chen, G. Z., Liao, B., Lu, J., Zhao, H. W., Chen, W., He, Y. L., Wang, H. Y., Xie, D., and Luo, J. H. (2013). Prognostic and predictive value of a microrna signature in stage ii colon cancer: a microrna expression analysis. *Lancet Oncology*, 14:1295–1306.
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.

- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.
- Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming rna-seq data to improve the performance of prognostic gene signatures. *PLoS One*, 8:e85150.