New Regularization Methods for Supervised Learning with High-Dimensional Data

by

Ziyang Ma

(Under the Direction of Jeongyoun Ahn)

Abstract

Supervised learning problems in high-dimensional settings have a wide range of applications across different disciplines, such as the predictions using high-throughput data from molecular biology. High dimensionality poses many challenges to traditional supervised learning problems and has captured great attention in the statistics and machine learning community. One solution is to use regularization methods. This dissertation considers new regularization approaches for high-dimensional data under the context of two supervised learning topics. The first topic concerns the ordinal classification problems which lie between standard classification and regression. We propose two novel methods that consider a new regularization idea, which weights the features by calculating their rank correlations with the class labels. In the first method, we incorporate the feature weights into the framework of linear discriminant analysis and add the group Lasso penalty to achieve sparse solutions. In the second method, we add the weights into sparse optimal scoring with an adaptive Lasso penalty. Both of the proposed methods can project the original data onto a lower-dimensional subspace which reveals the underlying ordinal structure. This distinguishes our methods from existing work which assume a strict underlying linear ordinality within the data. We also demonstrate the difference between linear and nonlinear ordinality and show that our methods are capable of detecting the nonlinear ordinality and applicable to high-dimensional data. Simulation studies and real data examples show that the proposed methods have superior performance

for ordinal classification with respect to various evaluation metrics. The second topic revisits the trace ratio optimization problems involved in dimension reduction. Solving the trace ratio optimization is not straightforward and it is conventionally replaced by a sub-optimal alternative, the ratio trace problem. We consider a trace regularization method and modify it in the scenario of high-dimensional canonical correlation analysis (CCA). Results from numerical studies demonstrate the efficiency of the modified trace regularization method, compared with other well-known high-dimensional CCA approaches.

Index words: Regularization, High-dimensional data, Ordinal classification, Feature weighting, Canonical correlation analysis, Dimension reduction

New Regularization Methods for Supervised Learning with High-Dimensional Data

by

Ziyang Ma

B.S., Nankai University, China, 2016

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

2021

New Regularization Methods for Supervised Learning with High-Dimensional Data

by

Ziyang Ma

Major Professor: Jeongyoun Ahn

Committee: Yuan Ke

Liang Liu

Justin D Strait

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2021

Acknowledgments

First of all, I would like to express my deepest gratitude to my Ph.D advisor Dr. Jeongyoun Ahn who gave me invaluable guidance, tremendous support and encouragement throughout the course of my Ph.D study. Her generous guidance and immense support make it possible for me to successfully finish this dissertation.

I would also want to express my sincere gratitude to my committee members Dr. Yuan Ke, Dr. Liang Liu and Dr. Justin D Strait, who provided me with their insightful comments and suggestions that help me finish the dissertation. I am also grateful for their time spent on reading my dissertation.

Last but not least, I deeply thank my parents for their undying love and limitless support that carry me on everywhere I go. I also want to give a big thank to my boyfriend, Qing Wang, for his patience, love, and support that help me survive the dark times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Learning from data is one of the key components in statistics. Within the family of statistical learning, supervised learning which refers to the task of making predictions of an outcome measurement (such as house price) based on a sets of features (such as house age and house area), has been paid in great attention for decades. Supervised learning can be further categorized according to the types of outcomes. For example, regression refers to the learning when the outcome is quantitative; classification refers to the task when the outcome is qualitative (or categorical). Examples of supervised learning include linear discriminating analysis (LDA), support vector machines, and k-nearest neighbors (Fisher, 1936; Fix, 1985; Hearst et al., 1998). Traditionally in a dataset, the number of observations n will be much larger than the number of features p. With the development of modern technologies, a gigantic number of features are able to be collected, which can far exceed the number of observations. For instance, text data usually have a ultra-high dimension of features (such as the words in a dictionary) with a small sample size (Aggarwal & Zhai, 2012). Also, the molecular biology techniques have produced large repositories of high-throughput data, such as the gene expression microarrays which usually involve tens of thousands of features for tens of patients (samples) (Aliferis et al., 2006; Y. Wang et al., 2008).

However, many traditional supervised learning algorithms will encounter great challenges when dealing with high-dimensional data. As the number of features increases, the algorithms may face the risk of 'over-fitting' (Hawkins, 2004), in which the model lacks the ability to generalize with new data. The

correlations among features might also be increased as it is likely to get redundant features, which will cause the issue of multicollinearity that hinders the performance of regression models (Farrar & Glauber, 1967). In addition, traditional methods based on the inverse of sample covariance matrix will become degenerate when dimension exceeds sample size, for the reason that the inverse does not exist, such as traditional LDA and traditional canonical correlation analysis (CCA). What is more, samples will become more are more sparse as the dimension increases, which will decrease the efficiency of some algorithms, such as the nearest neighborhood (see details in Chapter 3).

The challenges brought by high-dimensional data have captured great attention and many efforts have been made to accommodate the algorithms in a high-dimensional setting (Bühlmann & Van De Geer, 2011). A popular way is to use regularization techniques, which help solve ill-posed problems or prevent from overfitting by modifying the original algorithms. There are various types of regularization methods. A well-known regularization method that addresses the multicollinearity issue in regression is the ridge regression estimator by Hoerl and Kennard (1970). Later, ridge-type regularization has been widely used in learning algorithms (Hastie, 2020). The Lasso penalty by R. Tibshirani (1996) is also a popular regularization method that can produce sparse solutions to enhance the interpret-ability and perform feature selection. There are also similar regularization methods, such as the group Lasso, fused Lasso and elastic net (R. Tibshirani et al., 2005; Yuan & Lin, 2006; Zou & Hastie, 2005). These regularization techniques have demonstrated a strong performance in practice. However, the performance of a general regularization method might depend on the context of the problems, there still remains the need on the developments of new regularization methods when addressing specific learning problems.

In this dissertation, we consider new regularization methods for two supervised learning problems, ordinal classification and canonical correlation analysis. The first topic, ordinal classification, is different from standard classification in the sense that the outcome measurements are ordered categories. Compared with standard classification, less attention has been paid on classifying ordinal data when dimension exceeds sample size. The ordinal problems with high-dimensional data are common in real life, such as using gene expression micrioarrays to classify ordered clinical responses. However, some existing ordinal

classification methods make too simplistic assumptions which seem not reasonable. Therefore, it is motivated to address the difficulties under this area. The other topic, canonical correlation analysis (CCA) (Hotelling, 1992) is a method that learns the relationship between two sets of variables. CCA is closely related with both regression and classification. The high-dimensional CCA is still a challenging task with wide applications, such as understanding the relationship between DNA markers and gene expressions (Waaijenborg et al., 2008).

This dissertation is structured as following. In Chapter 2, we introduce the problem of ordinal classification along with the related existing work. We propose a new regularization idea which adds weights on the features that account for the ordinal associations between features and the class labels. We impose the feature weights into well-developed standard classification methods and propose two novel methods. The first method is to add feature weights into the within-class covariance matrix in the LDA framework. A group Lasso penalty is also added to achieve sparse solutions in high-dimensional setting. The second method is to use feature weights in the framework of sparse optimal scoring. We carry out simulation studies and real data examples to demonstrate the promising performance of the proposed methods compared with some well-known methods. It is also shown that our proposed methods can reveal the underlying ordinal structure of the data by projecting the original data onto a well-trained discriminating subspace.

In Chapter 3, the high-dimensional CCA problem is considered. We first introduce the trace ratio optimization problems involved in dimension reduction and then introduce how the CCA problem can be re-formulated into a trace optimization problem. We then propose to modify the trace regularization method by Ahn et al. (2020) such that it can be applied under the context of high-dimensional CCA. A low rank approximation is taken into account with the modification of the trace regularization method. Simulation studies demonstrate that the proposed work is competitive under certain circumstances, when compared with benchmark models.

Chapter 2

Ordinal Classification for High Dimension, Low Sample Size Data

2.1 Introduction

Data with ordinal outcomes are common in an overwhelming number of statistical problems, with broad applications in biomedical science, social science and so on. Examples of ordinal outcomes include: responses to a treatment in clinical studies that are classified as 'Complete Response', 'Partial Response', 'Minimum Response', 'No Change' or 'Progressive Disease' (BladÉ et al., 1998); tumor-node-metastasis (TNM) stages classified as 'Stage 0', 'Stage I','Stage II' ,'Stage III', or 'Stage IV'; customers' credit scores categorized as bad, fair, good or excellent. These ordinal labels are in contrast to nominal labels such as types of tumors, in that there are natural orderings among the classes. Ordinal classification, also known as ordinal regression, refers to the supervised learning task of predicting patterns into ordinal labels. However, as the face values of the labels are themselves meaningless, it cannot be considered as a regression problem, in which the response is a continuous variable taking real values. Also, in classification problems, the labels of the classes are purely categorical, in which there is no rank among the labels, such as classifying the species of flowers into several known categories. Therefore, ordinal classification lies between the framework of classification and regression. Generally speaking, given an input vector $\mathbf{x} \in R^p$, where

p denotes the dimensionality, the goal of an ordinal classification task is to learn a function (mapping) $f : \mathcal{X}$ (input space) $\rightarrow \mathcal{Y}$ (output space), which maps $\mathbf{x}$ into one of the classes $\mathcal{C}_k$, $k \in \{1, \ldots, K\}$, and $\mathcal{C}_k$ inherits the natural ordering where $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_K$ ($\prec$ denotes the ordering).

With the development of modern technologies, complex datasets with thousands of features are generated rapidly. Compared with traditional datasets, these types of datasets share a common characteristic of having enormous amount of features and limited number of observations, i.e., $p > n$, which are also known as high-dimension-low-sample-size (HDLSS) data. The high dimensionality brings great challenges to traditional statistical or machine learning methods, in the sense that either many methods become degenerate in such situation or many methods loss their interpret-abilities. However, the HDLSS data are benefiting many research areas. For example, multiple high-throughput platforms are providing a large repository of gene expression data that facilitate the biomedical research. The gene expression data can reveal many powerful features in predicting clinical response for modern therapies, in which traditional clinical and laboratory features have limited information (Mulligan et al., 2007). Therefore, the use of HDLSS data is inevitable. Compared with the decent amount of work on classifying HDLSS data with nominal outcomes, less attention has been paid on classifying the ordinal outcomes when the dimensions far exceeds the sample size. Thus, there remains an urge to develop novel methods which can accommodate HDLSS data to solve existing ordinal classification challenges.

In this chapter, we provide a literature review on the existing work for ordinal classification and related work on high-dimensional classification, given in Section 2.2. We introduce our proposed methods that use a new regularization idea in Section 2.3 and 2.4 . The methods are capable of handling the HDLSS data. The model evaluation metrics for ordinal classification are introduced in Section 2.5. In Section 2.6, we perform a comprehensive simulation study to test the performances of the proposed methods and related work under different scenarios. We also apply the methods on real data examples, as given in Section 2.7, including the predictions of tumor grades and clinical responses in the biomedical area. Finally, we conclude this chapter with discussions on the summary of the performances and future directions in Section 2.8. Supplementary materials are included in Section 2.9.

2.2 Literature Review

In this section, we provide a literature review on the existing work for ordinal classification and related work on classification for HDLSS data. Traditional methods for ordinal classification might become degenerate when the dimension exceeds the sample size, however, they serve as building blocks in the ordinal classification area and are worthwhile to be understood and discussed. We will also introduce some well-known classification methods which are suitable for high dimensional situation. These work can also be applied to ordinal classification even though the ordinal information within the labels are ignored. Next, we also discuss the existing work on ordinal classification for high-dimensional data and their limitations.

We use X to denote an $n \times p$ input data matrix, with n observations and p variables. Let $X_{n\times p} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)^T = (\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_p)$, where $\mathbf{x}_i \in R^p$ is the ith row vector, representing the ith observation and $\widetilde{\mathbf{x}}_j \in R^n$ is the jth column vector for variable (feature) j. Assume that each of the n observations falls into one of the K ordinal classes C_k, $k \in \{1, \ldots, K\}$, and C_k inherits the natural ordering where $C_1 \prec C_2 \prec \cdots \prec C_K$ ($\prec$ denotes the ordering). We use $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ to denote the vector that contains the class label for each observation, with $y_i \in \{1, 2, \ldots, K\}$.

2.2.1 Existing Approaches to Ordinal Classification

Traditional ordinal classification methods can be categorized into the following groups: thresholding (SVM-based) methods, regression methods, decomposition methods, cost-related methods and other types. These categories are a modified version of the categories defined by Gutierrez et al., 2015, who provided an overview and an experimental study of current ordinal classification methods.

Support Vector Machines (SVM)-based Methods

The first type is the SVM-based methods, in which we assume that there are a set of underlying threshold values that separate the ordered classes. The values of thresholds t_k are assumed to be ordinal such that $t_0 \leq t_1 \leq \cdots \leq t_K$, and observations from the class k are assumed to be bounded by t_{k-1} and t_k.

Support vector machines (Boser et al., 1992) has been widely used on binary classifications, which aims to find a hyperplane that separates the classes with maximum margin. Suppose the observations in the training dataset come from two classes, with notations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where y_i is either 1 or -1, representing the positive and negative class respectively. The target hyperplane in SVM will discriminate the positive samples from the negative samples, such that the distance between the hyperplane and the nearest sample from either group is maximized. Suppose that the hyperplane can be defined as:

$$\mathbf{w^T x} - b = 0,$$

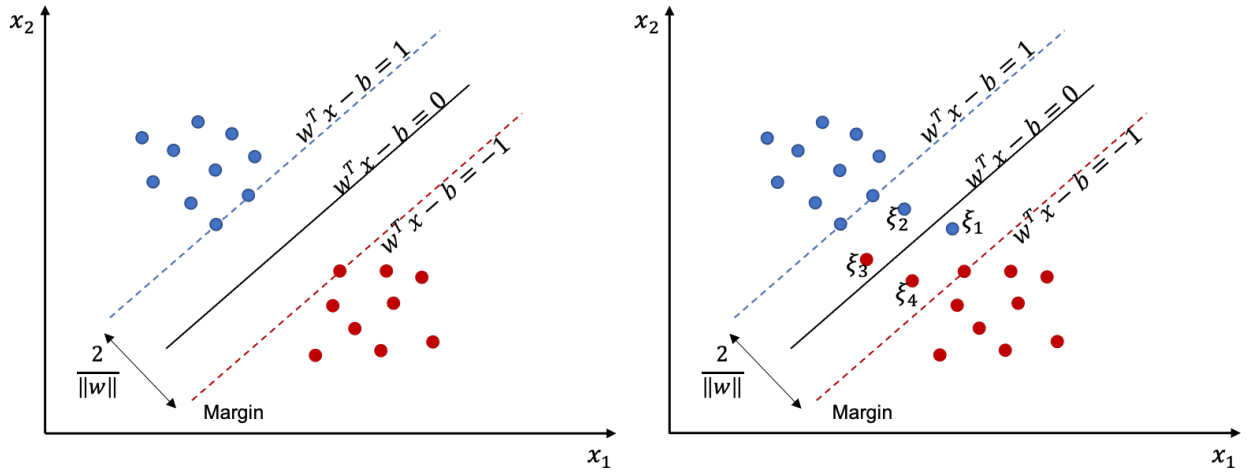
where $\mathbf{w}, \mathbf{x} \in R^p$, it can be shown that the margin of the hyperplane is $\frac{2}{\|\mathbf{w}\|_2}$. Then the objective of SVM can be written as:

$$\begin{aligned} &\max_{\mathbf{w} \in R^p} \frac{2}{\|\mathbf{w}\|_2}, \\ &\text{subject to } \; y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1, \text{ for } \; i = 1, \ldots, n. \end{aligned}$$

which is equivalent to:

$$\begin{aligned} &\min_{\mathbf{w} \in R^p} \frac{1}{2}\|\mathbf{w}\|_2^2, \\ &\text{subject to } \; y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1, \text{ for } \; i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where the constraints in (2.1) ensures the separability between the two classes, in the sense that positive samples will satisfy the condition that $\mathbf{w}^T \mathbf{x}_i - b \geq 1$ and negative samples will satisfy that $\mathbf{w}^T \mathbf{x}_i - b \leq -1$. The objective in (2.1) is also referred as the hard-margin SVM, which is applicable to the situation when data are linearly separable. Figure 2.1a gives a simple illustration of the hard-margin SVM when data are



(a) Hard-margin SVM

(b) Soft-margin SVM

Figure 2.1: Illustration of SVM

linearly separable. When data are not linearly separable, the concept of soft-margin is introduced, which allows for some points to be on the wrong side of the margin. The slack variables ξ_i are created for each sample point, such that samples on the right side have $\xi_i = 0$ and samples on the wrong side have $\xi_i > 0$. Figure 2.1b shows the example of four sample points that are on the wrong side with $\xi_i > 0$. The objective function of soft-margin SVM is give by:

$$\min_{\mathbf{w} \in R^p} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i,$$

$$\text{subject to } y_i(\mathbf{w}^T\mathbf{x}_i - b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, \text{ for } i = 1, \ldots, n,$$

where C is a hyper-parameter.

Although SVM was originally designed for binary classes, it has been widely extended to solve multi-class problems. One approach is to solve a series of binary classification problems, such as pairwise comparison or one-versus-all (Crammer & Singer, 2001; Dietterich & Bakiri, 1994); another approach is to apply SVM to multi-class problems directly (Y. Lee et al., 2004). In the context of classifying ordinal

data, Herbrich et al. (1999) proposed the first SVM-based method for ordinal classification. They derived new feature vectors based on the original features and took the ordinal problem as a binary classification problem in which they aimed to find parallel hyperplanes to classify the new training set. Specifically, the new training set consists of the differences between original input vectors and signs of the corresponding label differences, i.e., $\{(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}, z_i = \text{sign}(y_i^{(1)} \ominus y_i^{(2)}))\}$, where the notations $\mathbf{x}_i^{(1)}$ and $\mathbf{x}_i^{(2)}$ denote a pair coming from the original training set, and $\ominus$ denotes the rank difference. They assume that there is a linear function which maps the input space to the ranks, i.e., $U : X \to R, U(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, such that that there exists rank boundaries $\theta(r_k)$, for $k = 1, \ldots, K-1$, which separate $U(\mathbf{x})$ and satisfy the constraints: $\theta(r_1) \prec \cdots \prec \theta(r_{K-1})$. Figure 2.2 gives a visual illustration of the idea when the input data are two-dimensional. The goal is also to maximize the margins for the separating hyperplanes $\theta(r_k)$. Clearly when $U(\mathbf{x})$ incurs no error, there should be no inversion of $(U(\mathbf{x}_i^{(1)}), U(\mathbf{x}_i^{(2)}))$ compared with the original $(y_i^{(1)}, y_i^{(2)})$, i.e.,

$$z_i \mathbf{w}^T(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}) > 0.$$

Similar with the soft-margin SVM, if a small amount of violations are allowed, the constraints could be relaxed to

$$z_i \mathbf{w}^T(\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}) \geq 1 - \xi_i, \tag{2.2}$$

where ξ_i is the slack variable defined based on the new feature vectors $\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$. In short, Herbrich et al., 1999 proposed to minimize $\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{s} \xi_i$ with the constraints of (2.2), where s is the number of observations in the new training set. Their method is closely related to the soft-margin SVM.

Later, Shashua and Levin (2003) also utilized the principle of largest margin in SVM to find $K-1$ thresholds that map the ranks of classes to consecutive intervals on the real line. Geometrically, the idea is to find $K-1$ parallel hyperplanes separating the samples into K classes, which are defined by $\mathbf{w}^T\mathbf{x} - t_k = 0$, for $\mathbf{w} \in R^p$. An observation $\mathbf{x}$ satisfying $t_{k-1} < \mathbf{w}^T\mathbf{x} < t_k$ will be predicted to the class k. They proposed two strategies to maximize the margins, one is to maximize the margin between neighbor classes, the other is to maximize the sum of the individual margins. After that, Chu and Keerthi (2005) provided

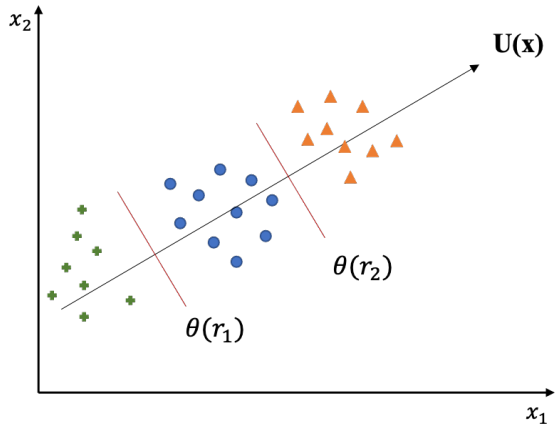


Figure 2.2: A simple illustration of the method by Herbrich et al. (1999) when the input data are two-dimensional. Samples from three classes are identified by different colors and shapes. $\theta(r_k)$ denotes the rank boundaries.

a modification by arguing that the thresholds obtained in the work of Shashua and Levin (2003) are not guaranteed to satisfy the ordering inequalities: $t_1 \leq t_2 \leq \cdots \leq t_{K-1}$. They added explicit constraints on the thresholds and their method is named as SVOREX. In addition, Chu and Keerthi (2005) proposed another new approach which integrates implicit constraints for the thresholds, named as SVMORIM.

The above discussed SVM-based methods for ordinal classification share the common characteristic of utilizing a distribution assumption of a latent variable and finding parallel hyperplanes which maximize the margins among classes. Specifically, they assume that there exists a function that maps the input space to the one-dimensional real space. Depending on the class labels, values of the function will fall into intervals on the real line, which are characterized by a sets of thresholds satisfying $t_1 \leq t_2 \leq \cdots \leq t_{K-1}$.

Regression Methods

The second type of existing ordinal approaches is the regression-type methods, which take the ordered class labels as continuous responses and fit a regression model. Regression models for ordinal data was first proposed by McCullagh (1980), which is referred as the proportional odds model. In the work of McCullagh (1980), the ordinal categories are thought of some contiguous intervals with continuous

scales with cutting points θ_k. It is assumed that the response variable y takes values ranging from 1 to K, and follows a multinomial distribution with probabilities $\{\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \ldots, \pi_K(\mathbf{x})\}$, given the input $\mathbf{x}$. For $k \in \{1, \ldots, K\}$, let $r_k(\mathbf{x})$ be the odds of $y \leq k$ given $\mathbf{x}$, then the proportional odds model makes the assumption that the logarithm of the odds $r_k(\mathbf{x})$ is proportional to the covariate $\mathbf{x}$, i.e., $r_k(\mathbf{x}) = c_k \exp\{-\boldsymbol{\beta}^T \mathbf{x}\}$. Then the ratio of the odds given two different covariates $\mathbf{x}_1$ and $\mathbf{x}_2$, is linearly related with the difference between the covariates, and is independent of k, i.e.,

$$\frac{r_k(\mathbf{x}_1)}{r_k(\mathbf{x}_2)} = \boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2).$$

Formally, the proportional odds model can be written as:

$$\log(\frac{P(y_i \leq k|\mathbf{x}_i)}{P(y_i > k|\mathbf{x}_i)}) = \log(r_k(\mathbf{x}_i)) = \theta_k - \boldsymbol{\beta}^T \mathbf{x}_i,$$

where $\theta_k = log(c_k)$ denotes the corresponding intercept, which is also the cutting point between adjacent categories. Then, a new sample will be classified to the class with the highest posterior probability.

Another related work is the continuation ratio (CR) model, which was first proposed by Fienberg (1980). Compared with the proportional odds model, the cumulative probabilities $P(y_i \leq k|\mathbf{x}_i)$ was replaced by $P(y_i = k|y_i \geq k, \mathbf{x}_i)$ in the continuation ratio model, i.e., the probability of y_i being in the category k given the condition of being in the categories that are equal or greater than k. The continuation ratio model can be expressed as:

$$\log \frac{P(y_i = k|y_i \geq k, \mathbf{x}_i)}{P(y_i > k|y_i \geq k, \mathbf{x}_i)} = \theta_k - \boldsymbol{\beta}^T \mathbf{x}_i, \text{ for } k = 1, 2, \ldots, K-1. \tag{2.3}$$

The model (2.3) is also known as the forward continuation ratio model and could have different set-ups. In contrast, if the conditional probability was replaced by $P(y_i = k|y_i \leq k, \mathbf{x}_i)$, (2.3) will become the backward continuation ratio model. Note that the proportional model will not be changed even when the coding of the response variable is reversed. In comparison, the CR model will be altered if the coding

is reversed and it pays more attention to the individual categories (Ananth & Kleinbaum, 1997; Liu et al., 2011).

Support vector regression (SVR), proposed by Drucker et al., 1997, as a regression analog to the support vector machines, could also be applied to ordinal data. Specifically, SVR predicts the labels by $\mathbf{w}^T\mathbf{x}_i$ and its objective function can be written as:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2,$$
$$\text{subject to } |y_i - \mathbf{w}^T\mathbf{x}_i| \leq \epsilon, \text{ for } i = 1, \ldots, n.$$

When allowing violations, the objective of SVR could be written as:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i,$$
$$\text{subject to } |y_i - \mathbf{w}^T\mathbf{x}_i| \leq \epsilon + \xi_i, \text{ for } i = 1, \ldots, n.$$

where $\xi_i \geq 0$ are the slack variables that are similar to those defined in support vector machine.

These regression-type methods are more sensitive to the numerical representations of the labels rather than the ordering information among them. Moreover, it is not reasonable to make assumption of equal distancing among the classes without prior knowledge of the true metric among classes. For example, consider predicting the weather conditions: cold $\prec$ warm $\prec$ hot, with labels coded as $\{1, 2, 3\}$ for $\{\text{cold}, \text{warm}, \text{hot}\}$, respectively, it is not precise or reasonable to say that the distance between hot and warm is 1, which is the same as the distance between warm and cold. Thus, naively treating the ordinal labels as a continuous response may add artificial information, which will hinder the performance of the model.

Decomposition Methods

Another main type of the ordinal approaches is the decomposition-type methods, which decompose the multi-class problem into several binary problems. Frank and Hall (2001) proposed to decompose the

original K-class classification problem into $K-1$ binary classification problems. Given the class labels ranging from 1 to K, each binary classifier will be discriminating a sample from a class with label no greater than k versus greater than k ($k \in \{1, \ldots, K-1\}$). In an example of four-class problem, three binary classifiers will be built on the classes: $\{C_1\}$ vs $\{C_2, C_3, C_4\}$, $\{C_1, C_2\}$ vs $\{C_3, C_4\}$ and $\{C_1, C_2, C_3\}$ vs $\{C_4\}$. After that, the posterior probabilities for a class given sample $\mathbf{x}_i$ are calculated in the following way:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= 1 - P(y_i > 1|\mathbf{x}_i), \\ P(y_i = k|\mathbf{x}_i) &= P(y_i > k-1|\mathbf{x}_i) \times P(y_i \leq k|\mathbf{x}_i)) \\ &= P(y_i > k-1|\mathbf{x}_i) \times (1 - P(y_i > k|\mathbf{x}_i)), \text{ for } 1 < k < K, \\ P(y_i = K|\mathbf{x}_i) &= P(y_i > (K-1|\mathbf{x}_i)). \end{aligned} \tag{2.4}$$

Then based on (2.4), the sample $\mathbf{x}_i$ will be assigned to the class with the largest posterior probability. Note that another way to calculate $P(y_i = k|\mathbf{x}_i)$ is: $P(y_i = k|\mathbf{x}_i) = P(y_i > k|\mathbf{x}_i) - P(y_i > (k-1)|\mathbf{x}_i)$, however, negative probabilities may occur in this case. One way to avoid negative probabilities is to replace the probability p with $\max(p, 0)$. Frank and Hall (2001) argued that this simple decomposition approach is applicable with many standard classification methods and they demonstrated that it can enhance the performance of the C4.5 decision tree learner (Quinlan, 2014) for ordinal classification. Later, Huhn and Hüllermeier (2008) conducted an experiment to evaluate the performance of decomposing approaches for ordinal classification and concluded that these methods are able to learn the ordinal information among the classes.

However, $K-1$ classifier are required to be built for the decomposition-type approaches, which will increase the complexity and introduce modeling error multiple times. Another limitation is that the predictions might be ambiguous due to the crossing of classification boundaries. There is a related work by Qiao (2015), in which the noncrossing constraints were proposed when constructing the boundaries simultaneously.

Cost Related Methods

The fourth type of methods are those utilizing specific loss functions designed for ordinal classification. As we know, under the context of ordinal data, it might be not reasonable to assume the same loss for the different types of misclassification regarding the distance between classes. For example, there should be more cost if we misclassified 'hot' as 'cold' than misclassifying 'hot' as 'warm', since 'hot' and 'warm' are more similar than 'hot' and 'cold'.

Kotsiantis and Pintelas (2004) proposed a cost-sensitive technique by introducing $c_{kl} = |k - l|$ as the relative cost for misclassifying class k to class l (or the opposite way). In such way, further classes suffer from higher cost when being misclassified compared with nearer classes. Explicitly, this idea minimizes the conditional risk by choosing k which minimizes $R(\hat{y}_i|\mathbf{x}_i) = \sum_l c_{lk} P(y = k|\mathbf{x}_i)$. Kotsiantis and Pintelas (2004) also conducted some experiments to show that the incorporation of this cost-sensitive technique can help achieve smaller mean absolute error (MAE) while not decreasing accuracy, when applied to learning algorithms which can output predicted probabilities, such as decision trees. This cost-sensitive technique provides a direction of making use of the ordinal information among the classes. There are also other potential criteria, such as $c_{kl} = |k - l|^2$. However, because of such flexibility, it is hard to find a suitable cost criterion which reflects the true metric among classes, without prior knowledge.

There are also other methods working on the cost function for ordinal classification. de La Torre et al. (2018) utilized the weighted kappa loss function and implemented it in the deep learning models for ordinal classification. The Kappa index is a statistics defined to measure the inter-rater agreement when classifying subjects into categories. The Weighted Kappa index is then defined to allow different weights for disagreements, and is used for ordinal categories. The Weighted Kappa is expressed as:

$$\kappa = 1 - \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl} o_{kl}}{\sum_{k=1}^{K} \sum_{l=1}^{K} w_{kl} e_{kl}},$$

where o_{kl} denotes the number of observations classified in the kth category and actually belong to the lth category, e_{kl} denotes the corresponding expected number of observations, and w_{kl} denotes the correspond-

ing weight. de La Torre et al. (2018) used three classification problems to demonstrate the improvements on model performance when using the weighted kappa loss function compared with using the logarithmic loss function.

These cost-sensitive approaches shed lights on a direction to deal with ordinal classification, which might be incorporated with standard classification methods to achieve better performance.

Other Types

At last, there are some methods that can not be categorized into the above categories, which we refer as other types. There is one method named the data replication method, proposed by Cardoso and Costa (2007). They introduced a new paradigm for ordinal classification by replicating data in an augmented feature space, in which the original data are projected into a $(p + K - 1)$-dimensional space. Then $K - 1$ parallel linear boundaries obtained from the binary classifiers on the augmented space will be sufficient to classify the data in the original space. Cardoso and Costa (2007) argued that the paradigm can be implemented with the framework of SVM and neural networks.

Another method is the kernel discriminant analysis (KDA) introduced by B.-Y. Sun et al. (2009). This method aims to find a projection which minimizes the within-class distance and maximizes the between-class distance, while satisfying the constraints that the average projection of samples from higher rank class is larger than that of lower rank class. The optimization function of KDA takes the form:

$$\begin{aligned} &\min J(\mathbf{w}, \rho) = \mathbf{w}^T \Sigma_w \mathbf{w} - C\rho, \\ &\text{s.t. } \mathbf{w}^T(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k) \geq \rho, \text{ for } k = 1, \ldots, K-1, \end{aligned} \tag{2.5}$$

where $\mathbf{w} \in R^p$ is the discriminant vector, Σ_w denotes the within-class covariance (see Section 2.2.2 for more details), C is a penalty coefficient. B.-Y. Sun et al. (2009) argued that their method has a lower computational complexity compared with the SVM-based methods, and showed that their method can preserve the ordinal information among the classes. However, the method in (2.5) is not applicable to the high-dimensional situation and such rank constraints are too rigid. Further, Cardoso et al. (2012)

conducted experiments on both synthetic and real data to compare the performances of KDA, kernel version of LDA with the decomposing idea from Frank and Hall (2001), and the data replication method with kernel version of LDA. The experiments did not show the superiority of KDA among the other two methods. However, the projection-based KDA still provides a direction for ordinal classification and it may be applicable to the high-dimensional situation if some regularizations are added.

2.2.2 Existing High-dimensional Classification Approaches

Ignoring the ordinal information among the class labels, nominal classification approaches could also be applied on ordinal classification problems. We will discuss some standard high-dimensional classification methods in this section, which shed lights on classifying ordinal data when facing high-dimensional challenges. We start with an overview of the linear discriminant analysis (LDA), a famous method for classification as well as dimension reduction. LDA is well-known for its simplicity and robustness and it also serves as a building block for many classification methods for high-dimensional data. In addition, optimal scoring, a regression analog for LDA, will be introduced, for the reason that many high-dimensional methods are based on a penalized version of optimal scoring. Then, we discuss the current work on classifying high-dimensional data.

Linear Discriminant Analysis

Fisher's linear discriminant analysis (Fisher, 1936) is a popular tool for classification, because of its simplicity and superior robust performance. The idea behind LDA can be expressed in two ways. On one hand, from the perspective of Bayes rule, LDA aims to predict the classes based on posterior probabilities. Let X be the $n \times p$ input matrix, with each observation belonging to one of the K classes with label k, for $k \in \{1, \ldots, K\}$, and assume the columns of X have been centered. Assume that the samples come from a multivariate Gaussian model, in which the observations from class k are distributed from $N(\boldsymbol{\mu}_k, \Sigma_w)$, where $\boldsymbol{\mu}_k \in R^p$ is the population mean vector for class k, and Σ_w is the common population within-class

covariance matrix. The density for $\mathbf{x}$ in class k takes the form:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_w|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_w^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\}. \tag{2.6}$$

Further, assume that the prior probability for class k is π_k, with $\sum_{k=1}^{K}\pi_k = 1$, then the posterior probability for class k can be calculated based on the Bayes rules, given as:

$$P(y=k|\mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{x})\pi_l}. \tag{2.7}$$

By calculating $\frac{\log(P(y=k|\mathbf{x}))}{\log(P(y=l|\mathbf{x}))}$ using (2.6) and (2.7) and cancelling out the same terms, the discriminant function for class k can be derived as:

$$\delta_k(\mathbf{x}) = \mathbf{x}^T\Sigma_w^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\Sigma_w^{-1}\boldsymbol{\mu}_k + \log\pi_k.$$

Then the predicted class label given $\mathbf{x}$ is $\hat{y} = \arg\max_k \delta_k(\mathbf{x})$. In practice, the prior probabilities, class mean vectors and the pooled within-class covariance matrix are estimated using the sample estimates, given as below:

$$\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n}, \\
\hat{\boldsymbol{\mu}}_k &= \frac{1}{n_k}\sum_{y_i=k}\mathbf{x}_i, \\
\hat{\Sigma}_w &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{y_i=k}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_k)^T,
\end{aligned} \tag{2.8}$$

where n_k is the number of observations in class k.

On the other hand, LDA can also be derived from the idea of finding a discriminant subspace on which the projected data are best separated, in terms of obtaining a maximum between-class covariance and a minimum within-class covariance. The sequence of discriminant vectors $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1})$ which

span the discriminant subspace can be obtained by the following optimization problem:

$$\begin{aligned} &\max_{\boldsymbol{\beta}_m \in R^p} \boldsymbol{\beta}_m^T \Sigma_b \boldsymbol{\beta}_m, \\ &\text{subject to } \boldsymbol{\beta}_m^T \Sigma_w \boldsymbol{\beta}_m = 1, \\ &\qquad \boldsymbol{\beta}_m^T \Sigma_w \boldsymbol{\beta}_l = 0, \ \forall\, l < m, \end{aligned} \tag{2.9}$$

where $1 \leq m \leq K-1$, Σ_b is the between-class covariance matrix that can be estimated as $\hat{\Sigma}_b = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^T$ and Σ_w can be obtained by (2.8). The solution to (2.9) can be obtained from a generalized eigenvalue problem: $\hat{\Sigma}_b \boldsymbol{\beta} = \rho \hat{\Sigma}_w \boldsymbol{\beta}$, with ρ being the generalized eigenvalue.

Optimal Scoring

As an analog of LDA, the idea of optimal scoring (Hastie et al., 1994) comes from recasting LDA as a regression problem. It assumes that there is a function that assigns real-value scores to the classes, in which the scores can be predicted by regressing on X. Let Y be a $n \times K$ matrix of dummy variables, indicating the class membership of each observation, i.e., $Y_{i,k} = 1$ if $\mathbf{x}_i$ belongs to the class k, otherwise, $Y_{i,k} = 0$, then, the objective function of optimal scoring takes the form:

$$\begin{aligned} &\min_{\boldsymbol{\theta}_m \in R^K,\, \boldsymbol{\beta}_m \in R^p} \|Y\boldsymbol{\theta}_m - X\boldsymbol{\beta}_m\|^2, \\ &\text{subject to } \frac{1}{n} \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ &\qquad \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0, \ \forall\, l < m, \end{aligned} \tag{2.10}$$

where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K-1}\}$ is a set of independent score vectors, $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}\}$ is a set of discriminant vectors, and $Y\boldsymbol{\theta}_m$ corresponds to the scores of classes. It is easy to show that given $\boldsymbol{\theta}_m$, $\boldsymbol{\beta}_m$ can be obtained by a least square estimate of regressing X on $Y\boldsymbol{\theta}_m$; given $\boldsymbol{\beta}_m$, $\boldsymbol{\theta}_m$ can be obtained by a restricted regression problem. Then the solution to (2.10) can be obtained by iterating between $\boldsymbol{\theta}_m$ and $\boldsymbol{\beta}_m$ until convergence.

It can be shown that optimal scoring is indeed equivalent to LDA (see details in Section 2.9.1). Given this fact, optimal scoring is also a popular choice for classification, and it is flexible to be incorporated with different penalties.

High-dimensional Classification Methods

However, the solutions to LDA and optimal scoring are based on the inverse of $\hat{\Sigma}_w$ (or the covariance matrix of X). As we know, when $p \gg n$ or the features are highly correlated with each other, the covariance matrix of X is not full rank and becomes singular. In such case, the inverse does not exist. Therefore, traditional LDA or optimal scoring can not be applied to HDLSS data. In addition, when p is large, it is hard to interpret the methods with a large amount of non-zero coefficients. Given these challenges, some efforts have been made to modify the original LDA such that it can be applicable for high-dimensional data.

J. H. Friedman (1989) suggested a method called regularized discriminant analysis (RDA), which can be viewed as a compromise between LDA and quadratic discriminant analysis (QDA). Different from LDA, QDA does not assume the same covariance matrix among different classes, instead, it estimates the covariance matrix for each class separately. Then, RDA uses a linear combination between the individual estimated class covariance matrix and the overall within-class covariance matrix. The estimated covariance matrix for class k from RDA is given by:

$$\hat{\Sigma}_k(\gamma, \lambda) = (1-\gamma)[(1-\lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma}_w] + \gamma I, \tag{2.11}$$

where $\gamma, \lambda \in [0, 1]$, $\hat{\Sigma}_k$ is the estimated covariance matrix for class k from QDA, and $\hat{\Sigma}_w$ is the estimated within-class covariance matrix from LDA, I is the identity matrix. A ridge-corrected LDA can be achieved by setting λ to be 1 and varying γ in (2.11). This is equivalent to say, instead of estimating the covariance matrix by a common within-class covariance matrix, ridge -corrected LDA is replacing Σ_w with $\Sigma_w + \alpha I$, where $\alpha = \frac{\gamma}{1-\gamma}$. Such regularization deals with the issue when there is a large number of highly correlated

features, in which the traditional LDA suffers from the risk of overfitting. In addition, this ridge-corrected LDA can also be applied to HDLSS data, since the issue of singularity will be resolved by adding αI.

Another penalized version of LDA was proposed by Hastie et al. (1995), which replaces Σ_w with $\Sigma_w + \lambda\Omega$, where Ω is a penalty matrix characterizing the 'roughness'. For example, for log-spectra data or image data, Ω is defined in the way such that nearby components of $\boldsymbol{\beta}_m$ are forced to be similar. Hastie et al. (1995) also demonstrated the equivalence between the penalized version of LDA and the penalized version of optimal scoring, in which the classification problem can be formulated with the framework of regression:

$$\min_{\boldsymbol{\theta}_m \in R^K,\, \boldsymbol{\beta}_m \in R^p} \|Y\boldsymbol{\theta}_m - X\boldsymbol{\beta}_m\|_2^2 + \boldsymbol{\beta}_m^T \Omega \boldsymbol{\beta}_m, \text{ subject to } \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1.$$

A related work is the method proposed by Clemmensen et al. (2011), named as sparse discriminant analysis (SDA), which uses the elastic net penalty in the framework of optimal scoring. Given a regression problem, the elastic net penalty (Zou & Hastie, 2005) solves the following problem:

$$\min_{\boldsymbol{\beta} \in R^p} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \gamma\|\boldsymbol{\beta}\|_2^2.$$

SDA finds the pair $(\boldsymbol{\theta}_m, \boldsymbol{\beta}_m)$ that optimizes:

$$\begin{aligned} \min_{\boldsymbol{\theta}_m \in R^K,\, \boldsymbol{\beta}_m \in R^p} & \|Y\boldsymbol{\theta}_m - X\boldsymbol{\beta}_m\|_2^2 + \gamma\boldsymbol{\beta}_m^T\Omega\boldsymbol{\beta}_m + \lambda\|\boldsymbol{\beta}_m\|_1, \\ \text{subject to } & \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ & \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0,\ \forall\, l < m, \end{aligned}$$

where Ω is a $p \times p$ positive definite matrix. By adding the penalty $\|\boldsymbol{\beta}_m\|_1$, SDA can perform the classification as well as feature selection at the same time to increase the interpret-ability in the high-dimensional setting.

Another similar work is the method by D. M. Witten and Tibshirani (2011), which adds additional penalty on the discriminant vectors given the framework of the LDA. The method is named penalized LDA, with the optimization criterion:

$$\begin{aligned} &\max_{\boldsymbol{\beta}_m \in R^p} \boldsymbol{\beta}_m^T \Sigma_b \boldsymbol{\beta}_j - P_1(\boldsymbol{\beta}_m), \\ \text{subject to } &\boldsymbol{\beta}_m^T \widetilde{\Sigma}_w \boldsymbol{\beta}_m \leq 1, \\ &\boldsymbol{\beta}_m^T \widetilde{\Sigma}_w \boldsymbol{\beta}_l = 0, \ \forall\, l < m, \end{aligned}$$

where $P_1(\boldsymbol{\beta}_m)$ is a convex penalty on $\boldsymbol{\beta}_m$, such as the Lasso or fused Lasso penalties, and $\widetilde{\Sigma}_w$ is a positive estimate of Σ_w, such as $\Sigma_w + \lambda\Omega$ or a diagonal estimate (Bickel, Levina, et al., 2004; Dudoit et al., 2002).

Recently, Mai and Zou (2015) provided a new approach which estimates the discriminant vectors simultaneously instead of sequentially. They utilized the Bayes rule for mutli-class linear discriminant analysis with a built-in penalty. Specifically, the method is comparing class 1 to the rest of the classes. The optimization criterion takes the following form:

$$\min_{\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_K} \sum_{m=2}^{K} [\frac{1}{2} \boldsymbol{\beta}_m^T \hat{\Sigma} \boldsymbol{\beta}_m - (\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_1)^T \boldsymbol{\beta}_m] + \lambda \sum_{l=1}^{p} \|\boldsymbol{\beta}_{\cdot l}\|_2^2,$$

where $\boldsymbol{\beta}_{\cdot l}$ is the vector consisting of the lth component of $\boldsymbol{\beta}_m$, for $m = 2, \ldots, K$, i.e., $\boldsymbol{\beta}_{\cdot l} = (\beta_{2l}, \ldots, \beta_{Kl})^T$. However, it seems that there is no rule of the choice of base class, and it is unsure whether choosing other base classes will result in different solutions.

Above all, these classification methods for high-dimensional data are not specifically designed for ordinal problems. However, they still can be applied to ordinal problems if ignoring the ordinal information, and they serve as building blocks for making ordinal classifications for HDLSS data.

2.2.3 Existing High-Dimensional Ordinal Approaches

Many of the ordinal classification methods discussed in Section 2.2.1 are typically not suitable for HDLSS data. Compared with the decent amount of work on classifying HDLSS nominal data or low-dimensional

ordinal data, less attention has been paid on classifying the ordinal outcomes when the dimension far exceeds the sample size. We have found some work that are designed to deal with HDLSS data in an ordinal setting. For example, Leha et al. (2013) proposed a method named hierarchical twoing (hi2), as an extension of the idea proposed by Frank and Hall (2001), which can perform ordinal classifications for high-throughput gene expression data. However, their work lacks theoretical justifications. Archer and Williams (2012) integrated the idea of Lasso and continuation ratio model to make a new approach that is applicable of classifying high-throughput gene expression data with ordinal classes. Their idea can be expressed as:

$$\max_{\boldsymbol{\beta} \in R^p} L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x})$ denotes the likelihood for the continuation ratio model and λ is the parameter controlling the degree of Lasso penalty. Archer and Williams (2012) applied their method on gene expression datasets and concluded that the non-zero coefficients obtained typically have a monotonic relationship with the ordinal responses, whose corresponding genes may serve as the information genes related to the progress of the disease. Similarly, Zhang et al. (2018) proposed to use a hierarchical ordinal regression model (BhGLM) to pridict the ordinal drug responses with gene expression profiles for patients with Multiple Myeloma, their model can be expressed as:

$$P(y_i = k) = \begin{cases} 1 - logit^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - c_1), \text{ for } k = 1, \\ logit^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - c_{k-1}) - logit^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - c_k), \text{ for } 1 < k < K, \\ logit^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - c_{k-1}), \text{ for } k = K, \end{cases}$$

in which a Cauchy prior was applied on the distribution of the coefficients. However, this line of approaches assume a strict linear ordinality among the classes. In other words, they all assume that the classes are linearly aligned with the orders from the geometrical perspective and one dimension is sufficient to separate the classes well. Figure 2.3 gives a simple geometrical illustration of a two-dimensional dataset coming from four classes showing different ordinal structures, from non-ordinal (nominal), non-

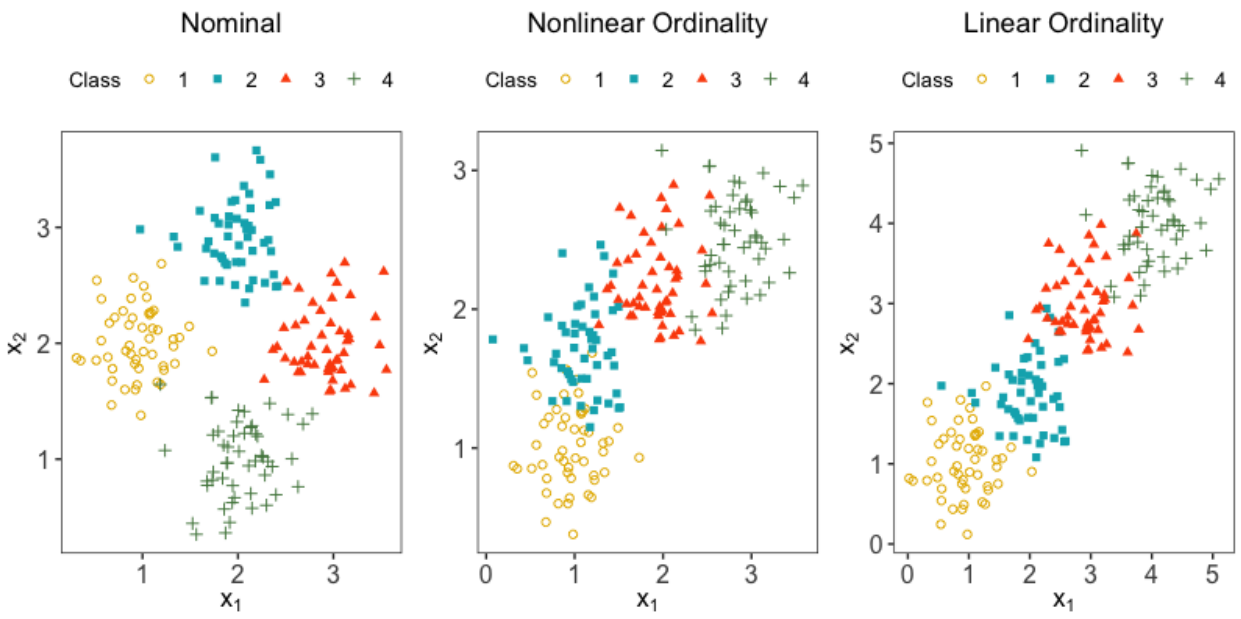


Figure 2.3: Two-dimensional toy data sets from four classes that are nominal, non-linearly ordinal and linearly ordinal, respectively. Data points from the four classes are identified by different colors and shapes. Classes are ordered with the labels ranging from 1 to 4.

linearly ordinal to strict linearly ordinal structures. The left panel shows a nominal case where there is no ordinal information among the classes; the middle panel shows a nonlinear ordinal structure in which the classes are following some ordinality which could not be separated using only one dimension; the right panel displays a strict linear ordinality among the classes, in which the data can be easily separated using one dimension. We reckon that the assumption of a strict ordinality is not realistic because of the complexity of the high-throughput data and such strict linear ordinality may not in accordance with the true ordinality of the data, which might more likely to be a nonlinear ordinality pattern as shown in the middle panel of Figure 2.3. We believe that instead of forcing a strict linear ordinality, the degree of ordinality should be learned from the data.

Given the current challenges of ordinal classification with high-dimensional data, we propose two novel ordinal classification methods that have the following advantages: 1) the true ordinal structure is learned from the data, thus nonlinear ordinality can be detected if it exists; 2) one can visualize the

estimated ordinality by projecting data onto a low-dimensional discriminant space; 3) they are scalable in the sense that it can run with HDLSS data as well as low-dimensional data; 4) they are sparse methods in that the classifier depends only on features relevant to ordinal classification. The summary of the strengths and limitations of each type of existing approaches and our proposed methods is given in Table 2.1.

Table 2.1: Summary of the strengths and weaknesses of existing approaches and the proposed methods.

Methods	Ordinal	High-dimensional	Linearly ordinal assumption
Traditional ordinal methods	Yes	No	Depends
High-dim classifiers	No	Yes	Not applicable
Existing high-dim ordinal methods	Yes	Yes	Yes
Proposed methods	Yes	Yes	No

We employ the concept of 'feature weighting' in machine learning (Cardie & Nowe, 1997) into some regularization in the proposed methods. The idea behind the feature weighting is that the features that are more concordant with the ordinal information should be weighted more compared with those that are less concordant. In the first proposed method, we use the framework of LDA, which has been shown to be an effective framework in HDLSS classification. To deal with the issue of high dimensionality, we add the group Lasso penalty to achieve a sparse solution for better interpretations. In the second proposed method, we incorporate the feature weights into optimal scoring and add the adaptive Lasso idea to achieve a sparse solution.

2.3 Feature-weighted Ordinal Classification

2.3.1 Feature Weighting

We first introduce the concepts of order-concordance and order-discordance. Suppose that the jth variable $\widetilde{\mathbf{x}}_j$ has a mixture distribution with K components with component means $\mu_{1j}, \ldots, \mu_{Kj}$. We call $\widetilde{\mathbf{x}}_j$ order-concordant if the component means are monotonically increasing or decreasing with class labels, i.e., $\mu_{1j} \leq \cdots \leq \mu_{Kj}$ or $\mu_{1j} \geq \cdots \geq \mu_{Kj}$. Otherwise $\widetilde{\mathbf{x}}_j$ is order-discordant. It is naturally to assume that order-concordant variables are more likely to be related to the ordinal information than order-discordant

ones. We propose to use the absolute value of the rank correlation between $\widetilde{\mathbf{x}}_j$ and the class labels as the weight for the jth variable: $w_j = |\text{rank corr}(\widetilde{\mathbf{x}}_j, \mathbf{y})|$. A rank correlation measures the ordinal association between two quantities. Two types of rank correlations are considered in this work, the Spearman's rank correlation and the Kendall's τ coefficient, whose definitions are given below:

Definition 1. *Given two sets of random variables: Z and Q, and the ranked variables (the original variables are ranked based on their values): rgZ and rgQ, the Spearman's rank coefficient between Z and Q is given by:*

$$r_s = \rho_{rgZ,rgQ} = \frac{cov(rgZ, rgQ)}{\sigma_{rgZ}\sigma_{rgQ}},$$

where $\rho_{rgZ,rgQ}$ denotes the Pearson correlation coefficient between rgZ and rgQ, $cov(rgZ, rgQ)$ is the covariance between rgZ and rgQ, σ_{rgZ} and σ_{rgQ} denotes the standard deviation of rgZ and rgQ, respectively.

Definition 2. *Let $(z_1, q_1), (z_2, q_2), \ldots, (z_n, q_n)$ be the sets of observations of the joint random variables Z and Q, the pair (z_{i_1}, q_{i_1}) and (z_{i_2}, q_{i_2}) ($i_1 \neq i_2$) is said to be concordant if $z_{i_1} > z_{i_2}$ and $q_{i_1} > q_{i_2}$ or $z_{i_1} < z_{i_2}$ and $q_{i_1} < q_{i_2}$ holds, and is said to be discordant if $z_{i_1} > z_{i_2}$ and $q_{i_1} < q_{i_2}$ or $z_{i_1} < z_{i_2}$ and $q_{i_1} > q_{i_2}$. Then the Kendall's τ coefficient between Z and Q is given by:*

$$\tau = \frac{1}{C_n^2} \times \{\textit{(number of concordant pairs)} - \textit{(number of discordant pairs)}\},$$

where $C_n^2 = \frac{n(n-1)}{2}$ is the total number of pairs. Note that ties can be considered as either concordant or discordant pairs.

According to Definition 1, Spearman's rank coefficient is designed to measure the monotonic relationship between two sets of variables by measuring the Pearson correlation coefficient between the ranked variables. A Spearman correlation of $+1$ or -1 indicates a perfect monotonic relationship between the two sets of variables. On one hand, a Pearson correlation coefficient of $+1$ (or -1) results a $+1$ (or -1) of Spearman's rank coefficient, on the other hand, a Spearman's rank coefficient of $+1$ (or -1) does not indicate a Pearson correlation of $+1$ (or -1). According to Definition 2, Kendall's τ coefficient is

designed to measure the ordinal association between two quantities, which takes the value between -1 and $+1$. A Kendall's τ coefficient of $+1$ indicates a perfect agreement and -1 indicates the perfect disagreement between the two quantities. A Kendall's τ coefficient of 0 indicates the independence between the two quantities. When there is a perfect monotonic relationship between the two sets of variables, both Spearman's rank correlation and Kendall's τ will be $+1$ or -1, depending on the direction of the association.

Weighting the features provides us with a way to quantify the ordinal information for the variables. It is flexible to incorporate the weights into the framework of standard classification methods. By regularizing the amount of the weights, we could achieve an appropriate degree of ordinality among the classes.

2.3.2 Objective Formulation

In the first proposed method, we propose to incorporate the feature weights into the framework of LDA. As discussed in Section 2.2.2, LDA aims to project the data onto a lower dimensional discriminant subspace such that the projected data are best separated, in terms of achieving a maximum between-class covariance and a minimum within-class covariance. We can rewrite (2.9) so that the objective becomes finding a matrix $B = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d]$, for a given $d \leq K - 1$ that optimizes

$$\max_{B \in R^{p \times d}} \text{trace}(B^T \Sigma_b B), \text{ subject to } B^T \Sigma_w B = I_d,$$

which leads to the following generalized eigenvalue problem (GEP):

$$\Sigma_b B = \Sigma_w B D, \tag{2.12}$$

where D is a diagonal matrix containing the generalized eigenvalues. To solve (2.12), we need to calculate Σ_w^{-1}, which does not exist when $p > n$. In order to solve this singularity issue, the regularized ridge-type LDA (J. H. Friedman, 1989) was proposed to replace $\hat{\Sigma}_w$ by $\hat{\Sigma}_w + \lambda I_p$, where λ is a non-negative regularization parameter. In fact, by adding the term λI_p, the eigenvalues of $\hat{\Sigma}_w$ are adjusted by adding

an amount of λ, which will ensure the positive definiteness of $\hat{\Sigma}_w + \lambda I$ and thus guarantee the non-singularity. We propose a regularization that incorporates the weights by using $\alpha\overline{W}$ instead, where α is a non-negative tuning parameter , $\overline{W}_{p\times p}$ is a diagonal matrix containing $\overline{w}_j = 1 - w_j$ and w_j's are the feature weights defined above. Thus if jth variable is order-concordant, $\overline{w}_j$ is likely to be smaller, it will be less penalized than order-discordant variables. Since $\overline{w}_j$ are all positive, using $\alpha\overline{W}$ will also solve the singularity issue. The objective function of this feature-weighted LDA is given as:

$$\max_{B\in R^{p\times d}} \text{trace}(B^T\Sigma_b B), \text{ subject to } B^T(\Sigma_w + \alpha\overline{W})B = I_d, \tag{2.13}$$

where $\alpha > 0$, whose solution satisfies the following GEP:

$$\Sigma_b B = (\Sigma_w + \alpha\overline{W})BD. \tag{2.14}$$

Once (2.13) is solved, we project the data onto the column space of B, and apply the standard LDA for class assignment. We note that if $\overline{W}$ is replaced by a 'roughness' penalty matrix, this approach can be seen as the penalized linear discriminant analysis by Hastie et al. (1995), who also showed the equivalence to a penalized optimal scoring. Thus, it is clear that the feature-weighted LDA is equivalent to finding the sequence of $\boldsymbol{\beta}$s using the objective:

$$\min_{\boldsymbol{\theta}\in R^K,\, \boldsymbol{\beta}\in R^p} \frac{1}{n}\{\|Y\boldsymbol{\theta} - X\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^T(\alpha n\overline{W})\boldsymbol{\beta}\}, \text{ subject to } \frac{1}{n}\boldsymbol{\theta}^T Y^T Y\boldsymbol{\theta} = 1, \tag{2.15}$$

where Y is an $n \times K$ indicator matrix whose columns corresponds to the dummy-variable coding of the K classes and $\boldsymbol{\theta}$ is the scoring vector in optimal scoring. From (2.15), it can be shown that $\overline{W}$ is actually imposing penalties on $\boldsymbol{\beta}$s, such that smaller $\overline{w}_j$ (corresponding to larger weight) will impose smaller penalties on the coefficients compared with larger $\overline{w}_j$.

When p is large, we would not want a discriminant vector with thousands of non-zero coefficients, therefore, feature selection is essential for the interpret-ability of the results. In order to achieve a sparse

solution, we add a group Lasso penalty (Yuan & Lin, 2006) on the GEP (2.14) so that it becomes a sparse generalized eigenvalue problem. Jung et al. (2019) proposed a framework for sparse GEP and suggested two algorithms to find a solution, namely POI (penalized orthogonal iteration) and fast-POI. Here, we apply the fast-POI algorithm to solve (2.14) with a group Lasso penalty: $p_\lambda(B) = \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\|_2$, with $\mathbf{b}_j$ being the jth row vector of B. The advantage of group Lasso over Lasso is that the former can achieve the sparsity at a group level. That is to say, whether or not a predictor will be dropped out of the model is consistent for all the dimensions in the discriminant subspace.

A sparse estimate of B can be obtained by solving the following:

$$\min_{B \in R^{p \times d}} \text{trace}\{\frac{1}{2} B^T (\Sigma_w + \alpha \overline{W}) B - B^T V\} + \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\|_2, \tag{2.16}$$

where V is a $p \times d$ matrix whose columns are the eigenvectors of Σ_b corresponding to the d largest eigenvalues, and $\lambda > 0$ is a parameter that controls the sparsity. We name this approach the feature-weighted ordinal classification (FWOC).

To solve (2.16), we apply the block coordinate descent algorithm, which updates one coordinate at a time (see Section 2.9.2). The jth row of B, $\mathbf{b}_j$, given $\mathbf{b}_t$ ($t \neq j$), is updated as the following until convergence:

$$\mathbf{b}_j = \frac{1}{s_{jj}} \left(1 - \frac{\lambda}{\|\mathbf{q}_j\|_2}\right)_+ \mathbf{q}_j, \tag{2.17}$$

where s_{jj} is the jth diagonal element of $(\Sigma_w + \alpha \overline{W})$, $\mathbf{q}_j = \mathbf{v}_j - \sum_{t \neq j} b_{jt} \mathbf{b}_t$, $\mathbf{v}_j$ is the jth row vector of V. The summary of the algorithm for FWOC is given in Algorithm 1.

2.3.3 Tuning Parameters

In the following, we discuss the roles of the tuning parameters in the proposed FWOC method. In the actual implementation, we re-parameterize so that $\Sigma_w + \alpha \overline{W}$ is replaced by $r\Sigma_w + (1 - r)\overline{W}$ so that the tuning range of r is bounded within $[0, 1]$. Clearly r controls how much the classifier depends on the ordinal information, in the sense that $r \approx 1$ will yield a solution that is more focused on maximizing the

Algorithm 1: Feature Weighted Ordinal Classification (FWOC)

I. Initialization: Given α, calculate the sample estimate $\hat{\Sigma}_b = \sum_{k=1}^{K} \frac{n_k}{n}(\hat{\mu}_k - \bar{\mathbf{x}})(\hat{\mu}_k - \bar{\mathbf{x}})^T$ and $\hat{\Sigma}_w = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{y_i=k}(\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T$, let $S = \hat{\Sigma}_w + \alpha\overline{W}$;

II. Calculation of β's:

- Initialize : V consists of the first $K-1$ eigenvectors of S, let $B = V$;
- **repeat**

For $j = 1, \ldots, p$;

(1) $\mathbf{q}_j = \mathbf{v}_j - \sum_{t \neq j} b_{jt}\mathbf{b}_t$, in which $\mathbf{v}_j$ is the jth row vector of V, b_{jt} is the (j,t)th element of B, $\mathbf{b}_t$ is the tth row vector of B ;

(2) $\mathbf{b}_j = \frac{1}{s_{jj}}(1 - \frac{\lambda}{\|\mathbf{q}_j\|_2})_+\mathbf{q}_j$, in which s_{jj} is the jth diagnoal element of S;

Update the jth row vector of B as $\mathbf{b}_j$;

until *Convergence or achieve the maximum iterations*;

III. Classification: Project X to the discriminant subspace spaned by the columns of B, i.e., $X_{proj} = XB$, then apply traditional LDA on X_{proj} to achieve the classification rules.

separation of the classes without regard to the ordinality. On the contrary, $r \approx 0$ will yield a solution more dependent on order-concordant variables than discordant ones for classification. In terms of feature selection, when $r = 1$, the variables that being valued more are the ones that contribute more to the separation of classes. In contrast, when $r < 1$, we add more weights to the variables with higher order concordance. Thus, if an order concordant variable will be filtered out because of a small contribution to discrimination, by adding weights, we are decreasing the chance that it is filtered out. Therefore, we propose to learn a good balance between separability and ordinality from the data and the weights would help us determine how much emphasis to put on those order concordant variables for them to 'stand out'.

To demonstrate the idea, we simulate a toy dataset with dimension of 100, in which the observations come from three ordinal classes, with equal sample size of 30 in each class. 40 of the 100 features are set to be signal features, among which there are 20 ordinal (order-concordant) features and 20 nominal (order-discordant) features. Assume that the observations from class k follow the multivariate normal

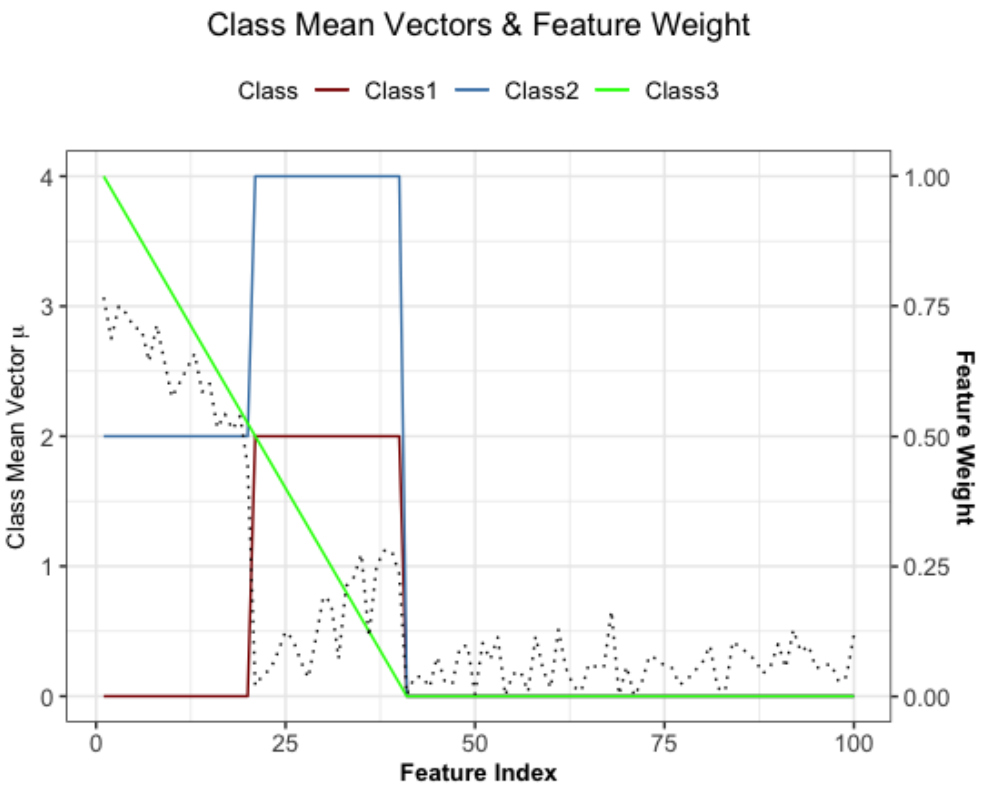


Figure 2.4: Class means for each class and corresponding feature weights for the 3-class dataset. The means for class 1, 2, 3 are colored with red, blue and green, respectively, scaled on the left y-axis. The black dotted line presents the corresponding feature weights, which is scaled on the right y-axis. The first 20 features are ordinal signal features, the 21-40 features are nominal signal features, the 41-100 features are noise features.

distribution $N_{100}(\boldsymbol{\mu}_k, \Sigma)$, where $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,100})^T$ and $\Sigma = I_{100\times 100}$. The first 20 components of the class means corresponding to the first 20 ordinal features follow the order: $\mu_{1,j} \leq \mu_{2,j} \leq \mu_{3,j}$, for $j \in \{1, \ldots, 20\}$; the 20-40 components correspond to the 20 nominal signal features which follow the order: $\mu_{3,j} \leq \mu_{1,j} \leq \mu_{2,j}$, for $j \in \{21, \ldots, 40\}$; the rest of the components of the class means are all zero. Figure 2.4 shows the mean vector $\boldsymbol{\mu}_k$ for each class along with the feature weights calculated based on Kendall's τ. We fix $\lambda = 0.10$ and record the selected features with respect to different values of r. Figure 2.5 shows the number of ordinal signal features selected by FWOC and the proportion of ordinal features selected among all the selected features with respect to different values of r. Clearly, both the

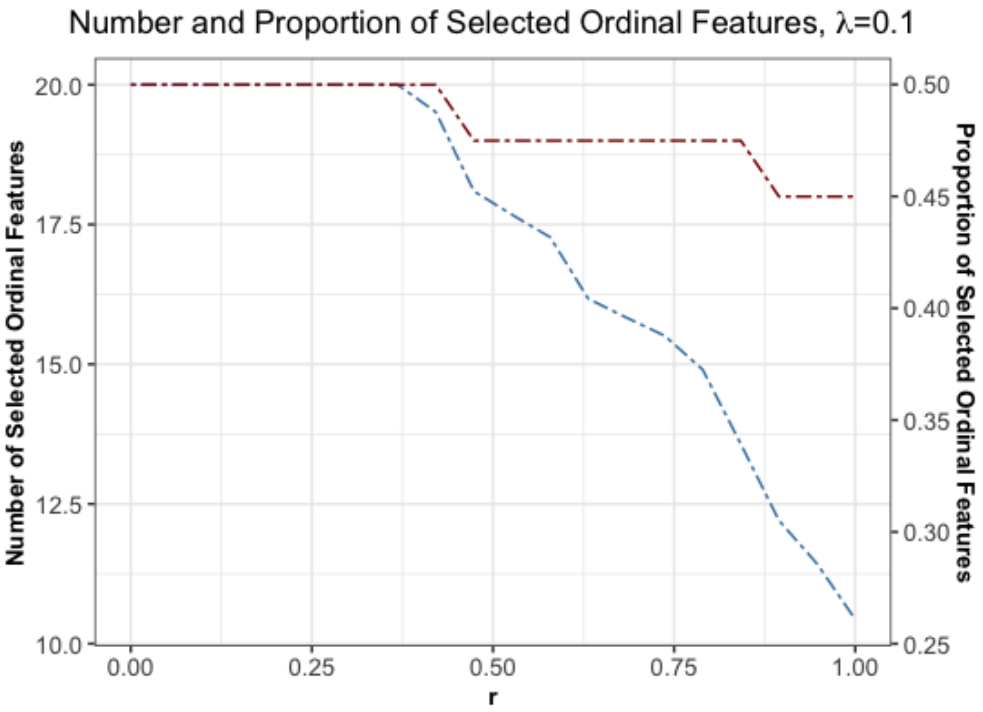


Figure 2.5: The number and proportion of selected ordinal features by FWOC when r changes, given $\lambda = 0.1$. The red dashed line presents the number of selected ordinal features, which is scaled on the left y-axis, the blue dashed line presents the proportion of selected ordinal features, which is scaled on the right y-axis. Both the number and proportion decrease as r increases.

number and the proportion decrease as r increases, indicating that the feature weights can help ordinal features stand out, which is in agreement with our expectations.

Another parameter λ controls the sparsity of the solution. A larger λ imposes a heavier penalty on the solution that would yield a more sparse solution. Note that there is an upper bound of λ that gives a nontrivial solution to (2.16), which is shown to be $\lambda_{max} = \max_{j\in\{1,\dots,p\}} \|\mathbf{v}_j\|_2$, where $\mathbf{v}_j$ is the jth row vector of V.

2.4 Weighted Sparse Discriminant Analysis

Here, we propose another method to incorporate the feature weights into the framework of sparse optimal scoring. It has been shown that optimal scoring is indeed equivalent to LDA subject to a constant factor. What is more, the sparse estimation from optimal scoring (Clemmensen et al., 2011) serve as a high-dimensional solution for classification tasks. However, just like other standard classification methods, the sparse optimal scoring is not designed for ordinal classes and does not take into account the additional ordinal information. Thus, we borrow the idea of adaptive Lasso and add the feature weights when penalizing the coefficients. We refer this proposed method as weighted sparse discriminant analysis (WSDA). The objective function of WSDA can be written as the following.

$$\begin{aligned} \min_{\boldsymbol{\theta}_m \in R^k,\, \boldsymbol{\beta}_m \in R^p} & \|Y\boldsymbol{\theta}_m - X\boldsymbol{\beta}_m\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_{mj}|}{w_j}, \\ \text{subject to } & \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ & \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0,\ \forall\, l < m, \end{aligned} \tag{2.18}$$

where Y is the $n \times K$ dummy matrix defined in optimal scoring, $\boldsymbol{\beta}_m = (\beta_{m1}, \ldots, \beta_{mp})^T$ is the mth discriminant vector, and λ is a complexity parameter which controls the shrinkage on the components of $\boldsymbol{\beta}_m$. The shrinkage depends on the magnitude of the weights w_j. It is obvious that given λ, a feature with smaller w_j tends to have a smaller $|\beta_{mj}|$, that is to say, more penalties are imposed on the features that contain less ordinal information. This criterion is similar to the idea of adaptive Lasso in regression proposed by Zou (2006), in which the coefficients are adjusted by the norm of ordinary least square estimates. In this way, we could differentiate the contribution of features based on how much ordinal information they contain. Note that when λ is zero, (2.18) becomes the original optimal scoring. As λ increases, the penalty on the coefficients will be increased, and the degree of increased penalty depends on the weights of the features.

To solve (2.18), let $\boldsymbol{\beta}_m^* = \frac{1}{\mathbf{w}} * \boldsymbol{\beta}_m$, where $\mathbf{w} = (w_1, w_2, \ldots, w_p)^T$ is the vector whose components are feature weights and $*$ indicates the element-wise multiplication, i.e., $\beta_{mj}^* = \frac{1}{w_j}\beta_{mj}$. Then the objective (2.18) will become:

$$\begin{aligned} & \min_{\boldsymbol{\theta}_m,\, \boldsymbol{\beta}_m} \|Y\boldsymbol{\theta}_m - X(\mathbf{w} * \boldsymbol{\beta}_m^*)\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_{mj}^*|, \\ \Leftrightarrow \quad & \min_{\boldsymbol{\theta}_m,\, \boldsymbol{\beta}_m} \|Y\boldsymbol{\theta}_m - X_w\boldsymbol{\beta}_m^*\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_{mj}^*|, \\ & \text{subject to } \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ & \qquad \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0,\ \forall\, l < m, \end{aligned} \tag{2.19}$$

where $X_w = (\widetilde{\mathbf{x}}_1 * \mathbf{w}, \ldots, \widetilde{\mathbf{x}}_p * \mathbf{w})$, given $\widetilde{\mathbf{x}}_j$ is the jth column vector of X. Therefore, we could modify the original data matrix by multiplying the weights on the features and solve the sparse optimal scoring problem on the modified data matrix X_w. After that, the solution to (2.19) will be transformed back to the original scale, i.e., $\boldsymbol{\beta}_{opt} = \mathbf{w} * \boldsymbol{\beta}_{opt}^*$, where $\boldsymbol{\beta}_{opt}^*$ is the optimal solution to (2.19) and $\boldsymbol{\beta}_{opt}$ is the optimal solution to (2.18).

The proposed (2.19) is closely related to the sparse discriminant analysis (SDA) criterion by Clemmensen et al., 2011. SDA is defined based on the framework of optimal scoring and it finds the mth pair $(\boldsymbol{\theta}_m, \boldsymbol{\beta}_m)$ by solving:

$$\begin{aligned} & \min_{\boldsymbol{\theta}_m, \boldsymbol{\beta}_m} \|Y\boldsymbol{\theta}_m - X\boldsymbol{\beta}_m\|_2^2 + \gamma\boldsymbol{\beta}_m^T\Omega\boldsymbol{\beta}_m + \lambda\|\boldsymbol{\beta}_m\|_1, \\ & \text{subject to } \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ & \qquad \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0,\ \forall\, l < m, \end{aligned}$$

where Ω is a positive definite matrix, γ and λ are tuning parameters. We carry the calculations from Clemmensen et al. (2011) to solve (2.18) using an iteration process.

In (2.19), at mth step, given $\boldsymbol{\theta}_m$, the optimal $\boldsymbol{\beta}_m^*$ solves the Lasso problem: $\min_{\boldsymbol{\beta}_m^*} \|Y\boldsymbol{\theta}_m - X_w\boldsymbol{\beta}_m^*\|_2^2 + \lambda\|\boldsymbol{\beta}_m^*\|_1$, whose solution can be obtained by the coordinate descent algorithm (see details in Section 2.9.2).

At mth step, given $\boldsymbol{\beta}_m^*$, the optimal $\boldsymbol{\theta}_m$ solves the problem:

$$\begin{aligned} &\min_{\boldsymbol{\theta}_m} \|Y\boldsymbol{\theta}_m - X_w\boldsymbol{\beta}_m^*\|_2^2, \\ \Leftrightarrow\ &\min_{\boldsymbol{\theta}_m} L = \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m - 2\boldsymbol{\theta}_m^T Y^T X_w \boldsymbol{\beta}_m^*, \\ \text{subject to } &g(\boldsymbol{\theta}_m) = \frac{1}{n}\boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_m = 1, \\ &h_l(\boldsymbol{\theta}_m) = \boldsymbol{\theta}_m^T Y^T Y \boldsymbol{\theta}_l = 0, \text{ for } l = 0, 1, \ldots, m-1. \end{aligned} \tag{2.20}$$

Let Q_m be the matrix that contains the trivial solution $\boldsymbol{\theta}_0$ and the set of $\boldsymbol{\theta}$s obtained from the previous steps, i.e., $Q_m = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m-1})$, where $\boldsymbol{\theta}_0 = (1, \ldots, 1)^T$. Denote $D = \frac{1}{n}Y^T Y$, then by the method of Lagrange multipliers, solving (2.20) is equivalent to solving the following set of equations:

$$\frac{\partial L}{\partial \boldsymbol{\theta}_m} = \lambda \frac{\partial g}{\partial \boldsymbol{\theta}_m} + \sum_{l=0}^{m-1} \mu_l \frac{\partial h_l}{\partial \boldsymbol{\theta}_m}, \tag{2.21a}$$

$$\boldsymbol{\theta}_m^T D \boldsymbol{\theta}_m = 1, \tag{2.21b}$$

$$\boldsymbol{\theta}_m^T D \boldsymbol{\theta}_l = 0, \text{ for } l = 0, 1, \ldots, m-1, \tag{2.21c}$$

where μ_ls are Lagrange parameters. The equation (2.21a) is equivalent to:

$$nD\boldsymbol{\theta}_m = Y^T X_w \boldsymbol{\beta}_m^* + \lambda D \boldsymbol{\theta}_m + \frac{1}{2}\sum_{l=0}^{m-1} \mu_l D \boldsymbol{\theta}_l. \tag{2.22}$$

Multiplying $\boldsymbol{\theta}_l^T$ on both sides of (2.22), and applying the conditions of (2.21b) and (2.21c) yields:

$$\begin{aligned} (n-\lambda)\boldsymbol{\theta}_l^T D \boldsymbol{\theta}_m &= \boldsymbol{\theta}_l^T Y^T X_w \boldsymbol{\beta}_m^* + \frac{1}{2}\mu_l, \\ \Rightarrow\ 0 &= \boldsymbol{\theta}_l^T Y^T X_w \boldsymbol{\beta}_m^* + \frac{1}{2}\mu_l, \\ \Rightarrow\ \mu_l &= -2\boldsymbol{\theta}_l^T Y^T X_w \boldsymbol{\beta}_m^*. \end{aligned}$$

Algorithm 2: Weighted Sparse Discriminant Analysis (WSDA)

I. Initialization: Let $D = \frac{1}{n} Y^T Y$, $X_w = X * \mathbf{w}$, $Q_1 = (1, \ldots, 1)^T$;

II. Calculation of $\boldsymbol{\beta}$'s: For $m = 1, \ldots, K-1$;

- Initialize $\boldsymbol{\theta}_m = (I_K - Q_m Q_m^T D)\boldsymbol{\theta}_r$, where $\boldsymbol{\theta}_r$ is a random vector with length of K, then apply normalization: $\boldsymbol{\theta}_m = \frac{\boldsymbol{\theta}_m}{\sqrt{\boldsymbol{\theta}_m^T D \boldsymbol{\theta}_m}}$;

- **repeat**

(1) $\boldsymbol{\beta}_m^*$ is the solution to: $\min_{\boldsymbol{\beta}_m^*} \ \|Y\boldsymbol{\theta}_m - X_w \boldsymbol{\beta}_m^*\|_2^2 + \lambda \|\boldsymbol{\beta}_m^*\|_1$;

(2) $\tilde{\boldsymbol{\theta}}_m = (I_K - Q_m Q_m^T D) D^{-1} Y^T X_w \boldsymbol{\beta}_m^*$, then apply normalization: $\boldsymbol{\theta}_m = \tilde{\boldsymbol{\theta}}_m / \sqrt{\tilde{\boldsymbol{\theta}}_m^T D \tilde{\boldsymbol{\theta}}_m}$;

until *Convergence or achieve the maximum iterations*;

- Let $Q_{m+1} = (Q_m : \boldsymbol{\theta}_m)$, and the optimal discriminant vector obtained at mth step is $\boldsymbol{\beta}_{m,opt} = \boldsymbol{\beta}_m^* * \mathbf{w}$.

III. Classification: Project X to the discriminant subspace spanned by the columns of $B = (\boldsymbol{\beta}_{1,opt}, \ldots, \boldsymbol{\beta}_{K-1,opt})$, i.e., $X_{proj} = X B_{opt}$, then apply traditional LDA on X_{proj}, such as the standard LDA to achieve the classification rules.

Then by substituting μ_l back into (2.22) and using the fact that $Q_m Q_m^T = \sum_{l=0}^{m-1} \boldsymbol{\theta}_l \boldsymbol{\theta}_l^T$, we can obtain:

$$
\begin{aligned}
(n - \lambda)\boldsymbol{\theta}_m &= D^{-1} Y^T X_w \boldsymbol{\beta}_m - Q_m Q_m^T Y^T X_w \boldsymbol{\beta}_m^*, \\
\Rightarrow (n - \lambda)\boldsymbol{\theta}_m &= (I - Q_j Q_j^T D) D^{-1} Y^T X_w \boldsymbol{\beta}_m^*.
\end{aligned}
$$

By using the constraints (2.21a), we can calculate the value of λ. Thus, given $\boldsymbol{\beta}_m^*$, the optimal $\boldsymbol{\theta}_m$ is proportional to $(I - Q_m Q_m^T D) D^{-1} Y^T X_w \boldsymbol{\beta}_m^*$, in which $Y^T X_w \boldsymbol{\beta}_m^*$ is the unconstrained solution to the least square problem. The solution to (2.19) is obtained by finding pairs $(\boldsymbol{\beta}_m^*, \boldsymbol{\theta}_m)$ iteratively for each m. As of FWOC, once (2.18) is solved, we project the data onto the column space of B, and apply the traditional LDA for class assignment. The summary of the algorithm of WSDA, as a modified version of the work in Clemmensen et al. (2011), is given in Algorithm 2.

2.5 Model Evaluation Metrics

In this section, we introduce the model evaluation metrics that are appropriate for ordinal classification. Generally, for standard classification problem, a common-used metric is the classification accuracy. Given a test set $\mathcal{T}$ and a classifier $f(\mathbf{x})$, the classification accuracy is defined as :

$$\text{Accuracy}(\mathbf{y}, f(\mathbf{x})) = \frac{\sum_{\mathbf{x}_i \in \mathcal{T}} I(y_i = f(\mathbf{x}_i))}{|\mathcal{T}|},$$

where $I(\cdot)$ is the indicator function and $|\mathcal{T}|$ is the number of samples in the test set. However, the limitation of classification accuracy becomes very obvious when classes are highly imbalanced. For example, if a two-class example includes 95 positive samples and 5 negative samples, the classification accuracy will be 95% even if we predict all the samples to the positive class, in which we actually achieve zero correct prediction in the negative class. Therefore, the classification accuracy is strongly biased for heavily imbalanced classes. There are also other popular metrics such as precision and recall, that can avoid the bias of classification accuracy. In the situation when there is a binary classification problem, the predictions and the actual class labels form a 2×2 contingency table, as given in Table 2.2.

Table 2.2: 2×2 contingency table for binary classification. True positives are samples classified as positive class by the classifier that are also actually positive. False negatives are samples classified as negative class that are actually positive. True negatives and false positives are defined similarly.

	Observation	
Prediction	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Then the precision and recall based on the binary classification are defined as the following.

$$\text{Precision} = \frac{\text{no. of TPs}}{\text{no. of TPs} + \text{no. of FPs}},$$

$$\text{Recall} = \frac{\text{no. of TPs}}{\text{no. of TPs} + \text{no. of FNs}}.$$

A precision score of 1 indicates that every sample that is classified as positive actually belongs to the positive class, a recall score of 1 indicates that every positive sample is classified as positive. Usually, the choice of precision and recall depends on the classification problem. For example, in a task of detecting cancer samples, ('positive' indicates patients with cancer, and 'negative' indicates patients without cancer), we would want to achieve a higher recall so that we will not miss a patient who should receive therapies. In another example of spam email detection ('positive' indicates spam and 'negative' indicates non-spam), we would want to achieve a higher precision so that for every spam email that is detected, it should be actually spam, to avoid putting important emails into the trash bin. A metric that combines the precision and recall is the F score, which is the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Note that the precision and recall was originally defined for binary classification problems. For the multi-class classification problems, we use the class-averaged precision and class-averaged recall which are given in the following:

$$\text{Class-averaged precision} = avg\{\text{precision for class } k\},$$
$$\text{Class-averaged recall} = avg\{\text{recall for class } k\},$$

where the precision (recall) for class k is calculated by taking 'class k versus other classes' as a binary classification problem and $avg\{\cdot\}$ denotes the average values.

In addition, to appropriately measure the performance of multi-class ordinal classification,we also define the weighted cost metric:

$$\text{Weighted cost} = \sum |y_i - f(\mathbf{x}_i)|^d,$$

where $f(\mathbf{x}_i)$ and y_i is the predicted class label and actual class label, respectively, and d is a positive integer. The weighted cost will favor the misclassification to nearer classes compared with the misclassification

to further classes, and as d increases, the differences between these misclassifications will become larger. When $d = 1$, the weighted cost is proportional to the mean absolute error (MAE).

For high-dimensional problems, we also consider sparsity as one measurement of the classifiers. Higher sparsity will increase the interpret-ability of the model. In short, we consider the following measurements as the evaluation metrics for the multi-class ordinal classification problems: (1) Classification accuracy; (2) Kendall's τ; (3) Weighted cost (when $d = 1$); (4) Class-averaged precision; (5) Class-averaged recall; (6) Number of selected features; (7) Proportion of selected signal features among all the selected features (only applied to situations where the signal features are known).

2.6 Simulation Studies

In this section, we carry out a simulation study to measure the empirical performance of the proposed methods and compared with other well-known ordinal classification and nominal classification methods.

2.6.1 Different Ordinal Settings

In the simulations, we simulate moderate high-dimensional datasets, with dimension of 500 and sample size of 150. The samples are assumed to come from four classes with natural ordering: $C_1 \prec C_2 \prec C_3 \prec C4$ and the sample sizes are distributed as $n_1 = 45, n_2 = 35, n_3 = 40, n_4 = 30$. We allow a small fluctuation of the sample size across the four classes but maintain a relative balance among them. The observations from class k are assumed to follow a multivariate normal distribution $N_p(\boldsymbol{\mu}_k, \Sigma_k)$, where $\boldsymbol{\mu}_k$ and Σ_k is the mean vector and covariance matrix for class k, respectively. Among the 500 features, we assume that there are 20 signal variables which contribute to the separation of classes and the rest are noise variables. We further divide the signal features into two categories, one consists of order-concordant variables, i.e., ordinal variables, the other consists of order-discordant variables, i.e., nominal variables. Figure 2.6 shows an example of ordinal variables, nominal variables and noise variables, respectively. Each panel of the figure presents two variables (dimensions) within the same category. In the left panel, the

concentration of the four classes aligns linearly in both x and y axis. In the middle panel, the concentration of the four classes does not follow an order, in which class 4 has smaller values than class 3. There is a clear separation among the classes in both ordinal dimensions and nominal dimensions. In the right panel, all the four classes are overlapped and there is no clear separation, which indicates the noise dimensions.

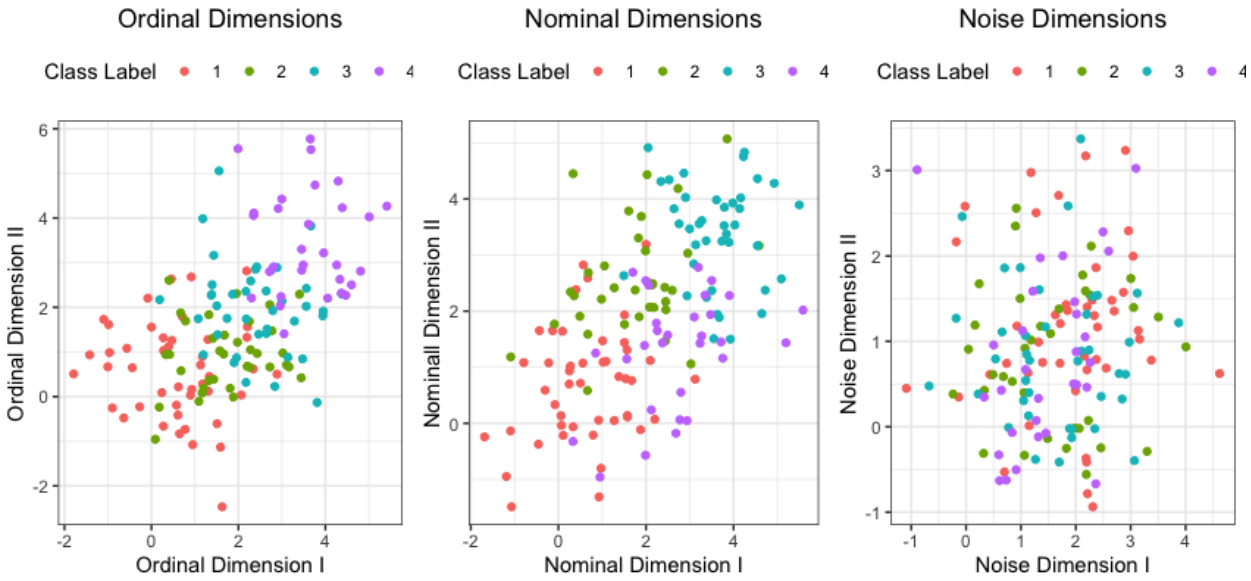


Figure 2.6: 2-Dimensional geometric representation of ordinal dimensions, nominal dimensions and noise dimensions, shown in the left, middle and right panel. Data points from four classes are identified by different colors.

In the following, based on how much ordinality is existing among the classes, we design four different scenarios that are characterized by different mean structures of the signal variables:

- linear ordinality
- nonlinear ordinality
- mixed of ordinal and nominal variables
- nominal situation

The signal variables in the scenario of 'linear ordinality' and 'nonlinear ordinality' are all ordinal variables, and the signal variables in the 'nominal situation' are all nominal variables. Let the mean vector

for class k be $\boldsymbol{\mu}_k = (e * \mathbf{m}_k, \mathbf{0})$, where e is the effect size, $\mathbf{m}_k$ is the mean structure for the signal variables, and $\mathbf{0}$ is a zero vector representing the mean structure for the noise variables. Details about the four scenarios are introduced in the following.

I. Linear Ordinality In the scenario with a linear ordinality, we assume that the mean values for the signal variables are monotonically increasing or decreasing with the class labels. As shown in Figure 2.6, in this case, one dimension would be sufficient to separate the classes. We allow ten increasing ordinal variables and ten decreasing ordinal variables in this scenario. The values of $\mathbf{m}_k$s are shown in Figure 2.7.

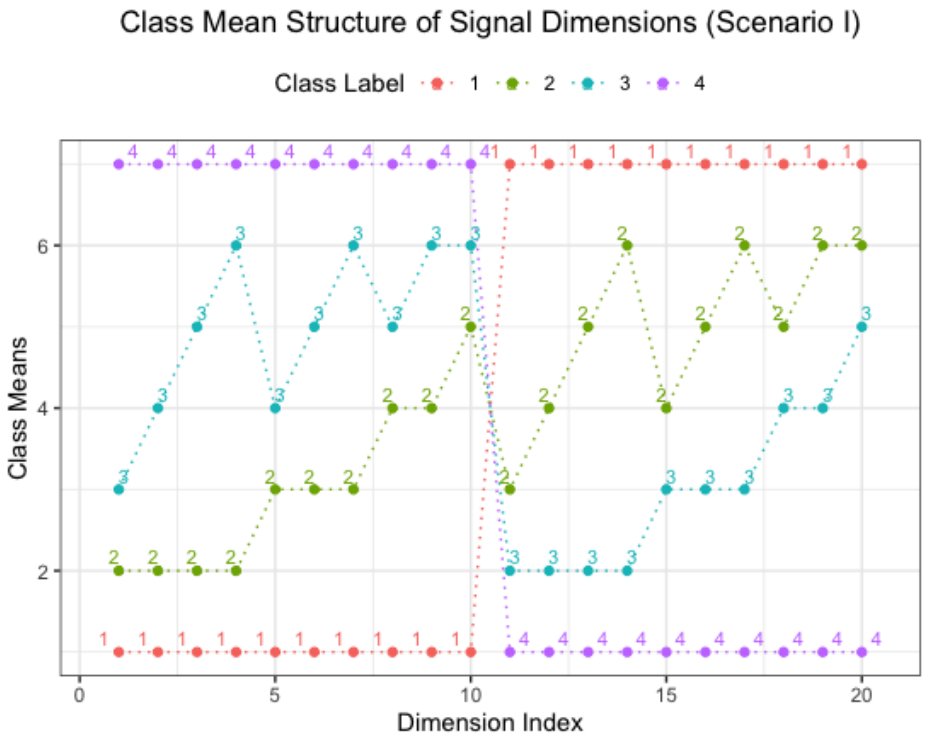


Figure 2.7: The class mean structures for the situation with a linear ordinality. The values are presented in the y-axis, the indexes of the signal variables are presented in the x-axis. Mean structures for the four classes are colored differently. The first ten dimensions represent a increasing trend of means as class labels increase, the last ten dimensions represent a decreasing trend of means. Note that we allow different distances between means from adjacent classes for different dimensions.

II. Nonlinear Ordinality Different from the linear case in which the mean values monotonically increase (decrease) with class labels, we allow the equivalence between the mean values for adjacent classes

in the scenario with a nonlinear ordinality, so that each dimension alone is not sufficient to separate all the four classes. For example, we consider the following order of the mean values for some dimensions: $1 = 2 < 3 < 4$, $1 < 2 = 3 < 4$, and $1 < 2 < 3 = 4$, where the numbers represent the means for the corresponding classes. As shown in Figure 2.6, the data will preserve a nonlinear ordinality in this case. Class mean structures $\mathbf{m}_k$s for the nonlinear case are shown in Figure 2.8.

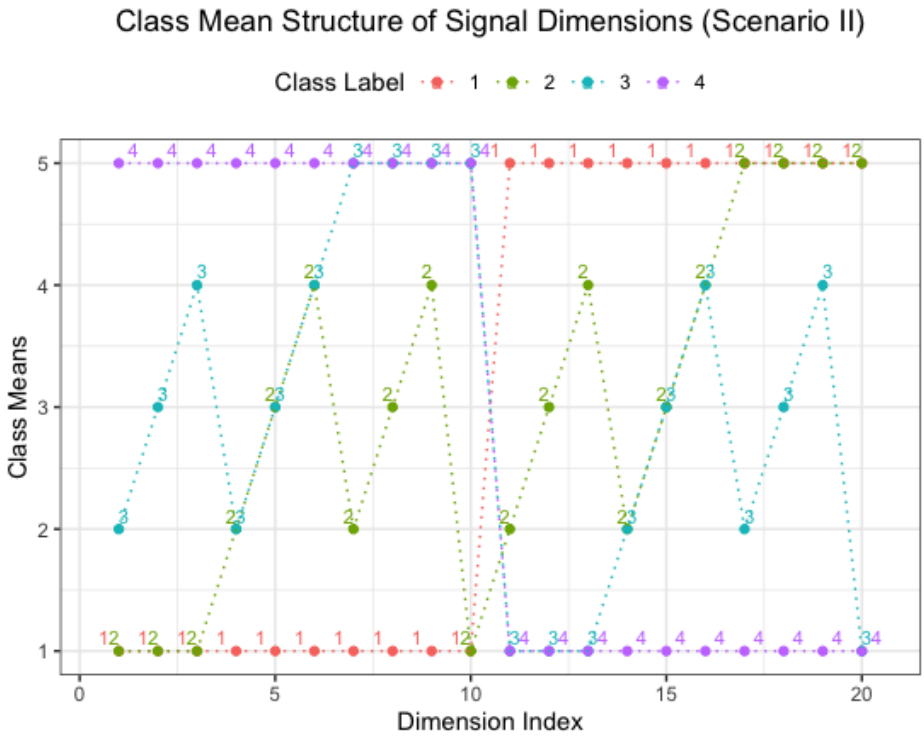


Figure 2.8: As of Figure 2.7, this figure shows the class mean structures for the situation with a nonlinear ordinality. The first ten dimensions present a non-decreasing trend of the means as class labels increase, the last ten dimensions present a non-increasing trend of means as class labels increase.

III. Mixed of Ordinal and Nominal Variables The third scenario is the situation when there is a mixture of ordinal variables and nominal variables. Among the 20 signal variables, there are ten ordinal variables, whose mean values increase monotonically as the class labels increases, the rest ten are nominal variables whose mean values are order-discordant. For example, the order of the mean values for the order-discordant variables could be $1 < 2 < 3 < 4$, where the numbers represent the means for the

corresponding classes. The mean structures for the signal variables for the 'mixed' scenario are shown in Figure 2.9.

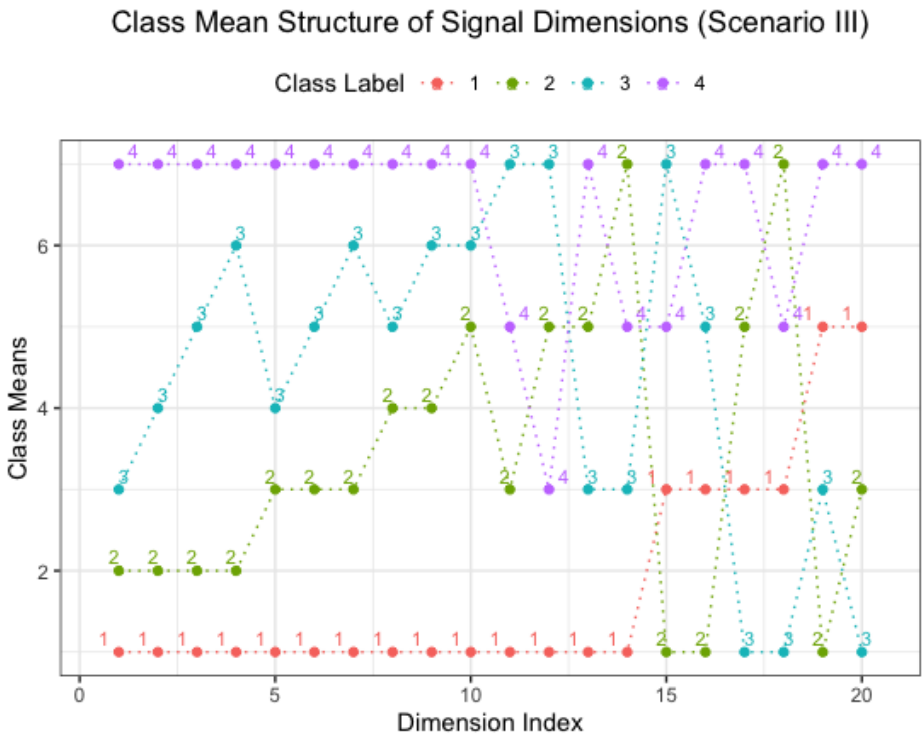


Figure 2.9: The class mean structures of the signal variables in the 'mixed' scenario. The first ten dimensions represent the ordinal variables, the last ten dimensions represent the nominal variables.

IV. Nominal Situation The fourth scenario will represent the nominal case, as shown in Figure 2.6, in which there is no ordinal information among the signal variables. The mean structures for this scenario are shown in Figure 2.10. The nominal situation represents the standard classification problem, in which we want to test whether the performance of ordinal approaches will be hindered and whether the performance of nominal approaches will be stronger.

Effect Size and Correlation Structure When other factors are fixed, the bigger the effect size is, the easier the classes will be classified. The effect size (e) over all the simulations is set to be 0.25. We fix the variance to be 1. For each scenario, we consider three correlation structures: 1) Identity matrix, which indicates that all the variables are independent with each other; 2) Block auto-correlation matrix, in which the

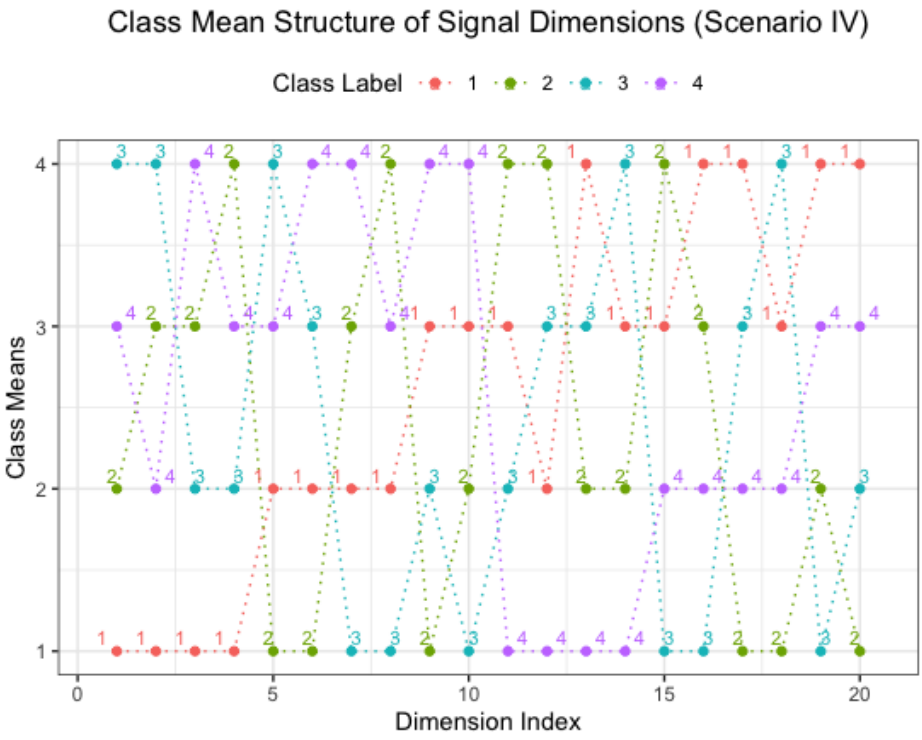


Figure 2.10: As of Figure 2.7, this figure shows the class mean structures for the situation of no ordinality. All the dimensions are order-discordant.

signal variables are assumed to correlated with each other with decreasing correlations, the noise variables are mutually independent, and at the same time the signal variables are independent with the noise variables; 3) Block compound symmetry matrix, in which the signal variables are assumed to correlated with each other with a given correlation coefficient, the noise variables are independent with each other, and at the same time the signal variables are independent with the noise variables. The block auto-correlation matrix Σ_{auto} and block compound symmetry matrix Σ_{cs} are given in the following:

$$\Sigma_{\text{auto}} = \begin{pmatrix} A(\rho_1)_{ns\times ns} & \mathbf{0}_{(p-ns)\times(p-ns)} \\ \mathbf{0}_{(p-ns)\times(ns)} & I_{(p-ns)\times(p-ns)} \end{pmatrix}, \Sigma_{\text{cs}} = \begin{pmatrix} C(\rho_2)_{ns\times ns} & \mathbf{0}_{(p-ns)\times(p-ns)} \\ \mathbf{0}_{(p-ns)\times(ns)} & I_{(p-ns)\times(p-ns)} \end{pmatrix},$$

where $ns = 20$ is the number of signal variables, ρ_1 and ρ_2 are the coefficients of auto-correlation and compound symmetry correlation, respectively. The (i, j)th element of $A(\rho_1)_{ns\times ns}$ is $a_{ij} = \rho_1^{i-j}$, the diagonal elements and off-diagonal elements of $C(\rho_2)_{ns\times ns}$ are 1 and ρ_2, respectively. In the simulations, we set $\rho_1 = 0.9$ and $\rho_2 = 0.7$.

2.6.2 Methods for Comparison and Practical Issues

We consider the following four methods to compare with the proposed FWOC and WSDA: Archer et al. (2014), Zhang et al. (2018), D. M. Witten and Tibshirani (2011), and Clemmensen et al. (2011). Theses methods have been introduced in Section 2.2.2 and 2.2.3. We will refer the first two work as PCRM (penalized continuation ratio model) and BhGLM, respectively. Note that both of PCRM and BhGLM are model-based approaches that assume linearly ordered classes. The third method by D. M. Witten and Tibshirani (2011) is a well-known multi-class sparse LDA that was originally developed for nominal multi-category classification. We call their method PLDA (penalized LDA) in this work. The last one for comparison is the sparse discriminant analysis (SDA) proposed by Clemmensen et al. (2011). Both of PLDA and SDA are projection-based approaches, in which they aim to project the data onto a lower-dimensional discriminant subspace and then apply the standard classification methods (such as LDA) on the projected data to obtain the classification rules. Also, the proposed FWOC and WSDA are projection-based approaches.

In practice, we use the 5-fold cross validation technique to tune the optimal parameters for the methods. In short, it works by splitting the whole training dataset into 5 folds, and taking each fold as the test set and the rest folds as the training set once a time. The process is repeated until each fold has been used as the test set. An illustration of 5-fold cross validation is given in Figure 2.11.

For FWOC, we select the optimal parameters (r_{opt}, λ_{opt}) based on the grid search. The tuning range of r and λ is $(0, 1)$ and $(0, \lambda_{max})$, respectively. The tuning range for WSDA depends on the data. For both FWOC and WSDA, we use the Kendall's τ as the cross-validation metric. Note that for FWOC, there are two tuning parameters, in practice, there might be situations when there are multiple winners

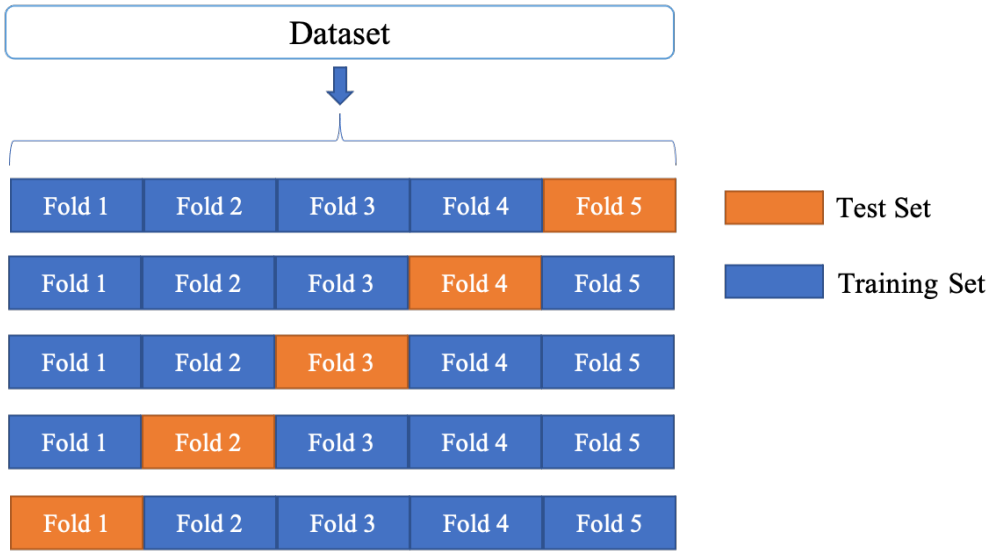


Figure 2.11: The illustration of 5-fold cross validation.

in the grid space. If that happens, we adapt a parsimonious rule of selecting the solution with the largest sparsity and largest ordinality which favors larger λ and smaller r. There are also other options of selecting parameters, such as the one standard error rule mentioned in J. Friedman et al. (2001). For FWOC, we also consider two criteria of selection, in which the first one is giving the priority to largest λ and the second one is giving the priority to smallest r. Figure 2.12 demonstrates the two criteria denoted as c1 and c2, when there are multiple winners in the grid space. For all the projection-based approaches, we use traditional LDA as the post-hoc classifier for projected data. For SDA, we only use the Lasso penalty and vary the tuning parameter in an log-scale grid on some predefined values and select the best tuning parameter based on 5-fold cross validation on classification accuracy. The SDA method is implemented by the R package **sparesLDA** (Clemmensen et al., 2011). For PLDA, we also use the Lasso penalty and use the built-in cross validation function in the R package **penalizedLDA** (D. Witten, 2015) to select the best tuning parameter in a range of $(0.0001, 0.001, 0.01, 0.1, 1, 10)$. PCRM is implemented using the R package **ordinalgmifs** (Archer et al., 2014) with $\epsilon = 0.01$. BhGLM is implemented via the R package

BhGLM (Yi, 2019) and its optimal parameter is selected from an equally 0.01-spaced grid on $[0.05, 0.25]$ via 5-fold cross validation.

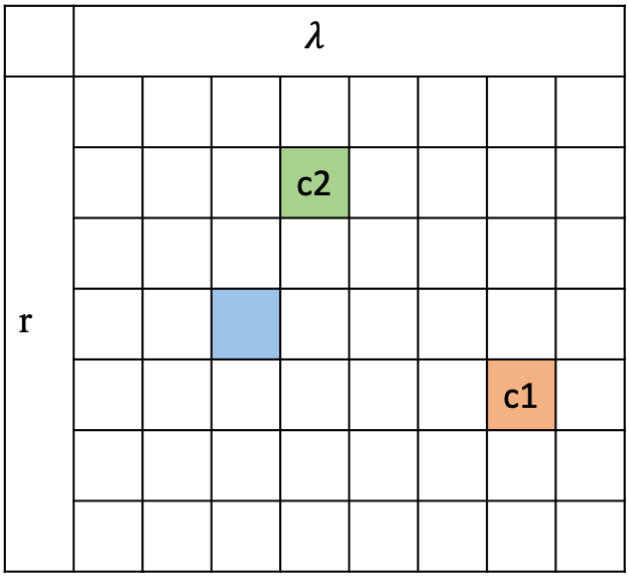


Figure 2.12: Demonstration of the tuning criteria. The figure shows a simple illustration of the grid space spanned by the range of λ and r, each grid represents a 2-dimensional point in the grid space. The figure shows the situation when there are three winners (colored grid) in the grid space with the highest Kendall's τ. Orange grid will be chosen by the c1 criterion, and green grid will be chosen by the c2 criterion.

As discussed in section 2.3.1, there are two rank correlation measurements that could be used by the proposed methods, the Spearman's rank coefficient and the Kendall's τ. Therefore, in the simulation study, to make a comparison between the rank correlations, these two weighting methods will be both incorporated with FWOC and WSDA, denoted as FWOC(sp), WSDA(sp), FWOC(k) and WSDA(k), respectively. In addition, we also include the two criteria for choosing a tuning parameter in case there is a tie in the 5-fold cross validation for FWOC. For example, FWOC(k)(c1) represents the situation when we use Kendall's τ as the weighting measurement and c1 as the tuning criterion. Another variant in FWOC is the number of discriminant vectors, which by default is $K - 1$, where K is the number of classes. In the scenario of linear ordinality, we also allow a reduction in the number of discriminant vectors in FWOC, in other words, we also compare the performance with two discriminant vectors versus that with three discriminant vectors. For example, FWOC(k)(c1)_2D means we use two discriminant vectors with FWOC(k)(c1). In the last three scenarios, we only consider three discriminant vectors.

2.6.3 Simulation Results

In each of the four scenarios, we compare three correlation structures which are mentioned before: block auto-correlation, block compound symmetry and identity. For each simulated dataset, we randomly split it to a training set with 70% observations and a test set with 30% observations. We use the training set for model building and the test set for model assessment. The test results are averaged over 100 repetitions. Results for each scenario are discussed in the following.

Linear Ordinality

The average classification accuracy, Kendall's τ and weighted cost (when $d = 1$) along with the standard deviations are presented in Figure 2.13. In general, when the correlation structure is the block auto-correlation, WSDA(sp) performs the best, with the highest classification accuracy, Kendall's τ and the lowest weighted cost. WSDA(k) is the second best followed by FWOC. PLDA performs significantly worse than all the other methods, and SDA performs the second worst. The performance of BhGLM and PCRM are worse but close to that of FWOC. When the correlation structure is the block compound symmetry or the identity matrix, FWOC and WSDA are still the top performers among all the methods. In addition, the performance of PLDA becomes stronger in these two correlation structures, but BhGLM becomes much more weaker. PLDA performs the best under identity correlation, in which the true correlation is close to the estimated correlation. SDA is still weak under these two correlation structures. Although both of PCRM and BhGLM are model-based approaches, PCRM performs better then BhGLM generally. In term of the variants within FWOC and WSDA, the Spearman's rank works better with WSDA than the Kendall's τ and they work similarly for FWOC. For FWOC, there is no significant difference between the tuning methods, mainly because that the simulated data are not too complicated such that the case when there is a tie among multiple tuning parameters might not happen. Also, using three discriminant vectors is only slightly better than using two discriminant vectors for auto-correlation, in which the data are generally harder to be separated than the other two correlations. The classifiers perform the best under the correlation structure of compound symmetry, in which the correlations among

signal variables are the highest. The results are the most variant in the situation with the auto-correlation covariance.

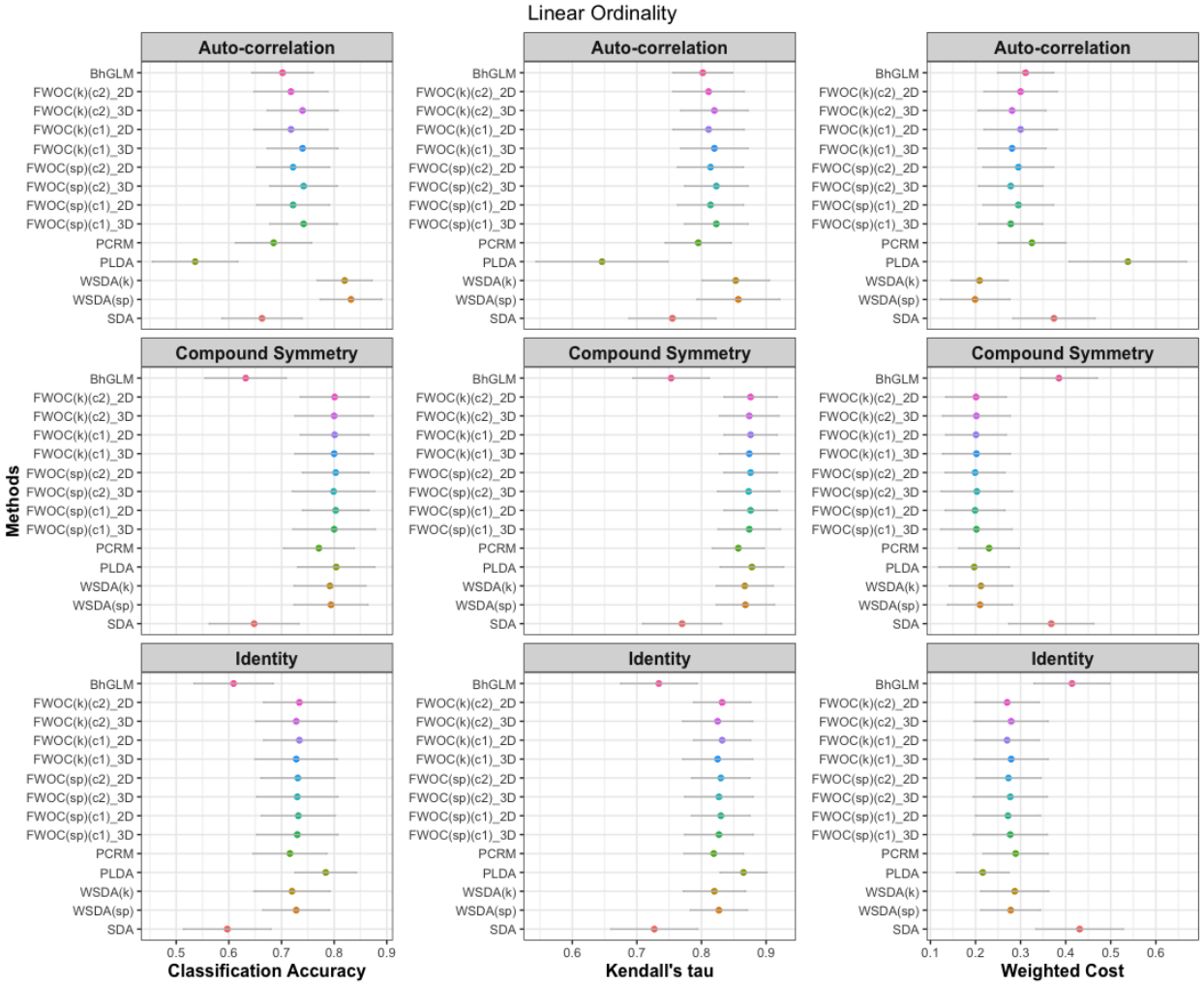


Figure 2.13: The average classification accuracy, Kendall's τ and weighted cost (when $d = 1$) over 100 simulated data sets for the scenario of linear ordinality. Standard deviations are represented by error bars. The three columns show the three metrics, whose values are displayed on the x axis. Different correlation structures under the scenario are presented in the rows.

In addition, we plot the number of selected features by each method, which is shown as the bar graph in Figure 2.14. Along with the bargraph, we also calculate the ratio of signal features among all the selected features, which is represented by the blue lines in Figure 2.14. According to Figure 2.14, BhGLM fails to achieve a sparse solution. Among all the methods, the most sparse solution is achieved by FWOC with the auto-correlation covariance, and WSDA with the compound symmetry and identity correlation

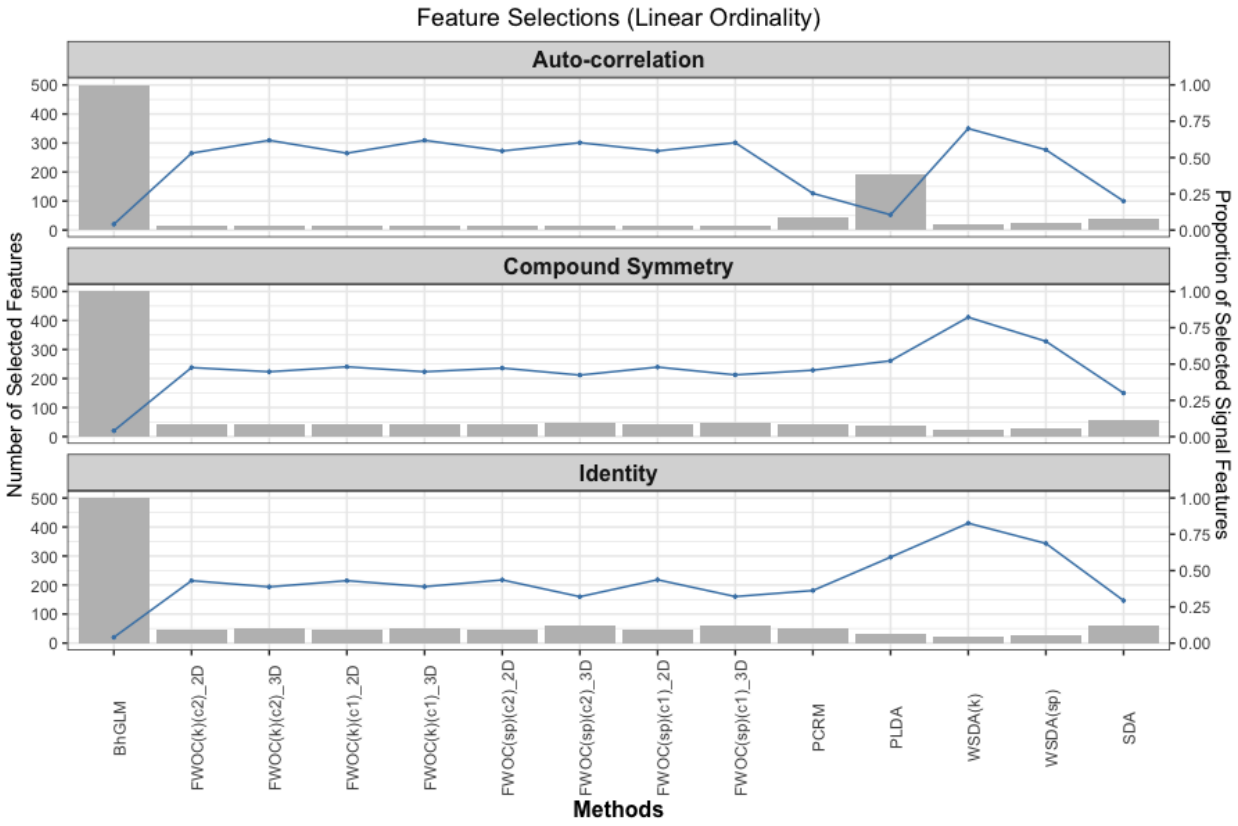


Figure 2.14: The bargraph shows the number of selected features by each method, which is scaled on the left y axis. The line plot shows the ratio of selected signal features among all the selected features, which is scaled on the right y axis. The three rows show the three correlation structures.

structure. In addition, WSDA achieves the highest signal ratio among all these methods. In comparison, the performance of SDA is much weaker than WSDA in the linear case. Above all, both FWOC and WSDA have a very strong performance in the scenario with a linear ordinality. BhGLM does not perform well might due to the reason that it does not achieve a sparse solution. The reason that PLDA and SDA are not competitive might be that they do not take into account of the strong ordinal information in this scenario.

Nonlinear Ordinality

The results in the scenario with a nonlinear ordinality are shown in Figure 2.15.

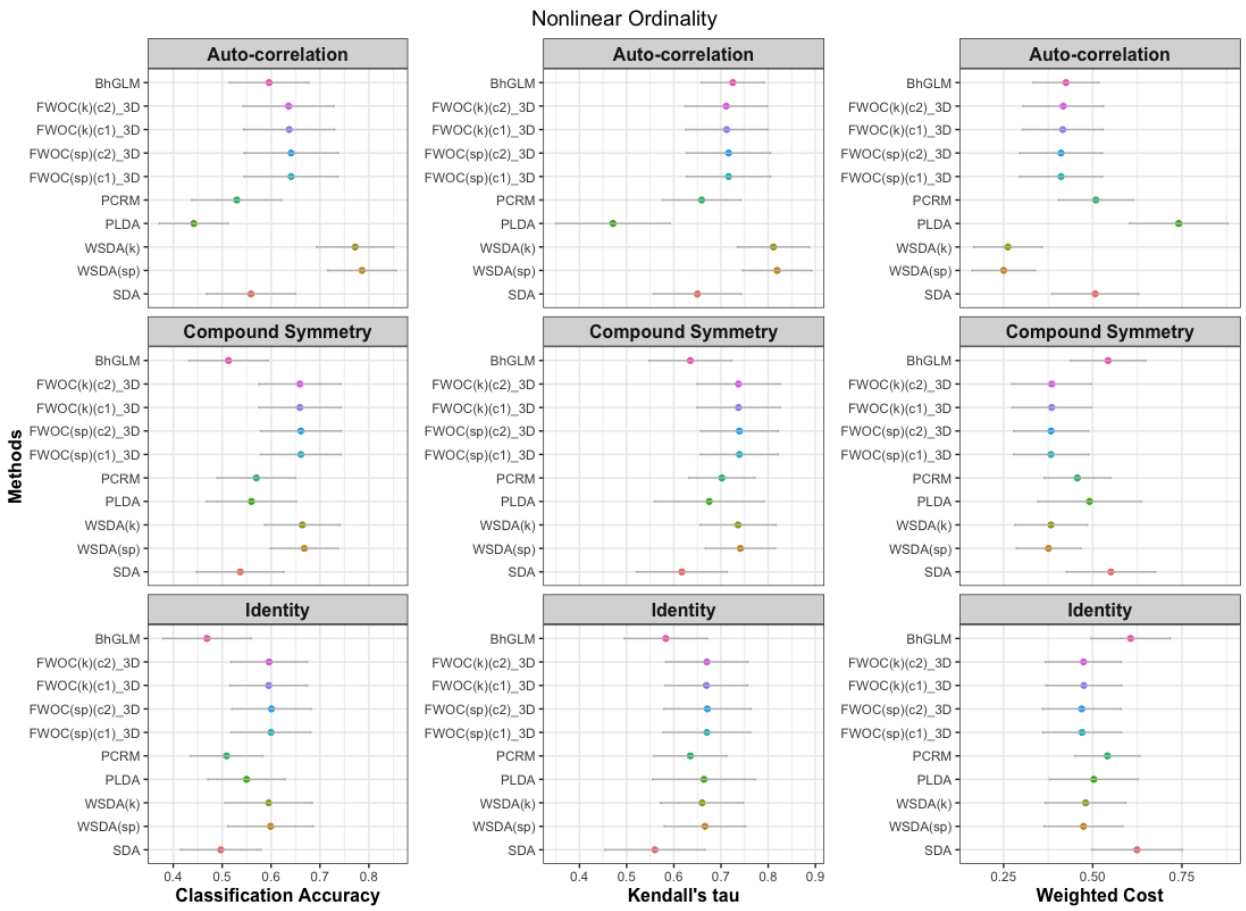


Figure 2.15: As in Figure 2.13, the x axis shows average classification accuracy, Kendall's τ and weighted cost (when $d = 1$) over 100 simulated data sets in the scenario of nonlinear ordinality. Different correlation structures are presented by rows.

For all the three correlation structures, WSDA(sp) performs the best, followed by WSDA(k). FWOC is always the second strongest method. Also, with the correlation structure of compound symmetry and identity, FWOC performs very competitive to WSDA. BhGLM performs better with auto-correlation covariance compared with the other two correlations. In addition, the performance of PCRM lies in the middle tier. Both of PLDA and SDA have a weak performance with all three correlation structures. In the scenario of nonlinear ordinality, FWOC and WSDA have far more better performance than the other methods.

Figure 2.16 shows the number of selected features and ratio of signal features. The pattern of signal ratios are very similar across the three different correlation structures, in which WSDA(k) achieves the highest ratio, followed by WSDA(sp). Again, BhGLM fails with feature selection. Among the rest of the methods, the most sparse one is achieved by WSDA(k). SDA does a good job in achieving a sparse solution, but selects less signal features. The signal ratios of PLDA and SDA are relatively low, which might explain why they have a weak performance in terms of the metrics in Figure 2.19. The reason that BhGLM and PCRM do not perform well might be that one dimension is not sufficient to separate the classes in this case. The reason that PLDA and SDA are neither promising might be that they do not take into account of the ordinal information, similar as the reason for that in the scenario of linear ordinality.

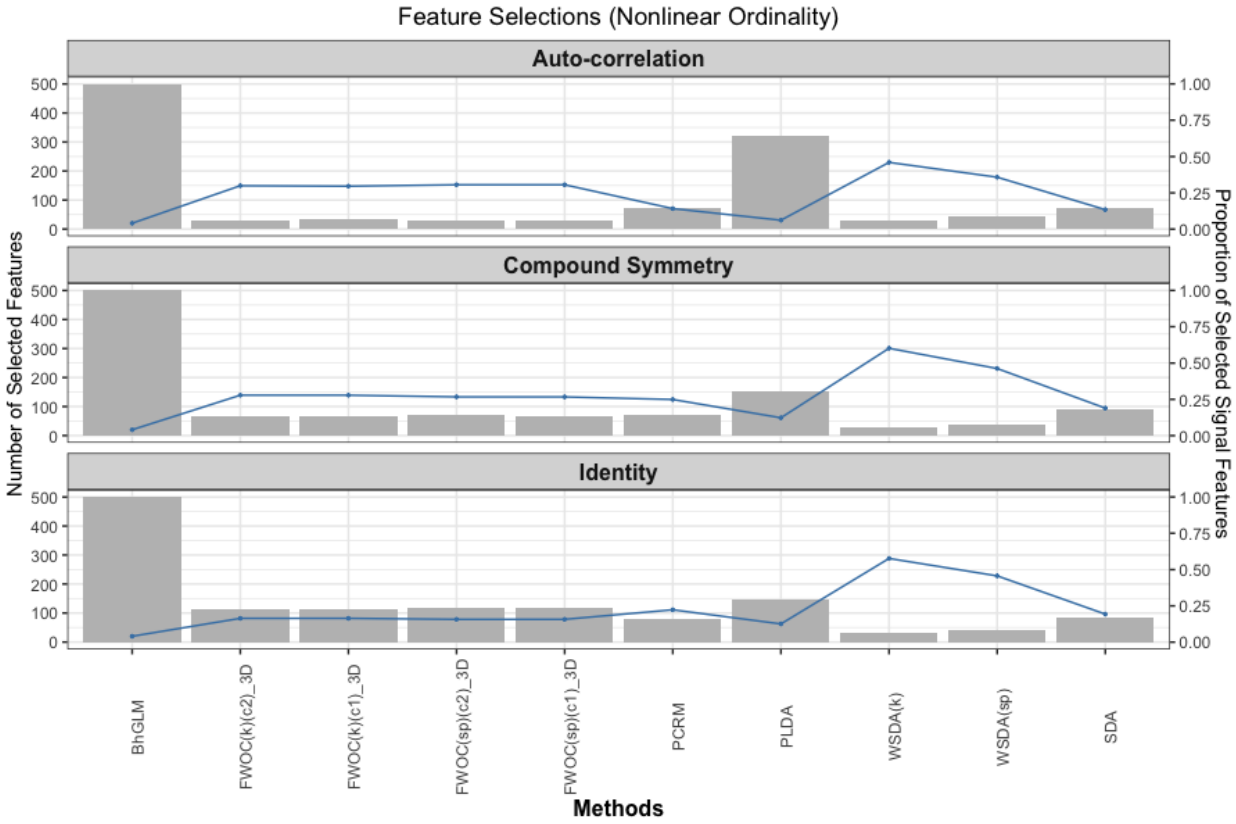


Figure 2.16: As in Figure 2.14, the bargraph shows the number of selected features and the ratio of selected signal features among all the selected features, with scales on the left and right y axis, respectively.

Mixed of Ordinal and Nominal Variables

Results for the 'mixed' scenario are presented in Figure 2.17 and Figure 2.18.

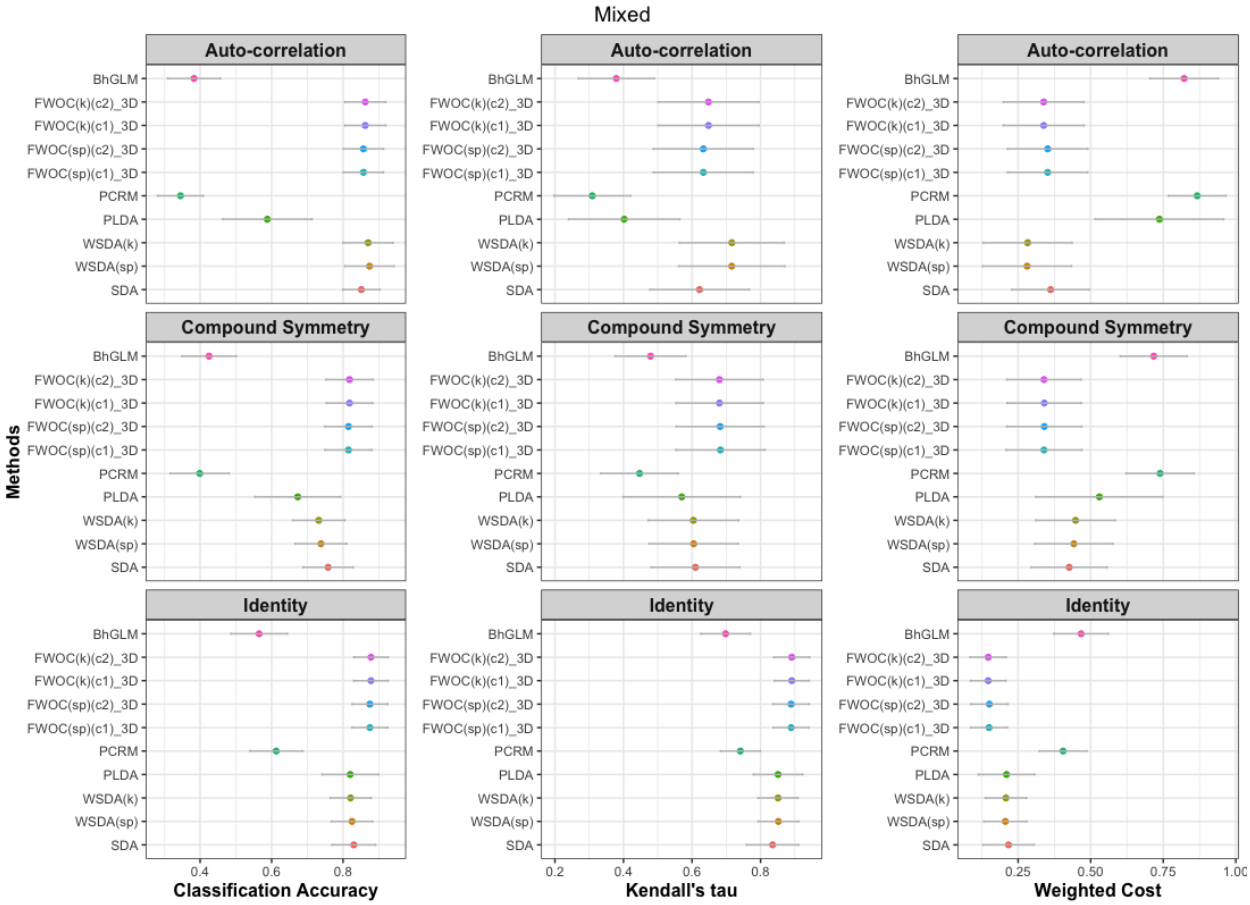


Figure 2.17: As in Figure 2.13, the x axis shows the average classification accuracy, Kendall's τ and weighted cost (when $d = 1$) over 100 simulated data sets in the 'mixed' scenario.

The top methods in this scenario are FWOC, WSDA and SDA. To be more specific, WSDA performs the best with the auto-correlation covariance, FWOC performs the best with the compound symmetry and identity correlation structures. SDA has a close performance with WSDA with the identity covariance. Note that although sometimes SDA has a higher accuracy than WSDA, it achieves a lower Kendall's τ than WSDA. The performances of BhGLM and PCRM are weak in this scenario. PLDA is also not strong and its performance lies in the middle tier.

In terms of feature selection, we are not only interested in the signal ratio but also interested in the ratio of ordinal variables among all the selected variables. Figure 2.18 presents the number of selected features, signal ratio and ordinal ratio. In terms of sparsity, WSDA, FWOC and SDA perform comparably well, in which the most sparse solution is achieved by WSDA(k). Also, the highest signal ratio and ordinal ratio are both achieved by WSDA(k) . SDA performs slightly better than FWOC in term of the two ratios. BhGLM still does not achieve a sparse solution. The ordinal ratio for PCRM is relatively high compared to its signal ratio, which indicates that PCRM does a good job in selecting ordinal variables.

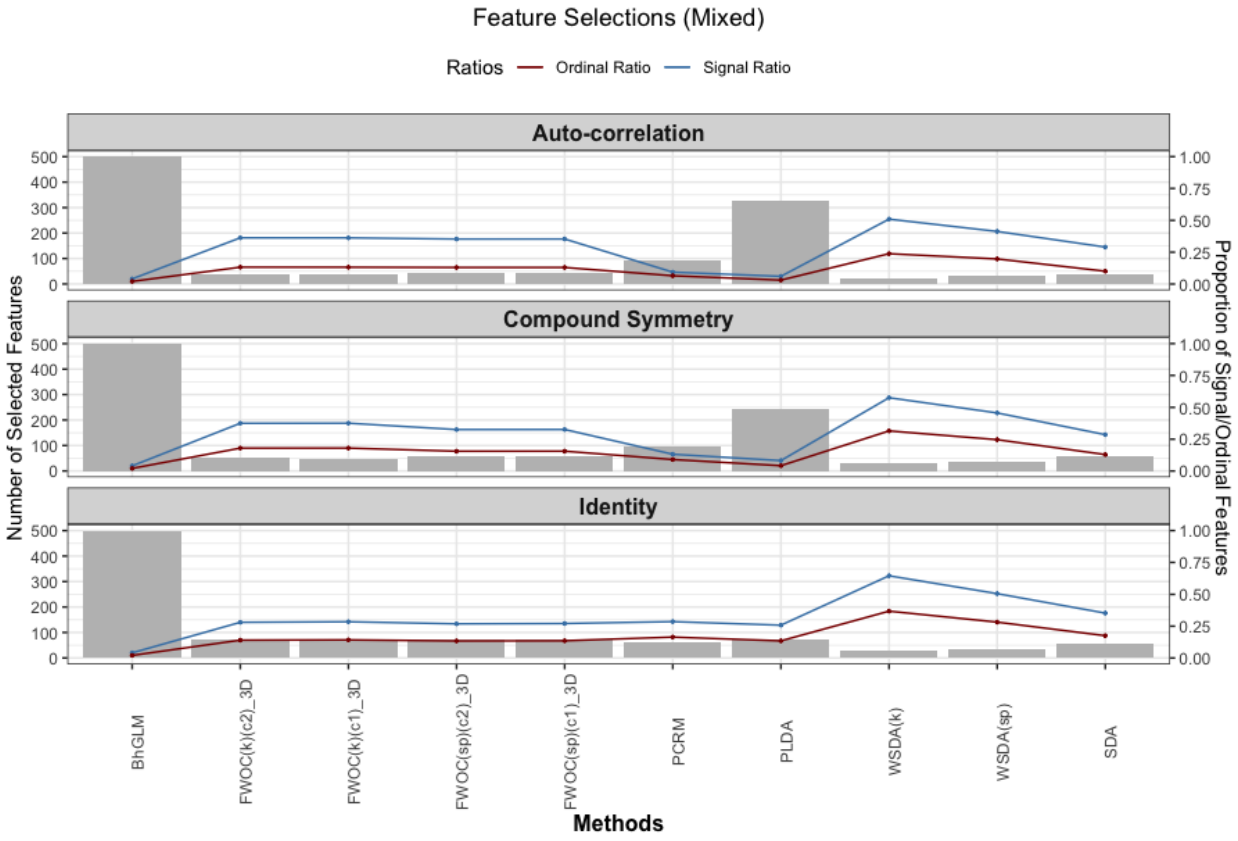


Figure 2.18: The bargraph shows the number of selected features, the ratio of signal features and the ratio of ordinal features among all the selected features. The scales of the numbers are on the left y axis, and the scales of the ratios are on the right y axis. The blue line represents the signal ratio and the red line represents the ordinal ratio.

Nominal Situation

Results of the quantitative metrics and feature selection the nominal scenario are shown in Figure 2.19 and Figure 2.20, respectively.

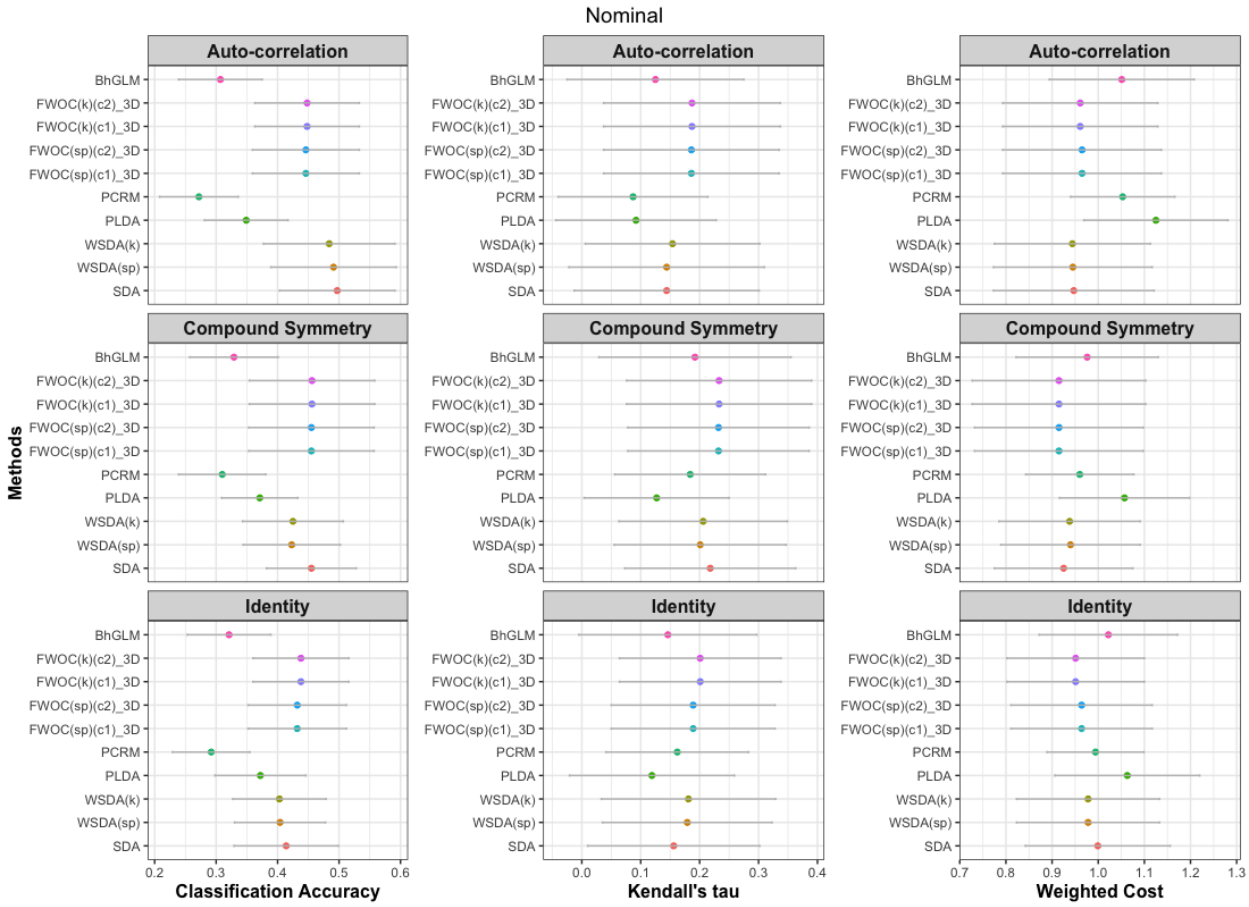


Figure 2.19: As in Figure 2.13, the x axis shows the average classification accuracy, Kendall's τ and weighted cost (when $d = 1$) over 100 simulated data sets in the nominal scenario.

The top methods are still FWOC, WSDA and SDA. FWOC is the strongest with the compound symmetry and identity covariance, and it also always achieves the highest Kendall's τ. SDA performs very competitive, especially in obtaining a high accuracy with the auto-correlation and compound symmetry covariance. However, in terms of Kendall's τ and weighted cost, SDA does not beat WSDA and FWOC. PLDA is still in the middle tier with respect to accuracy, and it is the weakest with respect to Kendall's τ

and weighted cost. Also, neither of BhGLM and PCRM performs well, due to the reason that the scope of these classifiers can not be extended to the standard classification problems.

According to Figure 2.14, WSDA is still the best in terms of having a most sparse solution with the three covariance structures. However, the advantage of WSDA in this scenario is not that obvious compared with previous scenarios. The signal ratios achieved by FWOC and SDA are not far more smaller than that of WSDA.

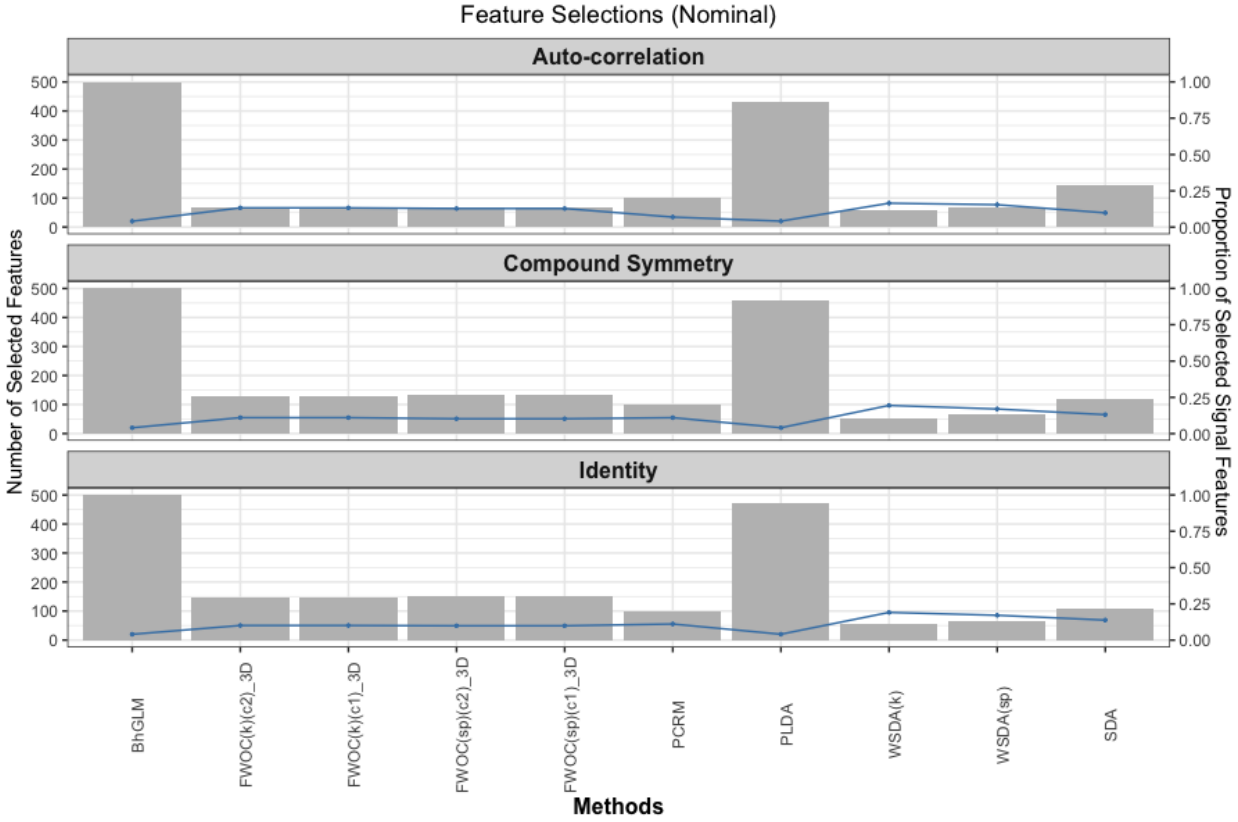


Figure 2.20: As in Figure 2.14, the bargraph shows the number of selected features and the ratio of selected signal features among all the selected features, with scales on the left and right y axis, respectively.

Based on the simulation results, the four scenarios, if ranked by the complexity from the highest to the lowest, would follow: nominal $\succ$ nonlinear $\succ$ mixed $\succ$ linear. Table 2.3 shows a summary of the performances of the methods with respect to the metrics shown in Figure 2.13, 2.15, 2.17 and 2.19. The performances are categorized as 'Strong' ,'Middle' and 'Weak' subjectively. Among the methods, WSDA generally performs the best for most of the times, not only in achieving the highest accuracy (or Kendall's

Table 2.3: Summary of the performances of the methods in different scenarios in the simulation study. 'CS' is short for compound symmetry correlation.

	Linear			Nonlinear		
	Auto-correlation	CS	Identity	Auto-correlation	CS	Identity
BhGLM	Middle	Weak	Weak	Middle	Weak	Weak
FWOC	Middle	Strong	Strong	Middle	Strong	Strong
PCRM	Middle	Strong	Strong	Middle	Middle	Weak
PLDA	Weak	Strong	Strong	Weak	Middle	Middle
WSDA	Strong	Strong	Strong	Strong	Strong	Strong
SDA	Middle	Weak	Weak	Middle	Weak	Weak
	Mixed			Nominal		
	Auto-correlation	CS	Identity	Auto-correlation	CS	Identity
BhGLM	Weak	Weak	Weak	Weak	Weak	Weak
FWOC	Strong	Strong	Strong	Strong	Strong	Strong
PCRM	Weak	Weak	Weak	Weak	Weak	Weak
PLDA	Middle	Middle	Strong	Middle	Middle	Middle
WSDA	Strong	Strong	Strong	Strong	Strong	Strong
SDA	Strong	Strong	Strong	Strong	Strong	Strong

τ and weighted cost) but also achieving the most sparse solution with the highest signal ratio. FWOC is also one of the strongest performers. Comparably, the performance of PLDA always lies in the middle tier, which is better than the model-based approaches but worse than WSDA and FWOC. Interestingly, in the 'linear+identity' and 'linear+compound symmetry' case, PLDA performs very good, probably due to the reason that data are much more easier to be separated in these cases. In addition, SDA performs not well with the case of linear ordinality and nonlinear ordinality, but works very well in the 'mixed' and nominal scenario.

As WSDA can be considered as an 'ordinal' modification version of SDA, the comparison of WSDA and SDA becomes very interesting. The simulation results demonstrate that WSDA works better than SDA when there is some ordinality within the dataset and the advantages of WSDA disappears when the data are purely nominal. The comparison between SDA and WSDA confirms that WSDA is more suitable for ordinal cases, and SDA is more suitable for nominal cases.

Furthermore, it is also interesting to discuss about the performance of the model-based approaches, BhGLM and PCRM. These model-based approaches make the assumption that the classes are linearly aligned with the orders, which means that one dimension is sufficient enough to separate the classes. Among the four scenarios, the performances of these model-based approaches are only acceptable in the 'linear ordinality' case. BhGLM and PCRM perform not well in other cases, which is not beyond our expectations. In terms of feature selection, PCRM perform well in achieving a sparse solution, but BhGLM fails with feature selection. In practice, data are complicated such that the classes have a high probability of being non-linearly aligned, in which these model-based approaches might fail. In terms of feature selection, both of WSDA and FWOC work better than the other methods, indicating that the incorporation of weights will help us select the ordinal features.

The simulation results also show that different rank correlation measurements have different impacts on the proposed methods. To be more specific, with Spearman's rank coefficient, WSDA can achieve a better classifier performance than with Kendall's τ; with Kendall's τ, WSDA does better in feature selection. However, FWOC is not sensitive with the two rank correlations in the simulations. In addition, although we consider two different tuning criteria for FWOC, it seems that the results are not significantly different. It is probably because that the simulated data are not that complex enough so that there might not be multiple winners in the grid space when tuning the parameters. Through simulations, it is also interesting to see that the two metrics, Kendall's τ and weighted cost ($d = 1$) have some symmetry between them, and the optimal methods indicated by different metrics are consistent for most of the times (The results of class-averaged precision and recall are presented in Table 2.4, 2.5, 2.6, and 2.7 in Section 2.9.3). This is also because that our simulated data are relatively balanced.

Above all, both FWOC and WSDA perform very well under the ordinal scenarios and they also show a strong performance in nominal scenarios, indicating that the proposed methods are capable of achieving a balance between the class separation and ordinality.

2.7 Classifications with Tumor Grades and Drug Responses

High-dimensional ordinal data are not uncommon in real world applications, especially in the field of biomedical research. Clinical responses and tumor-node-metastasis (TNM) stages are always categorized as ordinal groups. For example, the clinical responses to a treatment could be categorized as 'Complete Response', 'Partial Response', 'Minimum Response', 'No Change' and 'Progressive Disease' (BladÉ et al., 1998), depending on the degree of reaction. In addition, tumor grades are often grouped as 'Stage 0', 'Stage I', 'Stage II', 'Stage III' and 'Stage IV', depending on the severity of the disease. With the development of modern technologies, multiple high-throughput platforms are providing a large repository of data, such as the gene expression data, to facilitate the biomedical research. Among various types of biomarkers, many studies have shown that gene expressions are very powerful predictive features in predicting clinical responses and tumor grades. In this section, we test our high-dimensional ordinal classification methods FWOC and WSDA using four datasets from the biomedical area. Among the four datasets, two of them are associated with the predictions of tumor grades and the other two are related with the predictions of clinical responses. A brief description of these datasets is given in the following.

I. B-cell Acute Lymphoblastic Leukemia Data This dataset consists of a subset of the observations from the acute lymphoblastic leukemia (ALL) dataset (Chiaretti et al., 2004) which includes micro-arrays from 128 patients with acute lymphoblastic leukemia from the Ritz laboratory. ALL is a type of cancer that is due to the abnormalities of bone marrow cells and is the most common childhood cancer (Hunger & Mullighan, 2015). The whole dataset is available in the R package **ALL** (Li, 2019) and we consider the subset of it which includes the 90 patients who are in the group of B-cell ALL. Their tumor stages information lies in one of the four ordinal classes: B1, B2, B3, B4. This dataset has been normalized using RMA. The classes are ordered such that: $B1 \prec B2 \prec B3 \prec B4$.

II. Primary Human Glioma Data This dataset was originally from the study of L. Sun et al. (2006), on a research of the evaluation of SCF expression in primary human gliomas. Glioma is a type of cancer

that occurs in the brain and spinal cord. The gene expression dataset was generated using the Affymetrix HG-U133 Plus 2.0 Array and can be assessed from the Gene Expression Omnibus (GEO) database with accession number GDS1962. It consists of mRNA expressions of 157 primary human gliomas (Grade II, III, IV) and 23 non-tumor human brain samples. The classes are ordered such that: Normal $\prec$ Grade II $\prec$ Grade III $\prec$ Grade IV.

III. Multiple Myeloma with VTD Induction Therapy Data This dataset was generated from the study of Terragna et al. (2016) focusing on the prediction of complete response using genetic information for multiple myeloma (MM) patients who received the induction therapy. MM is a type of cancer which is characterized by the proliferation of bone marrow of plasma cells (Terragna et al., 2016). Like other cancers, genetic abnormalities play an essential role in the acquisition of MM. Although it is a relatively uncommon cancer, the overall 5-year survival rate for people with MM is not high. Modern treatments such as induction, consolidation and maintenance therapies for MM have emerged over the years (Terragna et al., 2016). However, the prognosis of MM still remains variable, partly due to the heterogeneity of patients' response to the treatments. The dataset can be assessed from the GEO database with accession number GSE68871. It was generated using the Affymetrix HG-U133 Plus 2.0 Array and consists of 118 primary tumor cell samples obtained from new MM patients who received the Bortezomib-Thalidomide-Dexamethasone (VTD) induction therapy. The clinical outcome in this study is a five-class ordinal outcome: Complete Response (CR), Near Complete Response (NCR), Very Good Partial Response (VGPR), Partial Response (PR) and Stable Disease (SD). The order of the classes follows CR $\prec$ NCR $\prec$ VGPR $\prec$ PR $\prec$ SD.

IV. Multiple Myeloma with Bortezomib Data This dataset was originally from a study by Mulligan et al. (2007), on the relationship between gene expression profiling and the treatment outcomes of proteasome inhibitor Bortezomib, for patients with MM. In the study of Mulligan et al. (2007), they aimed to predict the complete response of treatment with Bortezomib for MM patients. The dataset can be assessed from the GEO database with accession number GSE9782. It was generated using the Affymetrix HG-U133 A/B platform and consists of 169 pre-treated tumor cell samples from the patients with relapsed myeloma

who were enrolled in the phase 2 and phase 3 clinical trials of Bortezomib. The outcomes in this dataset are categorized as Complete Response (CR), Partial Response (PR), Minimal Response(MR), No Change (NC) and Progression Disease (PD), according to the European Group for Bone Marrow Transplantation criterion (BladÉ et al., 1998). The order of the classes follows: CR $\prec$ PR $\prec$ MR $\prec$ NC $\prec$ PD.

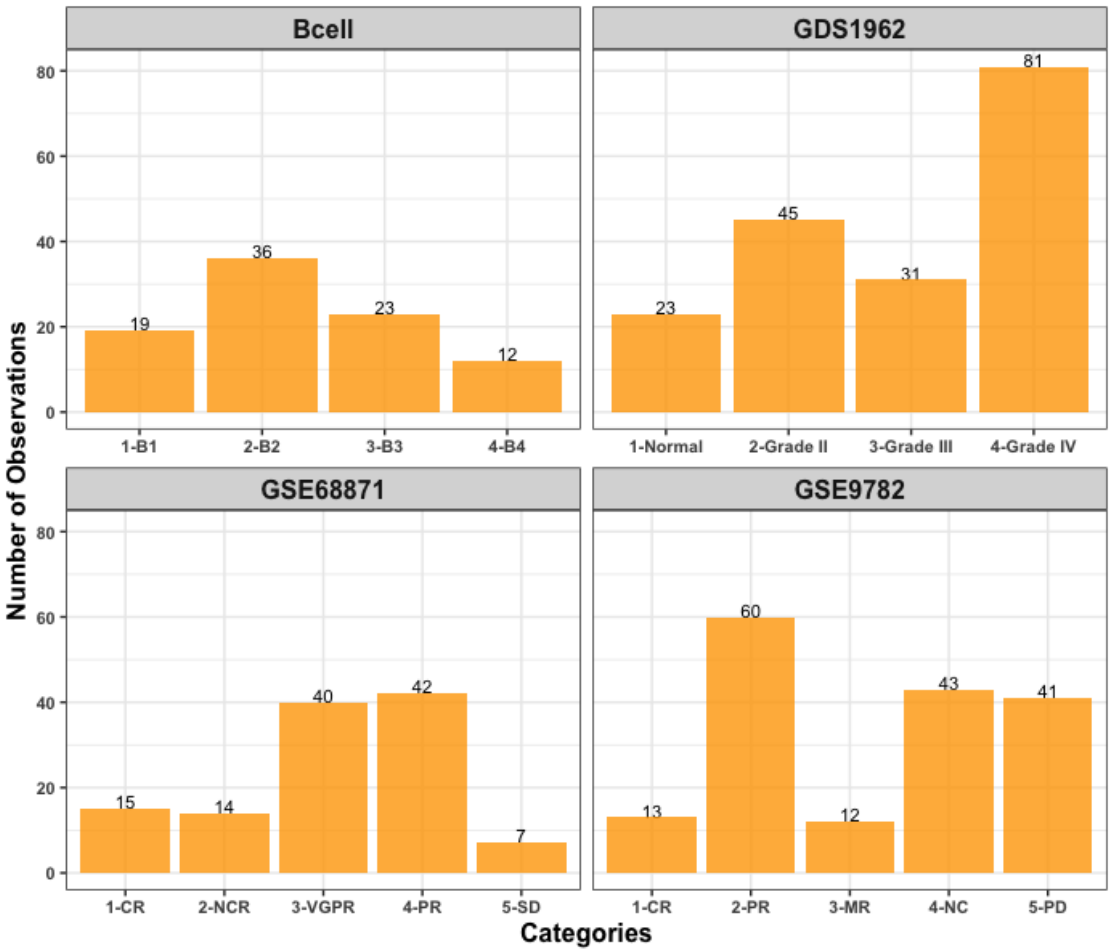


Figure 2.21: The bargraph shows the number of observations in the ordinal categories for the four datasets.

A summary of the four datasets is given in Figure 2.21. Both of the two tumor-grade datasets have four ordinal outcomes and both of the two treatment-response datasets have five ordinal outcomes. Note that the number of samples are not balanced in the datasets.

For each of the datasets, we randomly split it into a training set with 70% observations and a test set with 30% observations and repeat the random split for ten times. In each repetition, we pre-screen 500

probes using the univariate ordinal logistic regression model by Zhang et al. (2018) with the training set. To be specific, for the jth probe, an ordinal logistic regression model is fitted:

$$P(y_i = k) = \begin{cases} 1 - logit^{-1}(\phi x_{ij} + \beta_{1j}), \text{ for } k = 1, \\ logit^{-1}(\phi x_{ij} + \beta_{(k-1)j}) - logit^{-1}(\phi x_{ij} + \beta_{kj}), \text{ for } 1 < k < K, \\ logit^{-1}(\phi x_{ij} + \beta_{kj}), \text{ for } k = K, \end{cases}$$

where $logit^{-1}(z) = \frac{1}{1+exp(-z)}$, βs are the corresponding intercept terms under the constraints: $\beta_{1j} < \cdots < \beta_{Kj}$. After fitting the model, the probes are ranked based on the p values for testing the null hypothesis $H_0 : \phi = 0$ and the top 500 probes are selected based on the top smallest p values.

We test the performance of FWOC(k) and FWOC(sp) with two dimensions (as suggested by the average number of discriminant dimensions produced by PLDA), WSDA(k), WSDA(sp), and all the other methods for comparison in the simulation study.

2.7.1 Results

The results for the four datasets are averaged over ten repetitions and are presented in Figure 2.22 and 2.23. We present classification accuracy, Kendall's τ, class-averaged precision and class-average recall in the results. Note that due to the reason that classes are highly imbalanced, we can not rely on one metric. According to Figure 2.22, both FWOC and WSDA generally have a strong performance. The detailed discussions on the performances for each of the four datasets are given in the following.

For the classification of Bcell, FWOC(k) and FWOC(sp) achieve the highest accuracy, followed by WSDA(k), PLDA and BhGLM. PCRM obtains the highest Kendall's τ, followed by WSDA(k). WSDA(k) also obtains the highest class-averaged recall and class-averaged precision, indicating that it performs the best in identifying individual classes. Also, WSDA performs better than SDA for the Bcell data. BhGLM and PCRM performs well in terms of a high Kendall's τ.

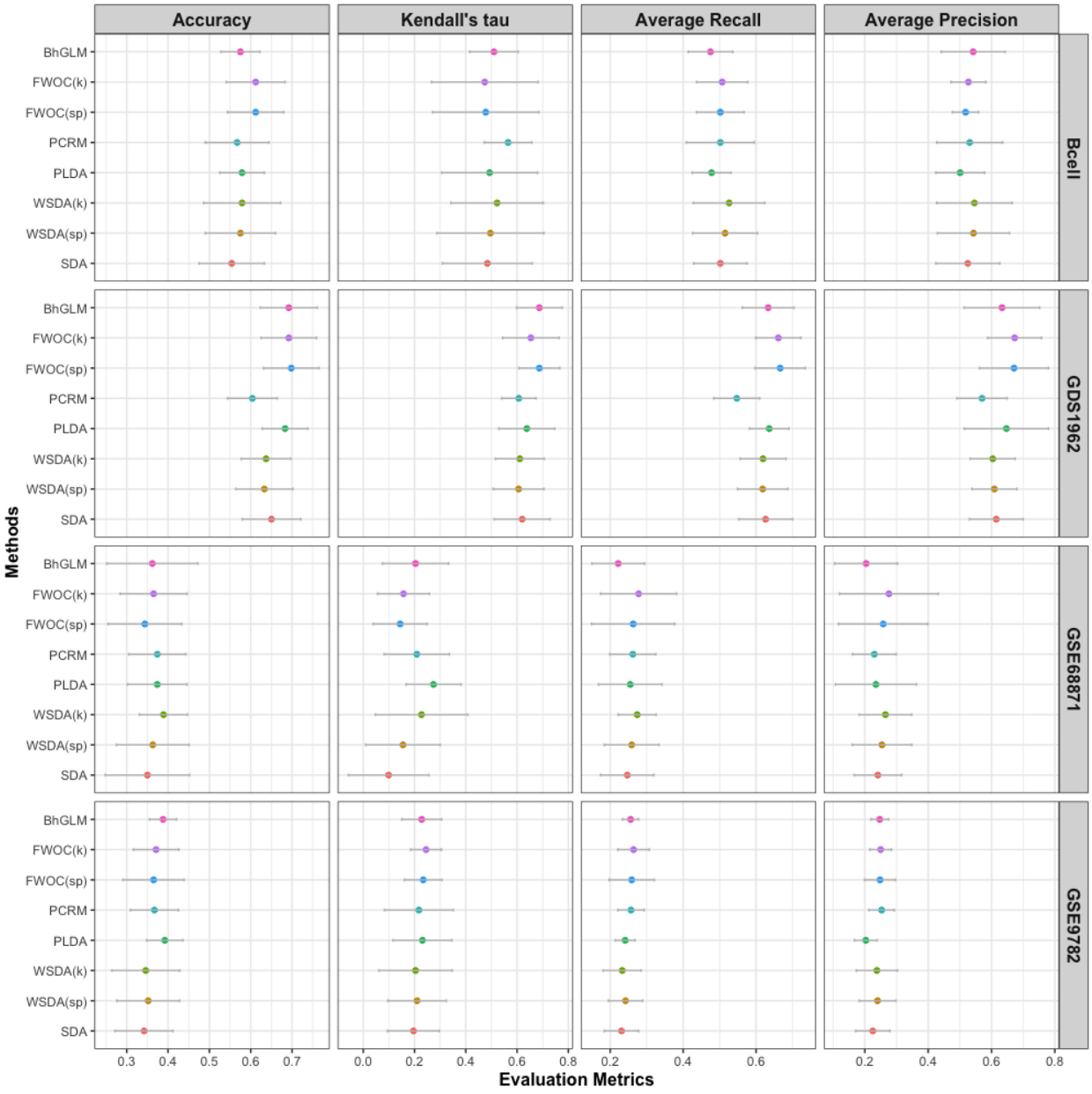


Figure 2.22: Results over ten repetitions for the four datasets. The classification performances are presented in the x-axis, measured by classification accuracy, Kendall's τ, class-average precision, and class-averaged recall. The standard deviations of the measurements are presented by the error bars. The performances for the four datasets are presented in the four rows, respectively.

For the GDS1962 dataset, FWOC(sp), FWOC(k) and BhGLM perform the best, with the highest accuracy and Kendall's τ. Also, FWOC(sp) and FWOC(k) achieve the highest class-averaged recall and class-averaged precision while BhGLM losses its advantage in class-averaged recall and precision. Comparably, PCRM does not perform well for this dataset. SDA is slight better than WSDA. PLDA lies in the middle tier.

In the data of the clinical responses to the induction therapy for MM patients (GSE68871), WSDA(k) has the strongest performance in accuracy and PLDA has the strongest performance in Kendall's τ. FWOC(k) achieves the highest class-averaged precision and recall. Although BhGLM achieves a relatively high Kendall's τ, it has the lowest class-averaged recall and precision, indicating that it may tends to classify classes to the majority class. Also, WSDA performs better than SDA for this dataset.

In terms of the data of the clinical responses to Bortezomib for MM patients (GSE9782), PLDA and FWOC(k) have the strongest performance in accuracy and Kendall's τ, respectively. WSDA works better than SDA. FWOC also performs the best in terms of class-averaged recall and class-averaged precision.

In addition, Kendall's τ works better with FWOC than Spearman's rank, except for GDS1962. According to Figure 2.23, generally the most sparse solution is achieved by PCRM. WSDA and SDA also achieve solutions with high sparsity. The solution of PLDA and BhGLM are generally not very sparse. FWOC performs in the middle tier in terms of feature selection.

Above all, both of FWOC and WSDA have a consistently strong performance in classifying these four ordinal datasets. In addition, methods perform better with four classes than five classes, in which the GDS1962 dataset is the easiest one to be classified.

As discussed earlier, FWOC, PLDA, WSDA and SDA estimate the discriminant subspace onto which we can project the data to see the pattern of classes. To visualize the underlying structure of the datasets, we further project the four datasets onto a two-dimensional subspace obtained from these methods, as shown in Figures 2.24, 2.25, 2.26, 2.27 for Bcell, GDS1962, GSE68871, GSE9782, respectively.

In general, the two-dimensional projections for the four datasets reveal that the classes have a nonlinear ordinality and one dimension is not sufficient to separate the classes. The projections from the

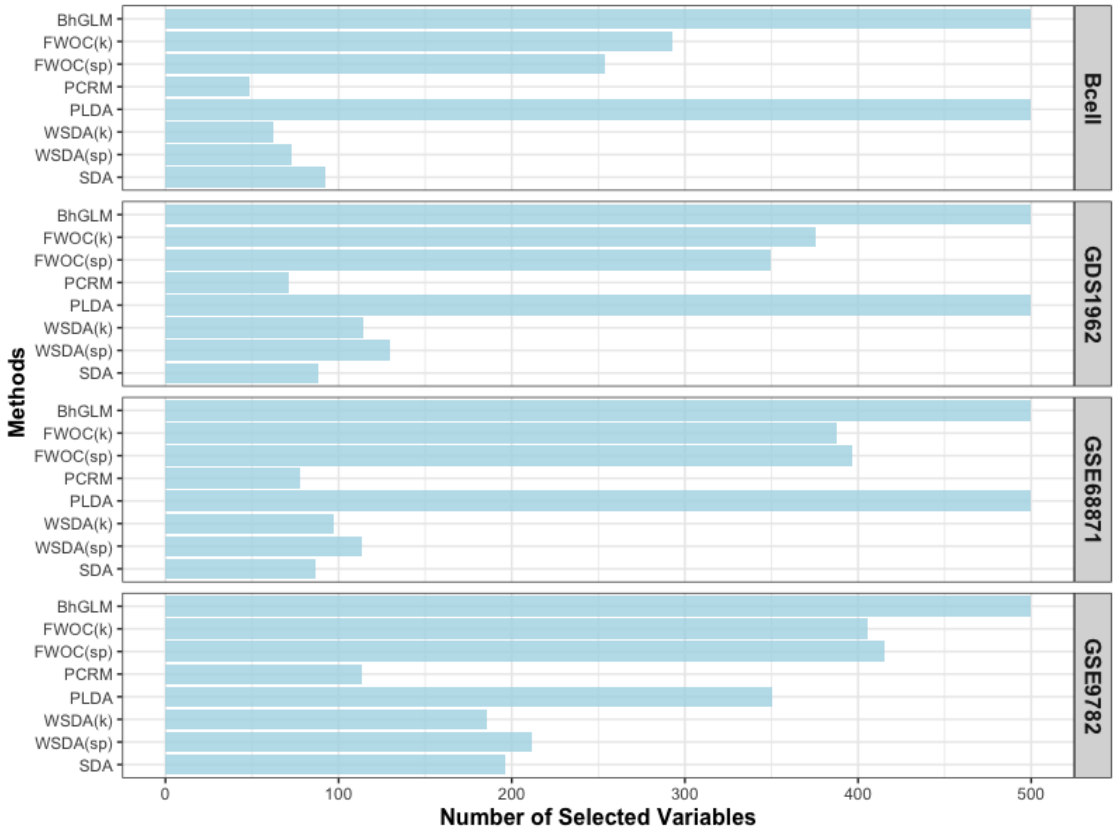


Figure 2.23: Results of feature selection over ten repetitions for the four datasets. The number of selected features is presented as bar-graphs in the x-axis. The four rows indicate four datasets.

four methods look similar but actually different. For Bcell, all of the four projections make a good job in separating class 1 (B1), class 2 (B2) versus class 3/4 (B3/B4). Classes of B3 and B4 are not very separable in these projections, which may indicate that the pathological differences between tumor grade B3 and B4 are not significant.

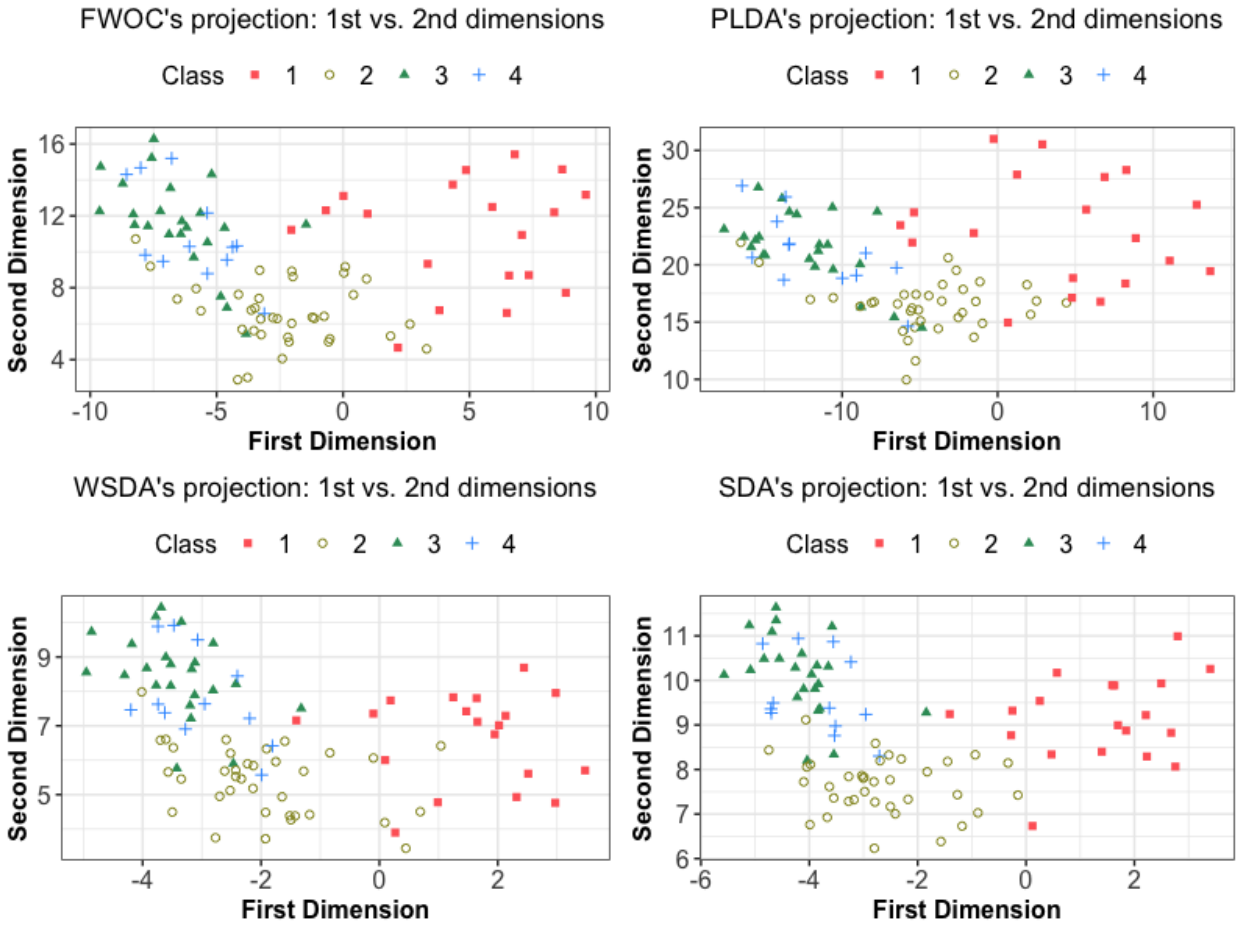


Figure 2.24: Two-dimensional projection for Bcell dataset. Classes are marked by different colors and shapes.

For GDS1962, FWOC works the best in separating the four classes. The projected class 4 (Grade IV) are more close to projected class 1 (Normal) and class 2 (Grade II), in the projections obtained from PLDA and SDA. This is in disagreement with the fact that 'Normal' samples should be far apart from 'Grade IV' samples. Also, there exists some ordinality in the first dimension of the FWOC's projection, in which the centers of class 1 to class 4 are aligned from the left to the right. In comparison, the center

of projected class 2, 3,and 4 are overlapped in the first dimension of projections from WSDA and SDA. This observations is also in agreement with the observation that FWOC works the best for GSE1962 in terms of the evaluation metrics. Also, the ordinality observed in the first dimension is consistent with the phenomenon that BhGLM also has a good performance in terms of evaluation metrics.

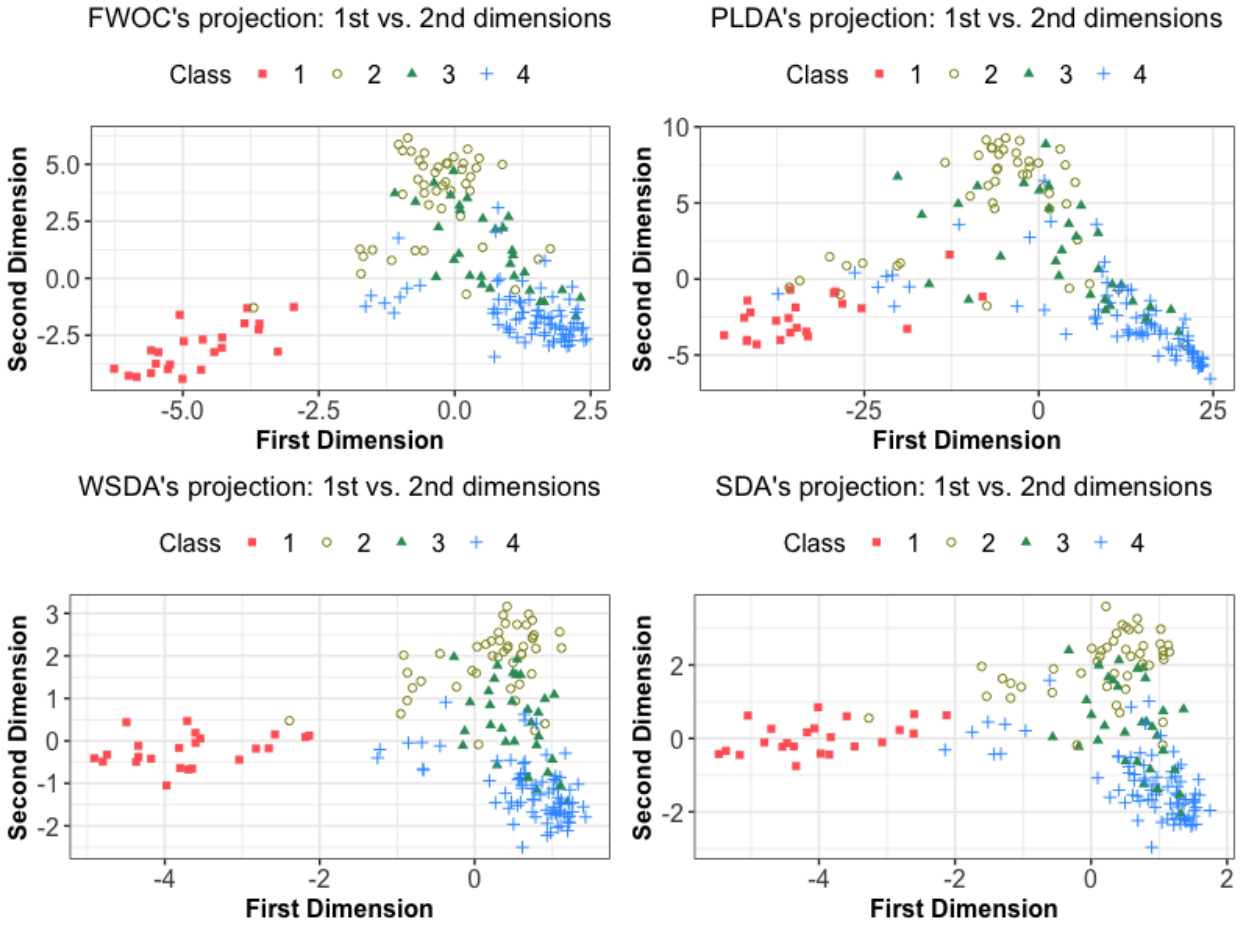


Figure 2.25: Two-dimensional projection for GDS1962 dataset.

The projections for the GSE68871 dataset are presented in Figure 2.26. FWOC has the best projection in terms of class separation. The classes overlap a lot in PLDA's projection. The projections from WSDA and SDA are similar, but we can observe that class 2 and class 3 are more separable in WSDA's projection. In addition, among the four projections, we can see that samples from class 1 (CR) and class 2 (NCR) are heavily overlapped, implying that the pathological differences between 'complete response' and 'near complete response' to the VTD therapy may be negligible.

For the dataset of GSE9782 shown in Figure 2.27, FWOC, WSDA and SDA show much better class separation than PLDA. The projected classes are more condensed in the projection from SDA. Also, the centers of the classes align from the left to the right, which is consistent with the ordering of the classes. Furthermore, we see that class 3 (MR) and class 4 (NC) are close to each other compared with the other three classes, which indicates that the pathological difference between 'no change' and 'minimum response' to Bortezomib therapy may be small.

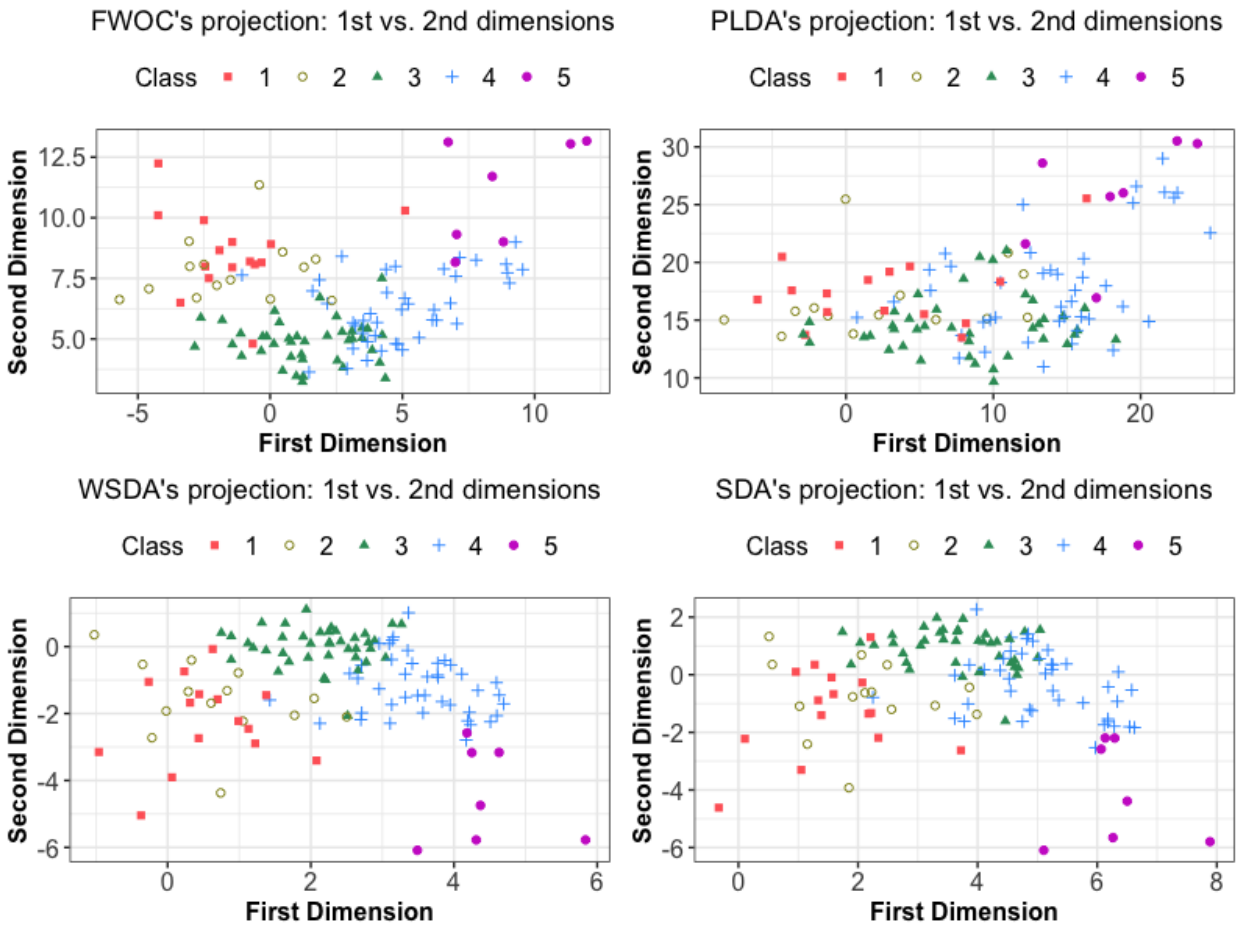


Figure 2.26: Two-dimensional projection for GSE68871 dataset.

Above all, both FWOC and WSDA produce good projections of the datasets. The model-based approaches are not capable of providing projections with two and more dimensions. We see that additional information such as the adjacency between certain classes are revealed by these projections, which will help in understanding the underlying similarities or differences among the classes.

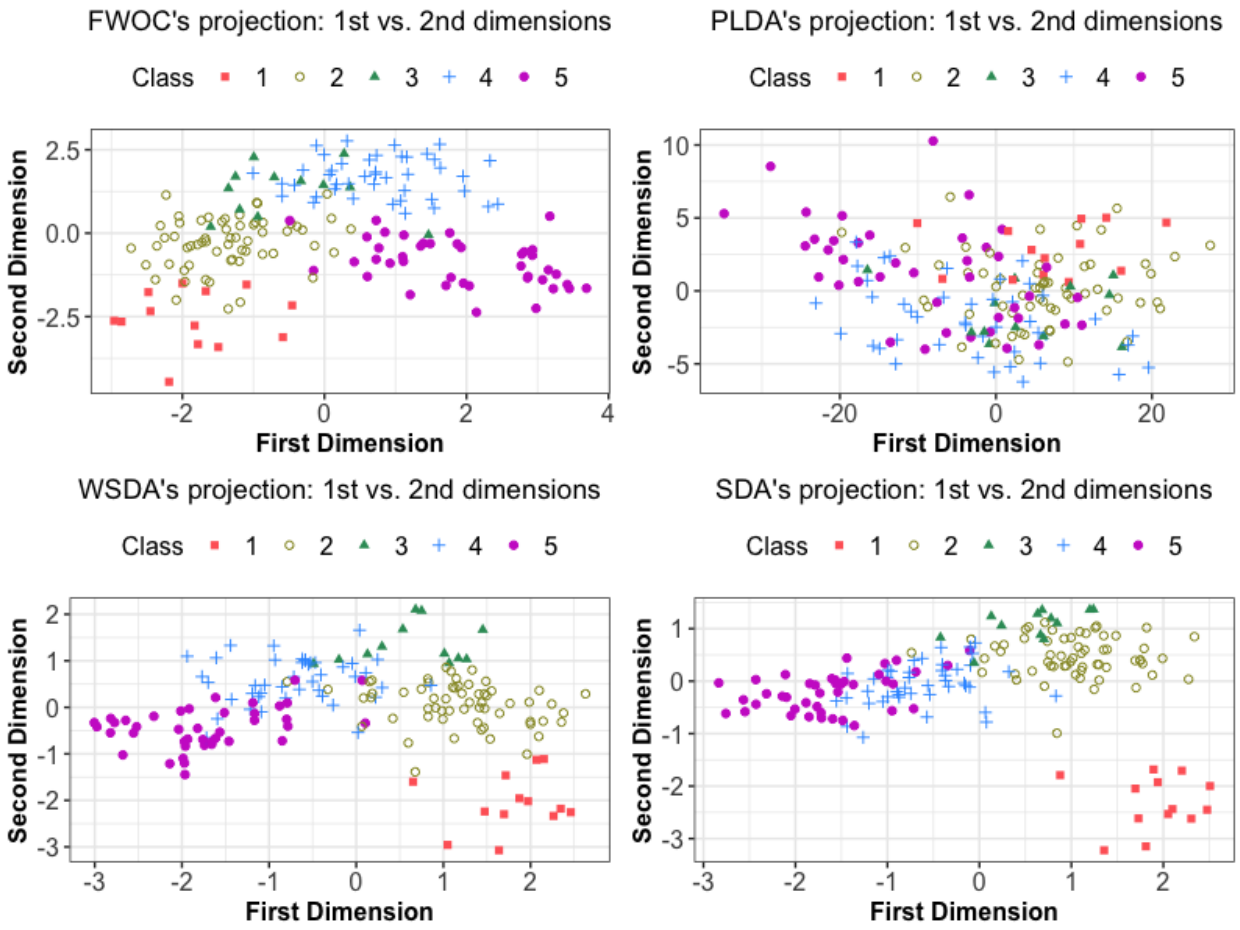


Figure 2.27: Two-dimensional projection for GSE9782 dataset.

2.8 Conclusions

In this work, we introduce the concept of 'feature weighting' in the context of ordinal classification, to quantify the contribution of individual features. Based on 'feature weighting', we propose two novel ordinal classification methods, namely feature-weighted ordinal classification method (FWOC) and weighted sparse discriminant analysis (WSDA). FWOC works by incorporating the feature weights into the framework of LDA and finding a balance between class separation and ordinality. FWOC works well when the dimension exceeds the sample size and is able to achieve a sparse solution through a group Lasso penalty. It also has been shown that FWOC is closely related to the well-known classification methods, such as ridge-corrected LDA and optimal scoring. In comparison, WSDA works by incorporating the feature weights into the framework of sparse optimal scoring via adaptive Lasso. It is also closely related to the work proposed by Clemmensen et al. (2011).

Different from the existing ordinal approaches which assume a linear ordinality among classes, both FWOC and WSDA are capable of learning the true structure of the ordinality and detecting a nonlinear ordinal structure if exists. Simulation studies demonstrate that both WSDA and FWOC have promising performances across different scenarios. It is also interesting to know that WSDA performs better than SDA in the ordinal scenarios and such advantages vanish when the data are purely nominal. This indicates that the incorporation of feature weights does help in the ordinal settings.

When applied to the real data examples, both FWOC and WSDA obtain competitive performances compared with other methods, in which their projections reveal that the underlying structures of the classes are not linearly ordinal. This indicates that imposing a strict linear ordinal assumption on the classes is not appropriate.

A main limitation of FWOC is that it is time-consuming to search for an optimal parameter pair in a 2-dimensional grid. A possible solution to reduce the computation time is to pre-define a smaller tuning range. In addition, in both FWOC and WSDA, we use pre-defined number of discriminant vectors, which can probably be considered as a tuning parameter in the future. However, adding another tuning

parameter will definitely increase the complexity and weaken the interpretations of the effects of individual parameter. Thus, a further investigation might be needed. A limitation of WSDA is that the tuning range is not bounded when we carry the calculations from the work by Clemmensen et al. (2011). Therefore, in practice, it may yield a trivial solution with a large tuning parameter. It remains efforts to study an appropriate range of the tuning parameter. What is more, the feature weights are only applicable with continuous features. When there are categorical features, new weighting criterion is needed to be explored. Also, there might be other choices of weights other than Spearman's rank correlation and Kendall's τ that could also be applied to the two methods, such as the Goodman and Kruskal's gamma (Goodman & Kruskal, 1959, 1963, 1972, 1979).

2.9 Supplement

This section contains the supplementary materials.

2.9.1 Connections between LDA and Optimal Scoring

In this section, we discuss the connections between LDA and optimal scoring, according to the work of Hastie et al. (1995). The connections have to be formulated via the link of canonical correlation analysis (CCA). Here, we assume that X has been centered in columns, and Y is the dummy indicator matrix defined in Section 2.2.2. We introduce some notations:

$$\begin{aligned}
\Sigma_{YY} &= \frac{1}{n} Y^T Y \quad \text{denotes the diagonal matrix of the class proportions,} \\
\Sigma_{XX} &= \frac{1}{n} X^T X \quad \text{denotes the total covariance matrix,} \\
\Sigma_{YX} &= \Sigma_{XY}^T = \frac{1}{n} Y^T X.
\end{aligned}$$

Then, CCA is introduced as follows. Given $X_{n\times p}$ and $Y_{n\times K}$ as two sets of observations of two random variables, CCA aims to find linear combinations of X and Y with the largest correlation with each other. Assume $X\boldsymbol{\beta}$ and $Y\boldsymbol{\theta}$ are the desired linear combinations of X and Y, the optimization problem of finding

sequential canonical pairs $(\boldsymbol{\theta}, \boldsymbol{\beta})$ is given by:

$$\begin{aligned} &\max_{\boldsymbol{\theta} \in R^K,\, \boldsymbol{\beta} \in R^p} \boldsymbol{\theta}^T \Sigma_{YX} \boldsymbol{\beta}, \\ &\text{subject to } \boldsymbol{\theta}^T \Sigma_{YY} \boldsymbol{\theta} = 1, \\ &\boldsymbol{\beta}^T \Sigma_{XX} \boldsymbol{\beta} = 1. \end{aligned} \tag{2.23}$$

Let $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_l)$ and $B = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_l)$ be the collections of optimal canonical vectors, where $l = \min(p, K)$. In fact, Θ and B are the left and right singular matrix obtained through the singular value decomposition of $\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1}$, with orthogonality constraints:

$$\begin{aligned} \Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1} &= \Theta D_d B^T, \\ \Theta^T \Sigma_{YY} \Theta &= I, \\ B^T \Sigma_{XX} B &= I, \end{aligned} \tag{2.24}$$

where D_d is a diagonal matrix with non-negative singular values d_i.

Next, we rewrite the optimization criterion in (2.10) as:

$$R(\boldsymbol{\theta}, \boldsymbol{\beta}) = \boldsymbol{\theta}^T \Sigma_{YY} \boldsymbol{\theta} - 2 \boldsymbol{\theta}^T \Sigma_{YX} \boldsymbol{\beta} + \boldsymbol{\beta}^T \Sigma_{XX} \boldsymbol{\beta}. \tag{2.25}$$

By (2.25), given $\boldsymbol{\theta}$, the optimal estimate of $\boldsymbol{\beta}$ is given by: $\boldsymbol{\beta}_{os} = \Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta}$. Then, if we substitute $\boldsymbol{\beta}_{os}$ into (2.25), the optimal estimate of $\boldsymbol{\theta}$ will minimize: $1 - \boldsymbol{\theta}^T \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta}$. A direct comparison of (2.25) and (2.23) suggests that optimal scoring is minimizing the same criterion as CCA when given the same constraints on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Also, it can be shown that, when given $\boldsymbol{\theta}$, the optimal estimate of $\boldsymbol{\beta}$ in CCA problem is given by:

$$\begin{aligned} \boldsymbol{\beta}_{cca} &= \frac{\Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta}}{\sqrt{(\Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta})^T \Sigma_{XX} (\Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta})}} \\ &= \frac{\boldsymbol{\beta}_{os}}{\sqrt{\boldsymbol{\beta}_{os}^T \Sigma_{XY} \boldsymbol{\beta}_{os}}}, \end{aligned} \tag{2.26}$$

which is the same as $\boldsymbol{\beta}_{os}$ up to a constant scale. Also, if we substitute $\boldsymbol{\beta}_{cca}$ into (2.23), the criterion will be the same as maximizing $\boldsymbol{\theta}^T \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY} \boldsymbol{\theta}$. Thus, CCA and optimal scoring obtain the same score vectors ($\boldsymbol{\theta}$) and discriminant vectors ($\boldsymbol{\beta}$) up to a scale factor.

What is more, by relating CCA to LDA using (2.24), it can be shown that $B_{cca} D_{(1-d^2)^{-\frac{1}{2}}}$ satisfies the criterion (2.9) in LDA, where B_{cca} denotes the optimal solution given by CCA and $D_{(1-d^2)^{-\frac{1}{2}}}$ is a diagonal matrix with elements $(1 - d_i^2)^{-\frac{1}{2}}$. Above all, the discriminant vectors obtained by LDA and optimal scoring are the same up to a constant scale, therefore, they are indeed equivalent.

2.9.2 Coordinate Descent with Lasso

Here, we introduce the coordinate descent algorithm, which is designed to find the solution that minimizes a function: $\min_{\mathbf{x} \in R^p} \ f(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_p)^T$. The main idea of coordinate descent is to minimize $f(\mathbf{x})$ by finding the minimum in each coordinate, which solves a univariate problem instead of a multivariate problem during the iterations. Specifically, the steps of coordinate descent are given below:

1. initial value: $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_p^0)^T$,

2. At k^{th}iteration step:

$$x_1^k = \arg \min_{x_1} \ f(x_1, x_2^{k-1}, x_3^{k-1}, \ldots, x_p^{k-1}),$$

$$x_2^k = \arg \min_{x_2} \ f(x_1^k, x_2, x_3^{k-1}, \ldots, x_p^{k-1}),$$

$$\ldots$$

$$x_p^k = \arg \min_{x_p} \ f(x_1^k, x_2^k, x_3^k, \ldots, x_p),$$

where $\mathbf{x}^k$ is the solution in k step. The updating rule for each coordinate is: $x_j = x_j - \alpha \frac{\partial f(\mathbf{x})}{\partial x_j}$, with α being the learning rate. Note that if a closed-form solution for the minimum at the coordinate exists, we can directly use the closed form instead of the updating rule. The steps in coordinate descent satisfy $f(\mathbf{x}^0) \geq f(\mathbf{x}^1) \geq f(\mathbf{x}^2) \geq \ldots$, which guarantee the convergence. What is more, the order of updating coordinates can be arbitrary, which is not necessary from 1 to p.

In the steps in Algorithm 2, we aim to find $\boldsymbol{\beta}_m^*$ as the solution to:

$$\min_{\boldsymbol{\beta}_m^*} \|Y\boldsymbol{\theta}_m - X_w\boldsymbol{\beta}_m^*\|_2^2 + \lambda\|\boldsymbol{\beta}_m^*\|_1, \tag{2.27}$$

where the coordinate descent algorithm can be applied. We first introduce the concept of sub-gradient as given below, since the absolute function in the Lasso penalty is not directly differentiable at 0.

Definition 3. *The sub-gradient of function* $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ *is defined by:*

$$\partial_\beta\|\boldsymbol{\beta}\|_1 = \frac{\partial \sum_{j=1}^p |\beta_j|}{\partial \beta_j} = \begin{cases} -1, \textit{ for } \beta_j < 0, \\ [-1, 1], \textit{ for } \beta_j = 0, \\ 1, \textit{ for } \beta_j > 0. \end{cases} \tag{2.28}$$

Next, the steps of finding the optimal solution of $\boldsymbol{\beta}_m^*$ is given as follows. Denote $Y^{'} = Y\boldsymbol{\theta}_m$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_m^* = (\beta_1, \ldots, \beta_p)^T$, then (2.27) is equivalent to the following:

$$\begin{aligned} & \min_{\boldsymbol{\beta}} \|Y^{'} - X_w\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \\ \Leftrightarrow\ & \min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T X_w^T X_w \boldsymbol{\beta} - 2Y^{'T} X_w \boldsymbol{\beta} + \lambda \sum_{j=1}^p |\beta_j|, \\ \Leftrightarrow\ & \min_{\boldsymbol{\beta}} L_1 + L_2, \end{aligned} \tag{2.29}$$

where $L_1 = \boldsymbol{\beta}^T X_w^T X_w \boldsymbol{\beta} - 2Y^{'T} X_w \boldsymbol{\beta}$ and $L_2 = \lambda \sum_{j=1}^p |\beta_j|$. Let $A = 2X_w^T X_w$, $\mathbf{d}^T = ZY^{'T} X_w$, then:

$$\begin{aligned} L_1 &= \frac{1}{2}\boldsymbol{\beta}^T A \boldsymbol{\beta} - \mathbf{d}^T\boldsymbol{\beta} \\ &= \frac{1}{2}(\sum_{j=1}^p a_{jj}\beta_j^2 + \sum_{l\neq j} a_{jl}\beta_j\beta_l) - \sum_{j=1}^p d_j\beta_j, \\ L_2 &= \lambda \sum_{j=1}^p |\beta_j|, \end{aligned} \tag{2.30}$$

where a_{jl} is the (j,l)th element of A and d_j is the jth component of $\mathbf{d}$. Then, by taking the derivatives, we have:

$$\frac{\partial L_1}{\partial \beta_j} = a_{jj}\beta_j + \sum_{\forall l \neq j} a_{jl}\beta_l - d_j,$$

$$\frac{\partial L_2}{\partial \beta_j} = \begin{cases} -\lambda, \text{ for } \beta_j < 0, \\ [-\lambda, \lambda], \text{ for } \beta_j = 0, \\ \lambda, \text{ for } \beta_j > 0. \end{cases}$$

and

$$\frac{\partial L_1}{\partial \beta_j} + \frac{\partial L_2}{\partial \beta_j} = \begin{cases} a_{jj}\beta_j + \sum_{\forall l \neq j} a_{jl}\beta_l - d_j - \lambda, \text{ for } \beta_j < 0, \\ [a_{jj}\beta_j + \sum_{\forall l \neq j} a_{jl}\beta_l - d_j - \lambda, a_{jj}\beta_j + \sum_{\forall l \neq j} a_{jl}\beta_l - d_j + \lambda], \text{ for } \beta_j = 0, \\ a_{jj}\beta_j + \sum_{\forall l \neq j} a_{jl}\beta_l - d_j + \lambda, \text{ for } \beta_j > 0. \end{cases}$$

By setting the derivatives to zero, we get:

$$\beta_j = \begin{cases} \frac{d_j + \lambda - \sum_{\forall l \neq j} a_{jl}\beta_l}{a_{jj}}, \text{ for } d_j - \sum_{\forall l \neq j} a_{jl}\beta_l < -\lambda, \\ 0, \text{ for } -\lambda \leq d_j - \sum_{l \neq j} a_{jl}\beta_l \leq \lambda, \\ \frac{d_j - \lambda - \sum_{\forall l \neq j} a_{jl}\beta_l}{a_{jj}}, \text{ for } d_j - \sum_{\forall l \neq j} a_{jl}\beta_l > \lambda. \end{cases}$$

Define

$$S_\lambda(c) = \begin{cases} c + \lambda, \text{ for } c < -\lambda, \\ 0, \text{ for } -\lambda \leq c \leq \lambda, \\ c - \lambda, \text{ for } c > \lambda. \end{cases}$$

Then the solution to (2.27) is $\beta_j = \frac{S_\lambda(d_j - \sum_{\forall l \neq j} a_{jl}\beta_l)}{a_{jj}}$. Clearly, when $-\lambda \leq d_j - \sum_{l \neq j} a_{jl}\beta_l \leq \lambda$, $\beta_j = 0$, i.e., when β's are all zero, $\lambda \geq |d_j - \sum_{l \neq j} a_{jl}\beta_l| \geq |d_j|$. That is to say, when $\lambda \geq \max_j |d_j| = 2\max_j \langle \mathbf{x}_{wj}, Y' \rangle$, the solution of $\boldsymbol{\beta}$ will be trivial, where $\mathbf{x}_{wj}$ is the jth column of X_w.

2.9.3 Tables for Simulation Results

Table 2.4: Class-averaged recall and precision in the scenario with a linear ordinality. Standard deviations are presented within parenthesis.

	Auto-correlation		Compound Symmetry		Identity	
	avg_recall	avg_prec	avg_recall	avg_prec	avg_recall	avg_prec
BhGLM	0.685	0.708	0.626	0.657	0.603	0.641
	(0.062)	(0.064)	(0.079)	(0.077)	(0.078)	(0.075)
FWOC(k)(c2)_2D	0.706	0.714	0.796	0.813	0.729	0.75
	(0.074)	(0.076)	(0.066)	(0.068)	(0.069)	(0.065)
FWOC(k)(c2)_3D	0.729	0.735	0.795	0.809	0.721	0.739
	(0.070)	(0.074)	(0.077)	(0.073)	(0.078)	(0.082)
FWOC(k)(c1)_2D	0.706	0.714	0.796	0.812	0.729	0.75
	(0.074)	(0.076)	(0.066)	(0.068)	(0.069)	(0.065)
FWOC(k)(c1)_3D	0.729	0.735	0.795	0.809	0.721	0.739
	(0.070)	(0.074)	(0.077)	(0.073)	(0.078)	(0.082)
FWOC(sp)(c2)_2D	0.71	0.717	0.798	0.815	0.726	0.747
	(0.073)	(0.077)	(0.065)	(0.065)	(0.070)	(0.067)
FWOC(sp)(c2)_3D	0.73	0.737	0.794	0.809	0.723	0.74
	(0.066)	(0.069)	(0.080)	(0.077)	(0.077)	(0.081)
FWOC(sp)(c1)_2D	0.71	0.717	0.799	0.815	0.726	0.747
	(0.073)	(0.077)	(0.065)	(0.065)	(0.070)	(0.067)
FWOC(sp)(c1)_3D	0.73	0.737	0.794	0.809	0.723	0.74
	(0.066)	(0.069)	(0.081)	(0.078)	(0.077)	(0.081)
PCRM	0.665	0.71	0.764	0.79	0.71	0.739
	(0.075)	(0.071)	(0.070)	(0.061)	(0.071)	(0.070)
PLDA	0.524	0.543	0.798	0.818	0.779	0.797
	(0.082)	(0.090)	(0.077)	(0.067)	(0.058)	(0.058)
SDA	0.646	0.652	0.64	0.649	0.589	0.596
	(0.081)	(0.084)	(0.086)	(0.094)	(0.084)	(0.086)
WSDA(k)	0.811	0.82	0.788	0.801	0.715	0.733
	(0.055)	(0.055)	(0.070)	(0.066)	(0.073)	(0.074)
WSDA(sp)	0.822	0.832	0.789	0.802	0.724	0.739
	(0.062)	(0.061)	(0.072)	(0.070)	(0.063)	(0.068)

Table 2.5: Class-averaged recall and precision in the scenario with a nonlinear ordinality. Standard deviations are presented within parenthesis.

	Auto-correlation		Compound Symmetry		Identity	
	avg_recall	avg_prec	avg_recall	avg_prec	avg_recall	avg_prec
BhGLM	0.578	0.589	0.503	0.538	0.46	0.507
	(0.082)	(0.089)	(0.083)	(0.093)	(0.090)	(0.106)
FWOC(k)(c2)_3D	0.62	0.628	0.65	0.663	0.586	0.6
	(0.095)	(0.101)	(0.087)	(0.090)	(0.081)	(0.085)
FWOC(k)(c1)_3D	0.621	0.629	0.65	0.663	0.585	0.599
	(0.095)	(0.100)	(0.087)	(0.090)	(0.081)	(0.086)
FWOC(sp)(c2)_3D	0.625	0.634	0.652	0.666	0.591	0.606
	(0.098)	(0.104)	(0.085)	(0.086)	(0.084)	(0.087)
FWOC(sp)(c1)_3D	0.625	0.634	0.652	0.666	0.59	0.605
	(0.098)	(0.104)	(0.085)	(0.086)	(0.084)	(0.088)
PCRM	0.515	0.551	0.561	0.61	0.502	0.553
	(0.090)	(0.105)	(0.080)	(0.078)	(0.075)	(0.082)
PLDA	0.423	0.436	0.544	0.573	0.534	0.567
	(0.069)	(0.100)	(0.096)	(0.116)	(0.080)	(0.094)
SDA	0.542	0.547	0.527	0.538	0.489	0.497
	(0.092)	(0.098)	(0.090)	(0.097)	(0.083)	(0.090)
WSDA(k)	0.76	0.775	0.656	0.674	0.587	0.602
	(0.083)	(0.083)	(0.079)	(0.083)	(0.091)	(0.096)
WSDA(sp)	0.773	0.79	0.66	0.678	0.592	0.608
	(0.074)	(0.074)	(0.071)	(0.074)	(0.089)	(0.093)

Table 2.6: Class-averaged recall and precision in the scenario of mixed ordinality. Standard deviations are presented within parenthesis.

	Auto-correlation		Compound Symmetry		Identity	
	avg_recall	avg_prec	avg_recall	avg_prec	avg_recall	avg_prec
BhGLM	0.366	0.394	0.414	0.449	0.557	0.594
	(0.070)	(0.085)	(0.075)	(0.085)	(0.081)	(0.081)
FWOC(k)(c2)_3D	0.861	0.874	0.812	0.826	0.875	0.887
	(0.059)	(0.057)	(0.068)	(0.064)	(0.051)	(0.049)
FWOC(k)(c1)_3D	0.862	0.874	0.812	0.826	0.875	0.886
	(0.059)	(0.057)	(0.067)	(0.064)	(0.050)	(0.049)
FWOC(sp)(c2)_3D	0.855	0.869	0.81	0.822	0.872	0.884
	(0.056)	(0.057)	(0.068)	(0.066)	(0.052)	(0.049)
FWOC(sp)(c1)_3D	0.855	0.869	0.81	0.823	0.872	0.883
	(0.056)	(0.057)	(0.068)	(0.066)	(0.052)	(0.048)
PCRM	0.339	0.376	0.391	0.44	0.604	0.645
	(0.062)	(0.088)	(0.082)	(0.099)	(0.074)	(0.074)
PLDA	0.572	0.642	0.662	0.694	0.814	0.83
	(0.133)	(0.126)	(0.126)	(0.114)	(0.084)	(0.077)
SDA	0.851	0.862	0.751	0.764	0.826	0.84
	(0.051)	(0.052)	(0.074)	(0.072)	(0.063)	(0.061)
WSDA(k)	0.872	0.881	0.73	0.744	0.816	0.832
	(0.071)	(0.071)	(0.075)	(0.074)	(0.059)	(0.058)
WSDA(sp)	0.877	0.885	0.736	0.749	0.82	0.835
	(0.070)	(0.069)	(0.076)	(0.073)	(0.060)	(0.061)

Table 2.7: Class-averaged recall and precision in the nominal scenario. Standard deviations are presented within parenthesis.

	Auto-correlation		Compound Symmetry		Identity	
	avg_recall	avg_prec	avg_recall	avg_prec	avg_recall	avg_prec
BhGLM	0.283	0.242	0.307	0.287	0.297	0.267
	(0.066)	(0.126)	(0.068)	(0.116)	(0.066)	(0.119)
FWOC(k)(c2)_3D	0.439	0.45	0.445	0.455	0.425	0.439
	(0.087)	(0.095)	(0.106)	(0.118)	(0.080)	(0.091)
FWOC(k)(c1)_3D	0.439	0.45	0.445	0.455	0.425	0.439
	(0.087)	(0.095)	(0.106)	(0.118)	(0.080)	(0.091)
FWOC(sp)(c2)_3D	0.437	0.446	0.444	0.454	0.42	0.429
	(0.090)	(0.098)	(0.105)	(0.119)	(0.082)	(0.094)
FWOC(sp)(c1)_3D	0.437	0.446	0.444	0.454	0.42	0.429
	(0.090)	(0.098)	(0.105)	(0.119)	(0.082)	(0.094)
PCRM	0.267	0.274	0.305	0.33	0.287	0.31
	(0.063)	(0.087)	(0.072)	(0.089)	(0.062)	(0.089)
PLDA	0.327	0.323	0.354	0.361	0.355	0.362
	(0.067)	(0.101)	(0.062)	(0.092)	(0.072)	(0.100)
SDA	0.491	0.505	0.447	0.457	0.404	0.409
	(0.095)	(0.098)	(0.072)	(0.078)	(0.084)	(0.090)
WSDA(k)	0.48	0.486	0.416	0.424	0.396	0.402
	(0.110)	(0.115)	(0.082)	(0.085)	(0.075)	(0.080)
WSDA(sp)	0.487	0.495	0.416	0.425	0.397	0.404
	(0.104)	(0.109)	(0.079)	(0.084)	(0.074)	(0.078)

Chapter 3

Trace Regularization in Dimension Reduction and its Applications on High-dimensional CCA

3.1 Introduction

High-dimensional data are unavoidable in machine learning and pattern recognition, which bring the curse of dimensionality (Bellman, 2015) in many model building processes. For example, as the dimension gets higher, the data will become more and more sparse since the volume of the data space increases very fast. This phenomena can be illustrated by the following. Suppose we have 100 data points that are uniformly distributed in an interval of unit length (dimension is one), then the average distance between points is 0.01. If these data points are distributed in an unit square (dimension is two), then the average area taken by each point is 0.01 and the average distance (edge of the area) between points is $0.01^{\frac{1}{2}} = 0.1$. For a unit space of dimension n, the distance would become $0.01^{\frac{1}{n}}$. Figure 3.1 shows the average distance between points with respect to different dimensions. Clearly, the distance grows exponentially with the dimensions and will converge to 1, indicating that the sample points will be close to the edges of the space eventually. Therefore, samples will become more and more sparse as dimension increases. In fact, the sparsity of samples is a problematic issue in many machine learning problems. Image the situation when we want to use the nearest neighborhood of a sample point to help us determine the class of it,

if the samples are distributed near the edges of the space, a very large edge of the neighborhood would be required in order to cover a sufficient amount of neighbors (J. Friedman et al., 2001), which will be undesirable. A high dimension will also increase the computational time of any algorithm, for example, calculating a $p \times p$ matrix will cause $O(p^2)$ of computation time.

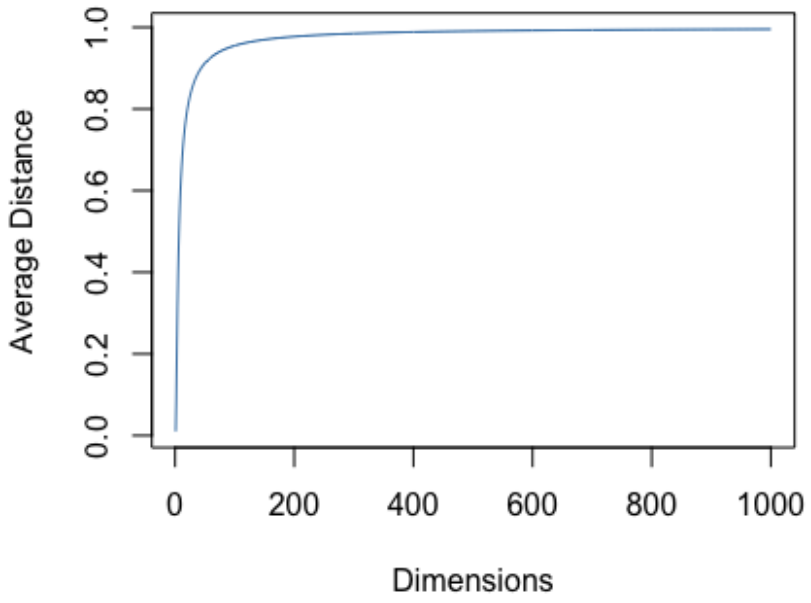


Figure 3.1: The average distance between sample points in a unit space with respect to the dimensions.

A proper reduction in the dimensionality can yield more accurate estimates, get rid of noise, save computational time and storage. Therefore, dimension reduction techniques are on demanding in the high-dimensional space, which refer to the methods that could transform the high-dimensional data to a low-dimensional representation such that the important information carried by the original data will be preserved to the maximum extent. Specifically, given an input data matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p] \in R^{n \times p}$, the goal is to find a faithful representation of X: $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_m] \in R^{n \times m}$, where $m \ll p$.

Many of the supervised and unsupervised algorithms in the family of dimension reduction can be formulated into the framework of a 'trace ratio' optimization problem, such as linear discriminant analysis

(LDA) (Fisher, 1936), canonical correlation analysis (CCA) (Hotelling, 1992) and principle component analysis (PCA) (Hotelling, 1992). Kokiopoulou et al. (2011) provided a comprehensive overview of some well-known dimension reduction techniques and elaborated on how the methods can be seen as trace ratio optimization problems. The trace ratio problem finds a matrix V that maximizes a trace ratio:

$$\max_{V^TV=I} Tr(V^TMV)/Tr(V^TQV), \tag{3.1}$$

where $Tr(\cdot)$ denotes the the trace of a matrix, M and Q are symmetric matrices that characterize the 'signal' and 'noise' of the algorithm, respectively. Take LDA as an example, the objective of LDA is to find a lower-dimensional representation of the original data matrix $X_{n\times p}$ by $Z_{n\times m} = XV$, in order to maximize the between-class scatter while minimizing the within-class scatter, where $m \ll p$ and $V \in R^{p\times m}$ is the desired transformation matrix. The between-class scatter is the signal part which can be expressed as $\sum_{k=1}^{K} \frac{n_i}{n}[V^T(\boldsymbol{\mu}_k-\boldsymbol{\mu})]^T[V^T(\boldsymbol{\mu}_k-\boldsymbol{\mu})]$ and the within-class scatter is the noise part which can be expressed as $\sum_{i=1}^{n}\sum_{y_i=k}\frac{n_i}{n}[V^T(\mathbf{x}_i-\boldsymbol{\mu}_k)]^T[V^T(\mathbf{x}_i-\boldsymbol{\mu}_k)]$, where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_k$ are the global class mean vector and class mean vector for class k, respectively (see other notations in Section 2.2.2). Therefore, maximizing the ratio of signal versus noise in LDA can be equivalently written as: $\max_{V\in R^{p\times m}} \frac{Tr(V^T\Sigma_bV)}{Tr(V^T\Sigma_wV)}$, where Σ_b and Σ_w are the between- and within- class covariance matrix defined in Section 2.2.2. Thus, LDA is indeed a trace ratio optimization problem under the conventional orthogonality constraint: $V^TV = I$.

However, solving the trace ratio optimization problem is not straightforward since it does not have a closed-form solution. Therefore, it is conventionally transformed to a 'ratio trace' optimization problem given as:

$$\max_{V^TV=I} \quad Tr[(V^TMV)^{-1}(V^TQV)],$$

which can be solved by a generalized eigenvalue decomposition (GEVD), i.e., $MV = \Lambda QV$, where Λ is a diagonal matrix consisting of the m largest generalized eigenvalues. Therefore, traditional LDA mentioned above actually solves the ratio trace problem, or equivalently, a problem of GEP: $\Sigma_bV = \Lambda\Sigma_wV$.

However, there are two main weaknesses of the ratio trace approach: 1. The ratio trace problem is indeed inferior to the trace ratio problem. H. Wang et al. (2007) argued that the solution to the ratio trace problem does not necessarily optimize the trace ratio in (3.1) and it may sacrifice the performance of the original trace ratio algorithm; 2. Solving the generalized eigenvalue decomposition will become degenerate when the dimension p is much higher than the sample size n. What is more, there have been some efforts that aim to solve a 'trace difference' problem, which is closely related with the trace ratio problem. According to the work by Kokiopoulou et al. (2011), solving (3.1) is indeed equivalent to finding the zero point of the trace difference function defined as:

$$f(\tau) = \max_{V^TV=I} Tr[V^T(M - \tau Q)V]. \tag{3.2}$$

In fact, let V_{opt} be the optimal solution to (3.1) which yields the maximum trace value $\theta = \frac{Tr(V_{opt}^T M V_{opt})}{Tr(V_{opt}^T Q V_{opt})}$, then it can be seen that for any V that satisfies $V^TV = I$, the condition of $\frac{Tr(V^TMV)}{Tr(V^TQV)} \leq \theta$ holds, i.e., $Tr[V^T(M - \theta Q)V] \leq 0$. The optimal solution V_{opt} will satisfy $Tr[V_{opt}^T(M - \theta Q)V_{opt}] = 0$. Thus, finding the optimal solution to the trace ratio problem is equivalent to finding the zero point of the (3.2). It is also easy to see that $f(\tau)$ is decreasing in τ when Q is positive definite. The existence of the zero point is also guaranteed, as $f(\tau) > 0$ when $\tau = 0$, and $f(\tau) < 0$ when τ is greater than the maximum generalized eigenvalue obtained in the GEP: $MV = \Lambda QV$.

Given the sub-optimal property of the ratio trace optimization, some efforts have been made in the machine learning community on optimizing the trace ratio problem directly by finding the root of $f(\tau)$. For example, Guo et al. (2003) proposed to optimize the trace ratio via finding the root of $f(\tau)$ by a heuristic bisection approach; H. Wang et al. (2007) proposed an iterative procedure to solve the trace ratio optimization problem and they argued that their method is superior to the work by Guo et al. (2003) in terms of a faster convergence and stronger theoretical justifications of the convergence issues; Jia et al. (2009) also introduced a method to solve the trace ratio problem based on the eigenvalue perturbation theory. However, these methods mentioned above generally aim to optimize the trace ratio itself while lacking statistical interpretations in the context.

The rest of this chapter is structured as the following. In Section 3.2, we introduce a trace regularization method that transforms the trace ratio problem into the well-defined generalized eigenvalue decomposition problem, which has shown its strong performance in classification. In Section 3.3 and Section 3.4, we introduce the canonical correlation analysis (CCA), a well-known dimension reduction method and the sparse CCA, a demanding topic designed for the high-dimensional situation where traditional CCA fails. In Section 3.5, we discuss how CCA can be formulated into the framework of a trace optimization problem and introduce how to use the trace regularization method to get sparse estimations of CCA. In Section 3.6, we carry out a simulation study to test the performances of the proposed method and related work under different settings. Finally, we conclude this chapter with some discussions and future directions in Section 3.7.

3.2 Trace Regularization

Under the context of multi-class discriminating problems, Ahn et al. (2020) proposed the trace regularization method as a new trace ratio optimization criterion that considers regularizing $Tr(V^TQT)$ when maximizing $Tr(V^TMV)$. In fact, in the high-dimensional setting where the dimension p exceeds the sample size n, Q will generally be rank-deficient. It can be shown that there exists an infinite number of V that satisfies $Tr(V^TQV) = 0$, which makes the trace ratio become infinite. Therefore, maximizing the trace ratio alone will naturally minimize $Tr(V^TQV)$ which might become degenerate. Thus, a regularization of $Tr(V^TQV)$ seems more reasonable.

The trace regularization method can be formulated as:

$$\max_{V^TV=I} Tr(V^TMV), \text{ subject to } Tr(V^TQV) \leq c, \tag{3.3}$$

where $c > 0$ is a regularization parameter, M and Q are both symmetric and semi-positive definite matrix. It can be shown that the solution to (3.3) lies at the boundary of the constraint, with a very mild condition

(See Lemma 1 in Ahn et al. (2020)), i.e., (3.3) can be thought of equivalent to:

$$\max_{V^TV=I} Tr(V^TMV), \text{ subject to } Tr(V^TQV) = c, \tag{3.4}$$

where $c > 0$. Further, the solution to (3.4) can be obtained via the method of Lagrange multipliers. Let

$$L = Tr(V^TMV) - \tau[Tr(V^TQV) - c] - Tr[O(V^TV - I)], \tag{3.5}$$

where $\tau > 0$ is a Lagrangian multiplier, O is a Lagrangian multiplier matrix that satisfies: $O^T = O$. Taking the partial derivatives of L with respect to V, τ and O and setting them to zero, it is easy to see that the solution to (3.5), $\hat{V}$, satisfies:

$$(M - \tau Q)\hat{V} = \hat{V}O, \tag{3.6a}$$

$$Tr(\hat{V}^TQ\hat{V}) = c, \tag{3.6b}$$

$$\hat{V}^T\hat{V} = I. \tag{3.6c}$$

Let the eigenvalue decomposition of O be: $O = RDR^T$, where $R^TR = I$ and D is the diagonal matrix consisting of the eigenvalues of O. Then, multiplying R on both sides of (3.6a) yields:

$$(M - \tau Q)\hat{V}R = \hat{V}RD.$$

Let $W = \hat{V}R$, then W contains the eigenvectors of $M - \tau Q$. Therefore, the solution $\hat{V}$ lies in the eigen-space of $M - \tau Q$ for some $\tau > 0$. It is known that the trace of V^TMV is maximized when V contains the eigenvectors of M associated with the m largest eigenvalues (Kokiopoulou et al., 2011), where m is the number of columns in V. Therefore, optimizing the trace regularization criterion (3.3) is then equivalent to:

$$\max_{V^TV=I} Tr(V^T(M - \tau Q)V), \tag{3.7}$$

where $\tau > 0$ is a tuning parameter. Denote the solution to (3.7) as $\hat{V}(\tau)$ for a given τ and $g(\tau) = Tr\{\hat{V}(\tau)^T M \hat{V}(\tau)\}$, $h(\tau) = Tr\{\hat{V}(\tau)^T Q \hat{V}(\tau)\}$ as the corresponding trace values. Let $l(\tau) = g(\tau) - \tau h(\tau)$ be the objective value. Since both M and Q are semi-positive definite, the sum of the first m largest eigenvalues of $M - \tau Q$ will be non-increasing when τ increases, therefore, $l(\tau)$ is non-increasing in τ. This observation implies that the tuning parameter τ has to be tuned independently from the objective. In addition, it can be seen that both $g(\tau)$ and $h(\tau)$ are non-increasing in τ.

What is more, as discussed by Ahn et al. (2020), the trace regularization (3.7) is also closely related with the ridge-type GEP:

$$MV = (Q + \delta I)V\Lambda, \tag{3.8}$$

where $\delta > 0$ is a tuning parameter. The ridge-type GEP has also been shown to maximize $Tr(V^T M V)$ while controlling $Tr(V^T Q V)$ when V is a vector (M. H. Lee et al., 2013), with the same objective as the trace regularization. Moreover, when $\delta \to \infty$ and $\tau \to 0$, the solutions to both ridge-type GEP and trace regularization will be close to the leading eigenvectors of M. On the other hand, when $\delta \to 0$ and $\tau \to \infty$, their solutions are close to $P_Q^{\perp} M$ where $P_Q^{\perp}$ is the projection matrix to the orthogonal complement of the column space spanned by Q. Thus, the ridge-type GEP and trace regularization are connected in the sense that there exist some tuning parameters δ and τ that make the solutions same to each other.

The trace regularization method is shown to work well in a wide range of problems under the context of classification, we are therefore interested in applying the trace regularization idea out of the framework of discriminating problems.

3.3 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multi-variate method that studies the correlations between two random vectors, proposed by Hotelling (1992). CCA has been widely applied to different research areas, such as information retrieving, psychology and genomics (Hardoon et al., 2004; Parkhomenko

et al., 2009; Weiss, 1972; D. M. Witten & Tibshirani, 2009). In the traditional setting, CCA aims to find linear combinations of two random vectors $\mathbf{x} \in R^p$ and $\mathbf{y} \in R^q$, denoted as $\mathbf{a}^T\mathbf{x}$ and $\mathbf{b}^T\mathbf{y}$, such that the correlation between $\mathbf{a}^T\mathbf{x}$ and $\mathbf{b}^T\mathbf{y}$ is maximized, where $\mathbf{a} \in R^p$ and $\mathbf{b} \in R^q$ are canonical vectors. Assuming that the covariance of $\mathbf{x}$ and $\mathbf{y}$ is Σ_{xx} and Σ_{yy}, respectively, and the covariance between $\mathbf{x}$ and $\mathbf{y}$ is Σ_{xy}, then the objective function of CCA can be written as:

$$\max_{\mathbf{a}\in R^p,\, \mathbf{b}\in R^q} \rho = \text{corr}(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{y}) = \frac{\mathbf{a}^T\Sigma_{xy}\mathbf{b}}{(\mathbf{a}^T\Sigma_{xx}\mathbf{a})^{1/2}(\mathbf{b}^T\Sigma_{yy}\mathbf{b})^{1/2}}, \tag{3.9}$$

where $\text{corr}(\cdot)$ denotes the Pearson correlation. Note that, scaling of either $\mathbf{a}$ or $\mathbf{b}$ will not affect the correlation in (3.9), therefore, we can further assume $\mathbf{a}^T\Sigma_{xx}\mathbf{a} = 1$ and $\mathbf{b}^T\Sigma_{yy}\mathbf{b} = 1$, such that (3.9) is equivalent to:

$$\max_{\mathbf{a}\in R^p,\, \mathbf{b}\in R^q} \rho = \text{corr}(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{y}) = \mathbf{a}^T\Sigma_{xy}\mathbf{b}, \text{ subject to } \mathbf{a}^T\Sigma_{xx}\mathbf{a} = 1, \mathbf{b}^T\Sigma_{yy}\mathbf{b} = 1. \tag{3.10}$$

To solve (3.10), we can apply the method of Lagrange multipliers. Let

$$L = \mathbf{a}^T\Sigma_{xy}\mathbf{b} - \frac{1}{2}\lambda(\mathbf{a}^T\Sigma_{xx}\mathbf{a} - 1) - \frac{1}{2}\mu(\mathbf{b}^T\Sigma_{yy}\mathbf{b} - 1), \tag{3.11}$$

where $\lambda > 0$ and $\mu > 0$ are Lagrangian multipliers. Setting the partial derivatives of L with respect to $\mathbf{a}$ and $\mathbf{b}$ to zero will yield

$$\frac{\partial L}{\partial \mathbf{a}} = \Sigma_{xy}\mathbf{b} - \lambda\Sigma_{xx}\mathbf{a} = 0, \tag{3.12a}$$

$$\frac{\partial L}{\partial \mathbf{b}} = \Sigma_{yx}\mathbf{a} - \mu\Sigma_{yy}\mathbf{b} = 0. \tag{3.12b}$$

Then if we left-multiply $\mathbf{a}^T$ and $\mathbf{b}^T$ on both sides of (3.12a) and (3.12b), respectively, and substitute the constraints in (3.10), we will get:

$$\mathbf{a}^T\Sigma_{xy}\mathbf{b} = \lambda\mathbf{a}^T\Sigma_{xx}\mathbf{a} = \lambda, \tag{3.13a}$$

$$\mathbf{b}^T\Sigma_{yx}\mathbf{a} = \mu\mathbf{b}^T\Sigma_{yy}\mathbf{b} = \mu. \tag{3.13b}$$

Clearly, $\rho = \lambda = \mu$ is the target correlation to be maximized. Next, based on the calculations from Izenman (2008), it can be shown that the solutions to (3.10) are:

$$\begin{aligned} \mathbf{a} &= \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}\mathbf{v}_1, \\ \mathbf{b} &= \Sigma_{yy}^{-1/2}\mathbf{v}_1, \end{aligned} \tag{3.14}$$

where $\mathbf{v}_1$ is the first eigenvector of $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ that corresponds to the largest eigenvalue.

Given the first pair of the canonical vectors, denoted as $(\mathbf{a}_1, \mathbf{b}_1)$, the subsequent canonical pairs can be found by maximizing the canonical correlations under the constraints that they are uncorrelated with the previous pairs. For example, the second pair of canonical vectors $(\mathbf{a}_2, \mathbf{b}_2)$ can be obtained by:

$$\max_{\mathbf{a}_2\in R^p,\ \mathbf{b}_2\in R^q} \rho_2 = \mathbf{a}_2^T\Sigma_{xy}\mathbf{b}_2, \text{ subject to } \mathbf{a}_2^T\Sigma_{xx}\mathbf{a}_2 = 1, \mathbf{b}_2^T\Sigma_{yy}\mathbf{b}_2 = 1, \mathbf{a}_2^T\Sigma_{xx}\mathbf{a}_1 = 0, \mathbf{b}_2^T\Sigma_{yy}\mathbf{b}_1 = 0.$$

We could get a total number of $m = min(p, q)$ canonical pairs. Further, it can be shown that the kth canonical pair is given by: $\mathbf{a}_k = \Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}\mathbf{v}_k$ and $\mathbf{b}_k = \Sigma_{yy}^{-1/2}\mathbf{v}_k$, where $\mathbf{v}_k$ is the kth eigenvector of $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ (See Izenman (2008)).

Consider two data matrices $X_{n\times p} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]$ and $Y_{n\times q} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T]$ whose rows vectors are the sets of observations of $\mathbf{x}$ and $\mathbf{y}$, respectively, then in practice, we will use the following sample

estimates of Σ_{xx}, Σ_{xy} and Σ_{yy} to calculate the canonical vectors:

$$\hat{\Sigma}_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$
$$\hat{\Sigma}_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T,$$
$$\hat{\Sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})^T,$$

where $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ are the sample means.

3.4 Sparse Canonical Correlation Analysis

When the dimensions are low, the canonical vectors can be directly estimated from the eigenvalue decomposition of $\hat{\Sigma}_{yy}^{-1/2}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2}$, or the singular value decomposition (SVD) of $\hat{\Sigma}_{yy}^{-1/2}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1/2}$, as discussed in Section 3.3. However, when the dimensions p and q are larger than the sample size n, $\hat{\Sigma}_{yy}$ and $\hat{\Sigma}_{xx}$ do not have full rank, indicating that taking the inverse will not yield unique solutions, and the traditional CCA problem will become degenerate. In addition, when the variables are highly correlated, the estimation of the sample covariance matrices will become unstable or undefined. What is more, when the dimensions are high, the linear combinations are difficult to interpret unless sparse solutions are obtained.

To solve theses issues, many efforts have been made by using penalties or regularization techniques on CCA. Vinod (1976) adapted the ridge-regression concepts (Horel, 1962) to overcome the collinearity issue in CCA and proposed a 'canonical ridge' model which adds penalties on the corresponding diagonals of the sample covariance matrices. Similar work includes replacing the covariance matrices with identity matrices and diagonal matrices (Parkhomenko et al., 2009; D. M. Witten et al., 2009). Waaijenborg et al. (2008) proposed to use the elastic-net penalties (Zou & Hastie, 2005) and the univariate soft-thresholding (UST) in CCA through a modification of the iterative partial least square algorithm (Horel, 1962). Later, Chalise and Fridley (2012) made a comparison of the different penalty functions imposed for sparse canonical correlation analysis, including the Lasso, elastic-net, SCAD and hard-thresholding. Their study favors

the SCAD penalty with BIC filter when applying the algorithms on genomic data. Most recently, Gao et al. (2014) proposed a two-stage approach to estimate the sparse canonical vectors, which consists of an initialization stage and a group-Lasso refinement stage. Chen et al. (2013) proposed an computational efficient algorithm named CAPIT, to estimate the sparse canonical vectors through iterative thresholding, after transforming the data using the estimators of the precision matrix Σ_{xx}^{-1} and Σ_{yy}^{-1}.

Among the decent amount of work on sparse canonical correlation analysis, we would like to elaborate several of them which are given in the following. The work by D. M. Witten et al. (2009) is one of the most popular methods which always works as a benchmark method for comparison. D. M. Witten et al. (2009) proposed the penalized matrix decomposition (PMD) which replaces Σ_{xx} and Σ_{yy} with the identity matrices and adds additional penalties on the canonical vectors for sparsity. Specifically, the estimated penalized canonical vectors from PMD solve the objective:

$$\max_{\mathbf{a}\in R^p,\, \mathbf{b}\in R^q} \mathbf{a}^T\Sigma_{xy}\mathbf{b}, \text{ subject to } \|\mathbf{a}\|_2^2 \leq 1, \|\mathbf{b}\|_2^2 \leq 1, P_1(\mathbf{a}) \leq c_1, P_2(\mathbf{b}) \leq c_2,$$

where $P_1(\cdot)$ and $P_2(\cdot)$ are convex penalty functions, such as the Lasso penalty and the fused Lasso penalty. D. M. Witten et al. (2009) stated that with sufficiently small c_1, c_2 and the Lasso penalty, PMD can yield sparse estimations of the canonical vectors $(\mathbf{a}, \mathbf{b})$. However, Chen et al. (2013) argued that PMD is not consistent when the true covariance structure is not close to identity matrix. A similar idea to PMD is the work proposed by Suo et al. (2017), which can be expressed as :

$$\begin{aligned} \min_{\mathbf{a}\in R^p,\, \mathbf{b}\in R^q} & -\text{Cov}(X\mathbf{a}, Y\mathbf{b}) + \tau_1\|\mathbf{a}\|_1 + \tau_2\|\mathbf{b}\|_1, \\ \text{subject to } & \text{Var}(X\mathbf{a}) \leq 1,\ \text{Var}(Y\mathbf{b}) \leq 1, \end{aligned} \tag{3.15}$$

where τ_1 and τ_2 are tuning parameters. The objective in (3.15) is a biconvex problem and can be addressed by an iterative process of alternately solving $\mathbf{a}$ and $\mathbf{b}$.

The next method to be elaborated is a very recent one proposed by Mai and Zhang (2019). They introduced a sparse CCA method (SCCA) by recasting the CCA problem into a least square problem

and imposing Lasso penalty to obtain sparse solutions. The kth pair of sparse canonical vectors $(\mathbf{a}_k, \mathbf{b}_k)$ from SCCA could be obtained by:

$$\min_{\mathbf{a}_k \in R^p,\, \mathbf{b}_k \in R^q} \{\frac{1}{2n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{a}_k - \mathbf{y}_i^T\mathbf{b}_k)^2 + \mathbf{a}_k^T(\sum_{l<k}\hat{\rho}_l\hat{\Sigma}_{xx}\hat{\mathbf{a}}_l\hat{\mathbf{b}}_l^T\hat{\Sigma}_{yy})\mathbf{b}_k + \lambda_1\|\mathbf{a}_k\|_2 + \lambda_2\|\mathbf{b}_k\|_1\},$$
$$\text{subject to } \hat{\mathbf{a}}_k^T\hat{\Sigma}_{xx}\hat{\mathbf{a}}_k = 1,\ \hat{\mathbf{b}}_k^T\hat{\Sigma}_{yy}\hat{\mathbf{b}}_k = 1, \tag{3.16}$$

where $\hat{\rho}_l = \hat{\mathbf{a}}_l^T\hat{\Sigma}_{xy}\hat{\mathbf{b}}_l$, and λ_1, λ_2 are positive tuning parameters. As mentioned by Mai and Zhang (2019), the first term $(\mathbf{x}_i^T\mathbf{a}_k - \mathbf{y}_i^T\mathbf{b}_k)^2$ is a least square term that measures the dependence between $\mathbf{x}_i^T\mathbf{a}_k$ and $\mathbf{y}_i^T\mathbf{b}_k$, and smaller values will yield larger correlations. The second term in (3.16) is adjusting the variability that has already been explained by the first $k-1$ canonical pairs, which is similar to the purpose of orthogonality constraints in the traditional CCA.

3.5 Trace Regularization on High-dimensional CCA

Next, we will discuss the application of the trace regularization method by Ahn et al. (2020) on CCA problems. Similar to LDA, the traditional CCA can also be formulated into a trace ratio optimization problem. Let $\mathbf{z}_1 = [\mathbf{a}_1^T, \mathbf{b}_1^T]^T$ and set

$$M_{d\times d} = \begin{bmatrix} \mathbf{0}_{p\times p} & \Sigma_{xy} \\ \Sigma_{yx} & \mathbf{0}_{q\times q} \end{bmatrix}, \quad Q_{d\times d} = \begin{bmatrix} \Sigma_{xx} & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \Sigma_{yy} \end{bmatrix}, \tag{3.17}$$

where $d = p + q$. Then according to (3.12a) and (3.12b), the desired first canonical pairs will satisfy:

$$M\mathbf{z}_1 = \rho Q\mathbf{z}_1, \tag{3.18}$$

whose solution will maximize $Tr(\mathbf{z}^T M\mathbf{z})$ with respect to the constraint: $\mathbf{z}^T Q\mathbf{z} = 1$. Assume that the number of population canonical pairs is m, let $Z_{d\times m} = [A^T, B^T]^T$, where $A_{p\times m} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ and

$B_{q\times m} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$ are the collections of the canonical vectors, then the objective of CCA can also be written as:

$$\max_{Z\in R^{d\times m}} \frac{Tr(Z^T M Z)}{Tr(Z^T Q Z)}, \tag{3.19}$$

which is clearly a trace ratio optimization problem in a dimension of $p+q$. Therefore, as discussed in Section 3.2, we can apply the trace regularization method here and achieve the target canonical vectors by:

$$\max_{Z\in R^{d\times m}} Tr(Z^T M Z), \text{ subject to } Tr(Z^T Q Z) = c. \tag{3.20}$$

3.5.1 Low Rank Approximation

Different from high-dimensional LDA, in which the between-class scatter and within-class scatter are linearly independent with probability one, the two matrices in the trace optimization for high-dimensional CCA have exactly the same column space, as shown by lemma 3.5.1. Therefore, the idea of maximizing $Tr[Z^T M Z]$ while regularizing $Tr[Z^T Q Z]$ seems not reasonable.

Lemma 3.5.1. *When $p, q \gg n$ (high-dimensional setting), the column space of M and Q defined in CCA* (3.17) *are exactly the same.*

Proof. Based on matrix algebra, we can see that:

$$\begin{aligned}
\text{rank}(\Sigma_{xx}) &= \min(n-1, p) = n-1;\\
\text{rank}(\Sigma_{yy}) &= \min(n-1, q) = n-1;\\
\text{rank}(\Sigma_{xy}) &= \text{rank}(\Sigma_{yx}) = \min(n-1, p, q) = n-1;\\
\text{rank}(M) &= \text{rank}(\Sigma_{xy}) + \text{rank}(\Sigma_{yx}) = 2(n-1);\\
\text{rank}(Q) &= \text{rank}(\Sigma_{xx}) + \text{rank}(\Sigma_{yy}) = 2(n-1);\\
\text{rank}([M, Q]) &\geq \text{rank}(Q).
\end{aligned}$$

Since

$$\begin{aligned}
\text{rank}([M,Q]) &= \text{rank}(\begin{bmatrix} \mathbf{0}_{p\times p} & \Sigma_{xy} & \Sigma_{xx} & \mathbf{0}_{p\times q} \\ \Sigma_{yx} & \mathbf{0}_{q\times q} & \mathbf{0}_{q\times p} & \Sigma_{yy} \end{bmatrix}) \\
&= \text{rank}(\begin{bmatrix} \mathbf{0}_{p\times p} & \Sigma_{xy} & \Sigma_{xx} & \mathbf{0}_{p\times q} \end{bmatrix}) + \text{rank}(\begin{bmatrix} \Sigma_{yx} & \mathbf{0}_{q\times q} & \mathbf{0}_{q\times p} & \Sigma_{yy} \end{bmatrix}) \\
&= \text{rank}(\begin{bmatrix} \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}) + \text{rank}(\begin{bmatrix} \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}) \\
&\leq \min(n-1,p) + \min(n-1,q) = 2(n-1) = \text{rank}(Q),
\end{aligned}$$

and we also have $\text{rank}([M,Q]) \geq \text{rank}(Q)$, thus $\text{rank}([M,Q]) = \text{rank}(Q)$, so $\mathcal{C}(M) \subset \mathcal{C}(Q)$.

On the other hand, $\text{rank}(M) = \text{rank}(Q) = \text{rank}([M,Q])$, thus $\mathcal{C}(Q) \subset \mathcal{C}(M)$.

Above all, $\mathcal{C}(M) = \mathcal{C}(Q)$, i.e., M and Q have exactly the same column space. □

However, although the column spaces of M and Q are the same, the interpretations of M and Q are actually different in the sense that M accounts for the correlations between X and Y, and Q consists of the information of the variance of X and Y. Then, it is reasonable to achieve the linearly independent piece of the column space of M and Q, where an appropriate low rank approximation will help. A basic low-rank approximation problem finds a matrix $\widetilde{W}$ that is close to the original matrix W but with a lower rank r, i.e., it solves the objective:

$$\min_{\widetilde{W}} \ \|W - \widetilde{W}\|_F, \text{ subject to } \text{rank}(\widetilde{W}) \leq r, \tag{3.21}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The analytic solution to (3.21) is the truncated singular value decomposition at the first r terms (Eckart & Young, 1936).

According to the canonical pair model proposed by Chen et al. (2013), the sequence of canonical vectors $(\mathbf{a}_k, \mathbf{b}_k)$ satisfies $\Sigma_{xy} = \Sigma_{xx}(\sum_{k=1}^{m} \rho_k \mathbf{a}_k \mathbf{b}_k^T)\Sigma_{yy}$ where m is the number of population canonical correlations. It is then reasonable to assume that Σ_{xy} and Σ_{yx} can be well-approximated by a lower rank matrix with rank m. Thus, M can be well-approximated by a lower rank matrix M_{low} with rank $2m$.

Consequently, Q can be approximated by a lower rank matrix Q_{low} with rank $2(n-1)-2m$ such that the approximates of M and Q are independent.

The approximations of M and Q are obtained by replacing Σ_{xy}, Σ_{xx} and Σ_{yy} in (3.21) with their corresponding lower rank approximations. Let $r_M = m$ and $r_Q = (n-1)-m$. Take Σ_{xy} as an example, given the SVD of $\Sigma_{xy} = \sum_k d_k \mathbf{u}_k \mathbf{v}_k^T$, where d_ks are singular values, $\mathbf{u}_k$ and $\mathbf{v}_k$ are left and right singular vectors, then Σ_{xy} can be approximated by $\widetilde{\Sigma}_{xy} = \sum_k^{r_M} d_k \mathbf{u}_k \mathbf{v}_k^T$, i.e., the first r_M terms of SVD. Q_{low} can be obtained similarly. After that, we replace M and Q with their low-rank approximations in the trace regularization method (3.20) and solve it by a generalized eigenvalue decomposition of $M_{low} - \tau Q_{low}$.

3.5.2 Row-wise Soft Thresholding for Sparsity

When dimensions are high, we would like to achieve a sparse estimation of the canonical vectors for better interpretations. We implement the sparse version of DTR method proposed in Ahn et al. (2020), which uses the group-wise soft-thresholding. Specifically, given the non-sparse solution to (3.20) denoted as $\tilde{Z}$, the sparse estimation denoted as $\hat{Z}$ can be obtained by:

$$\hat{\mathbf{z}}_j = \max(0, 1 - \frac{\lambda}{\|\tilde{\mathbf{z}}_j\|_2})\tilde{\mathbf{z}}_j, \tag{3.22}$$

where $\hat{\mathbf{z}}_j$ and $\tilde{\mathbf{z}}_j$ are the jth row vector of $\hat{Z}$ and $\tilde{Z}$, respectively, and $\lambda > 0$ is a tuning parameter. It is then easy to see that the maximum value of λ that yields a non-trivial solution to (3.22) is $\lambda_{max} = \max \|\tilde{\mathbf{z}}_j\|_2$.

In practice, we re-parameterize the method to make the tuning range bounded. The eigenvalue decomposition with be performed on $(1-\alpha)M_{low} - \alpha Q_{low}$ with $\alpha \in [0, 1]$.

Above all, the details of the trace regularization on high-dimensional CCA can be found in Algorithm 3. Note that there is no iterative process in the algorithm and the computation time is less compared with other algorithms that require convergence.

Algorithm 3: Trace Regularization on High-dimensional CCA

Given α and λ, and the number of desired canonical pairs m,

I. Initialization: Calculate the sample covaraince estimate: $\hat{\Sigma}_{xx}$ $\hat{\Sigma}_{yy}$ and $\hat{\Sigma}_{xy}$;

II. Low rank approximation: Construct M, Q and get their low rank approximations M_{low}, Q_{low} ;

III. Eigen-value decomposition: Let $S = (1-\alpha)M_{low} - \alpha Q_{low}$, obtain $\tilde{Z}$ that consists of the first m eigenvectors of S ;

IV. Sparse Estimation: Obtain the jth row of the sparse solution $\hat{Z}$ by:

$$\hat{\mathbf{z}}_j = \max(0, 1 - \frac{\lambda}{\|\tilde{\mathbf{z}}_j\|_2})\tilde{\mathbf{z}}_j$$

Then $\hat{A}$ consists of the first p row vectors of $\hat{Z}$, $\hat{B}$ consists of the $p+1$ to $p+q$ rows of $\hat{Z}$.

3.6 Simulation Studies

In this section, we carry out a simulation study to measure the empirical performance of the trace regularization method (TR) on estimating high-dimensional canonical vectors and compared with the penalized matrix decomposition (PMD) by D. M. Witten et al. (2009) as well as the sparse CCA (SCCA) by Mai and Zhang (2019).

We evaluate the methods with respect to the following metrics. (1) Error of the estimated canonical vectors (Mai & Zhang, 2019): $\text{Err}(\hat{A}) = \|P_{\hat{A}} - P_{A_0}\|_F$ and $\text{Err}(\hat{B}) = \|P_{\hat{B}} - P_{B_0}\|_F$, where $\hat{A}$ (or $\hat{B}$) and A_0 (or B_0) denotes the estimated and true canonical vectors, respectively, $P_A = A(A^TA)^{-1}A^T$ is the projection matrix onto the column space of A. (2) Number of variables selected in the estimated canonical vectors: $\text{NZ}(\hat{A})$ and $\text{NZ}(\hat{B})$. Note that when there are more than one column vectors in $\hat{A}$ (or $\hat{B}$), $\text{NZ}(\hat{A})$ (or $\text{NZ}(\hat{B})$) will consider the variables as long as they are selected by one column vector in $\hat{A}$ (or $\hat{B}$). (3) Matthew correlation coefficient (MCC) (Matthews, 1975), which is essentially a correlation measurement between the predicted and observed class labels in a binary classification. Under the context

of high-dimensional CCA, we define the following:

$$TP(TN) : \text{number of signal (noise) variables that are selected,}$$

$$FP : \text{number of noise variables that are selected,}$$

$$FN : \text{number of signal variables that are not selected.}$$

Then MCC is expressed as:

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

which is bounded in the interval of $[-1, 1]$. A MCC of 1 implies a perfect prediction of the signal and noise variables; a MCC of 0 implies a prediction that is no better than random guessing; a MCC of -1 implies a total disagreement between predictions and observations. Note that, when there are multiple canonical pairs, we only consider the MCC evaluated at the first canonical vector pair. (4) When the true canonical correlation is known, we also evaluate how far the estimated correlation is apart from the true canonical correlation.

Next, we discuss the tuning process of each method. Based on the results form empirical tries, for TR, we vary α in an equally spaced grid on $[0.1, 0.5]$ with 10 values and λ in a log-scale grid on $[0, \lambda_{max}]$ with 30 values. Note that depends on the specific setting, we may slightly reduce the tuning range and the number of grids for computational efficiency. For each training set (X_{train}, Y_{train}), we generate a separate tuning set (X_{tune}, Y_{tune}) which has exactly the same distribution as the training set, and select the optimal parameter pair $(\alpha_{opt}, \lambda_{opt})$ that minimizes:

$$|\text{corr}(X_{train}\hat{A}_1, Y_{train}\hat{B}_1) - \text{corr}(X_{tune}\hat{A}_1, Y_{tune}\hat{B}_1))|,$$

where $\hat{A}_1$ (or $\hat{B}_1$) is the first estimated canonical vector. For SCCA, we vary the tuning parameter in an equally space grid on $[0.01, 0.2]$ with 20 values and select the best tuning parameter that yields the largest first canonical correlation on the tuning set. For PMD, we use the Lasso penalty on both A and B , the

optimal tuning parameter is chosen by permutation tests implemented by the function CCA.permute in the R package **PMA** (D. Witten et al., 2020).

In the following, we will introduce how the simulations are designed.

Simulation I The first simulation uses the same settings implemented in the work of Safo (2014) and the number of population canonical pairs is fixed to be 1. The sample size is $n = 80$ and the dimensions of X and Y are $p = 200$ and $q = 150$, respectively. In addition, the population covariance matrices Σ_{xx}, Σ_{yy} and Σ_{yx} are pre-defined and the true canonical vectors $\mathbf{a}$ and $\mathbf{b}$ will be obtained as the first left and right singular vector from the SVD of $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$. The true canonical correlation is then the first singular value from the SVD. The numbers of non-zero coefficients (signal variables) in $\mathbf{a}$ and $\mathbf{b}$ are set to be 10% of the total number of variables, i.e., 20 and 15, respectively. The data matrices $(X_{n\times p}, Y_{n\times q})$ are generated from a multivariate normal distribution $\text{MVN}(\mathbf{0}_{p+q}, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}.$$

We consider the same covariance structures as Safo (2014), which are introduced in the following. Let $pn = p - 20$, $qn = q - 15$ be the number of noise variables for X and Y, respectively. For simplicity, we use J to denote a matrix whose elements are all 1. We use $C(s)_l$ to denote a compound symmetry matrix with correlation s and dimension $l \times l$:

$$C(s)_l = \begin{bmatrix} 1 & s & \dots & s \\ s & 1 & \dots & s \\ \vdots & \vdots & \dots & \vdots \\ s & s & \dots & 1 \end{bmatrix}.$$

The following four settings of the population covariance structures in the Simulation I consider different correlations among signal variables and noise variables.

- Setting 1. The signal correlations in X and Y are 0.7, the noise correlations are both 0.1.

$$\Sigma_{xx} = \begin{bmatrix} C(0.7)_{20} & \mathbf{0} \\ \mathbf{0} & C(0.1)_{pn} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} C(0.7)_{15} & \mathbf{0} \\ \mathbf{0} & C(0.1)_{qn} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.6J_{20\times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

- Setting 2. The signal correlations in X and Y are 0.7, the noise correlations are both 0.

$$\Sigma_{xx} = \begin{bmatrix} C(0.7)_{20} & \mathbf{0} \\ \mathbf{0} & I_{pn} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} C(0.7)_{15} & \mathbf{0} \\ \mathbf{0} & I_{qn} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.6J_{20\times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

- Setting 3. The signal correlations in X and Y are 0.5 and 0.3, respectively. The noise correlations are 0.3 and 0.1, respectively.

$$\Sigma_{xx} = \begin{bmatrix} C(0.5)_{20} & \mathbf{0} \\ \mathbf{0} & C(0.3)_{pn} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} C(0.3)_{15} & \mathbf{0} \\ \mathbf{0} & C(0.1)_{qn} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.35J_{20\times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

- Setting 4. The signal correlations in X and Y are 0.5 and 0.3, respectively. The noise correlations are both 0.

$$\Sigma_{xx} = \begin{bmatrix} C(0.5)_{20} & \mathbf{0} \\ \mathbf{0} & I_{pn} \end{bmatrix}, \Sigma_{yy} = \begin{bmatrix} C(0.3)_{15} & \mathbf{0} \\ \mathbf{0} & I_{qn} \end{bmatrix}, \Sigma_{xy} = \begin{bmatrix} 0.35J_{20\times 15} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Simulation II In the second simulation, we keep the same population covariance structures as Simulation I, but change the underlying distribution from multivariate normal to multivariate t distribution with $df = 3$.

Simulation III In the third simulation, we consider the settings in D. M. Witten et al. (2009) and extend it to t distribution as well. Specifically, the sample size is set to be $n = 50$, and the dimensions are set to be $p = q = 100$. There are two true canonical pairs, $(\mathbf{a}_1, \mathbf{b}_1)$, $(\mathbf{a}_2, \mathbf{b}_2)$ with the coefficients

defined as follows:

$$\mathbf{a}_1 = (\mathbf{1}_{20}, -\mathbf{1}_{20}, \mathbf{0}_{60})^T,$$

$$\mathbf{a}_2 = (-\mathbf{1}_{10}, \mathbf{1}_{10}, -\mathbf{1}_{10}, \mathbf{1}_{10}, \mathbf{0}_{60})^T,$$

$$\mathbf{b}_1 = (\mathbf{0}_{60}, -\mathbf{1}_{20}, \mathbf{1}_{20})^T,$$

$$\mathbf{b}_2 = (\mathbf{0}_{60}, \mathbf{1}_{10}, -\mathbf{1}_{10}, \mathbf{1}_{10}, -\mathbf{1}_{10})^T.$$

Let $\mathbf{w}_1, \mathbf{w}_2$ be two orthonormal vectors with length of n that connect X and Y. We consider both normal distribution and non-central t distribution.

- Simulation III - Setting 1.

$$X_{ij} \sim N(w_{1i}a_{1j} + w_{2i}a_{2j}, 0.09),\ Y_{ij} \sim N(w_{1i}b_{1j} + w_{2i}b_{2j}, 0.09),$$

  where w_{1i} and w_{2i} are the ith elements from $\mathbf{w}_1$ and $\mathbf{w}_2$, respectively.

- Simulation III - Setting 2.

$$X_{ij} \sim t(df = 10, \delta = w_{1i}a_{1j} + w_{2i}a_{2j}), \quad Y_{ij} \sim t(df = 10, \delta = w_{1i}b_{1j} + w_{2i}b_{2j}),$$

  where δ is the non-centrality parameter in the t distribution.

3.6.1 Results

For each simulation setting, we generate 100 training sets and 100 tuning sets, and calculate the evaluation metrics averaged on the 100 training sets.

Figure 3.2 shows the average estimated error, number of selected variables and the MCC values under the four settings in Simulation I. In setting 1, 2 and 4, we see that TR obtains the lowest error which is significantly lower than the other two methods. Also, the MCC values from TR are the highest in setting 1, 2 and 4. In setting 3, PMD has the lowest error, SCCA and TR have relatively larger and similar error and SCCA has the largest MCC values. In general, SCCA yields the most sparse estimation in setting 1, 3

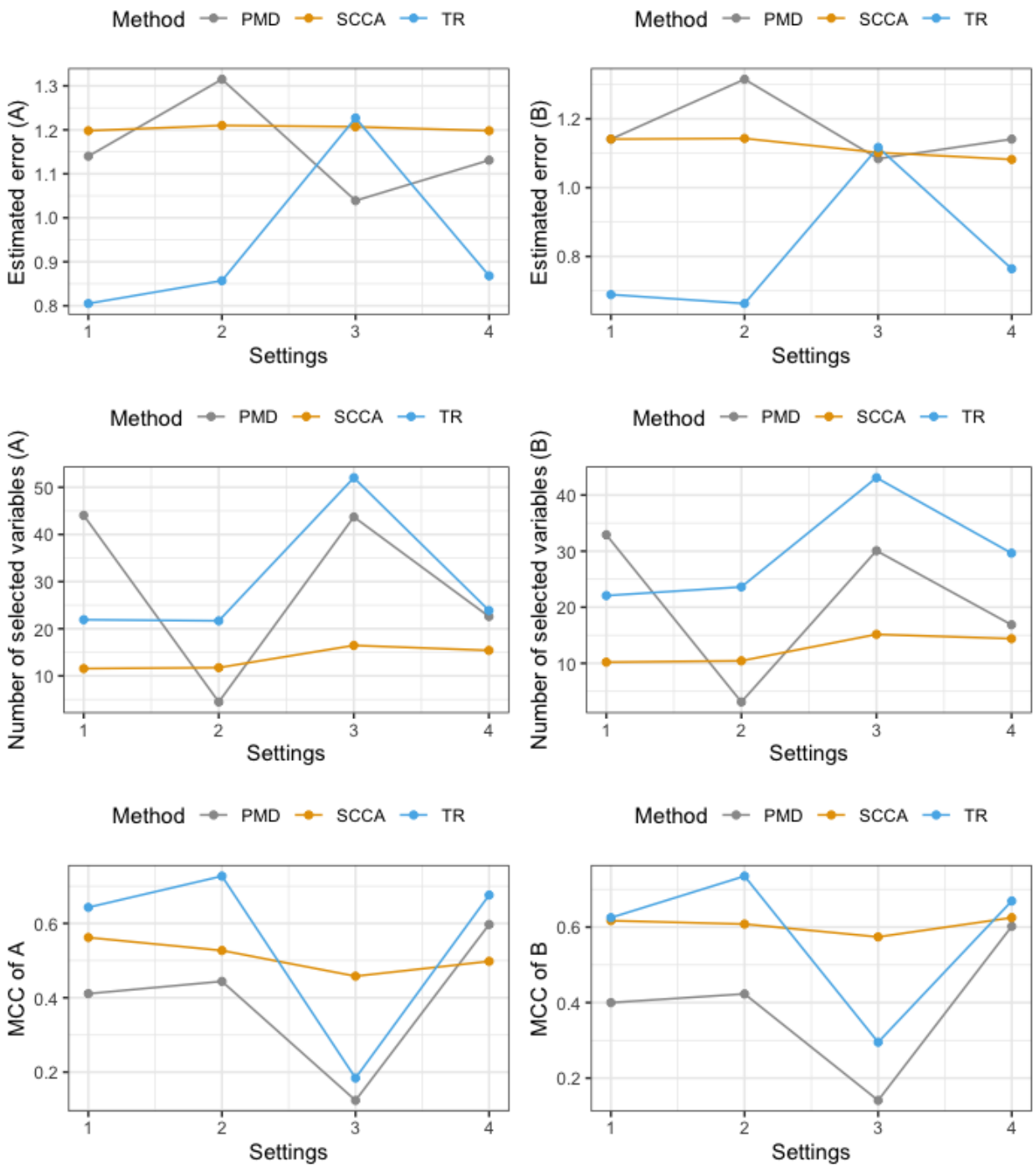


Figure 3.2: Results of Simulation I. The average estimation error, number of selected variables, and MCC over 100 simulated data sets. The three rows show the three metrics on the y-axis, with setting number presented in the x-axis.

and 4, presented by the second row in the figure and PMD achieves the most sparse one in setting 2. TR does not achieve the most sparse solution but produces the highest MCC in setting 1, 2 and 4, indicating that it has a superior performance in variable selectivity in these settings. TR does not show its superior performance in setting 3, in which the noise variables are higher correlated compared with other settings. We also see that the performance of SCCA is comparably stable across the four settings compared with the other two methods, indicating that the correlations among signal and noise variables would make less influence on the estimations from SCCA.

Figure 3.3 shows the results under the four settings in Simulation II. TR achieves the lowest estimated error in setting 1 and 2, but a higher error in setting 3 and 4. In terms of MCC, TR does not perform well, which may due to the fact that less sparsity is obtained in TR compared with the other two methods. Comparably, under the t distribution, TR still works better when the correlations among noise variables are smaller. Different from settings of the normal distribution in which SCCA is rather stable, with t distribution, SCCA performs worse when the noise correlations are relatively large. In terms of sparsity, the estimations from SCCA are almost the most sparse ones except for setting 2, in which PMD is the most sparse one (same as Simulation I). In generally, all the methods perform worse when the underlying distribution is changed from normal distribution to t distribution.

We also compare the estimated first canonical correlations for the three methods, as shown in Figure 3.4. In Simulation I and II, SCCA overestimates the true correlation. Comparably, TR and PMD obtain more close estimations to the true canonical correlation, and PMD is more conservative. Notice that the estimated canonical correlations from TR and PMD both drop in setting 3 when the noise correlations are relatively large, indicating the impacts of noise correlations.

Table 3.1 shows the results for Simulation III. In general, all the three methods perform better in the normal scenario (setting 1) than the t scenario (setting 2). In both settings, TR has the highest MCC values and its superior performance is much more obvious in setting 1. The solutions from SCCA are the still the most sparse ones.

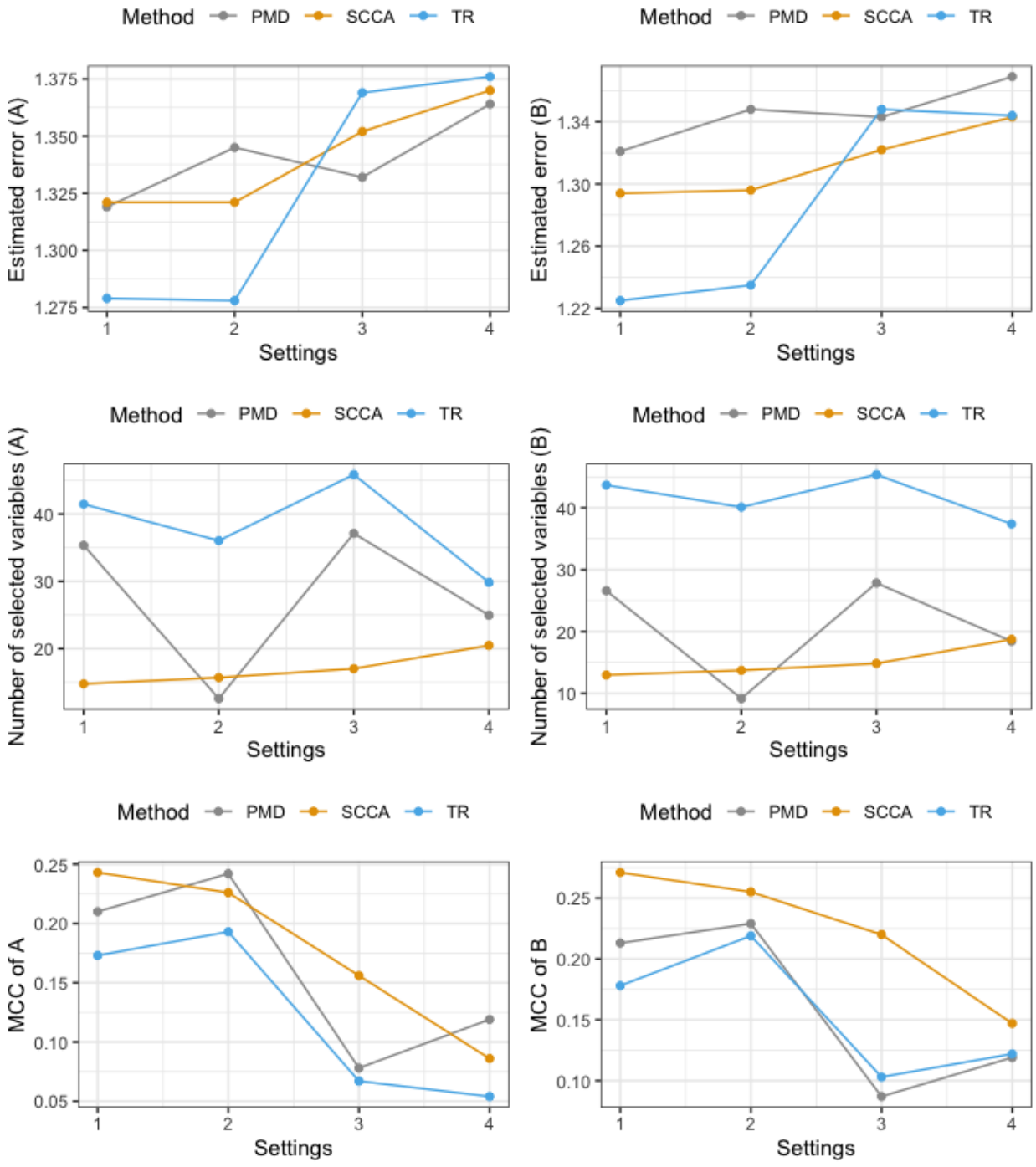
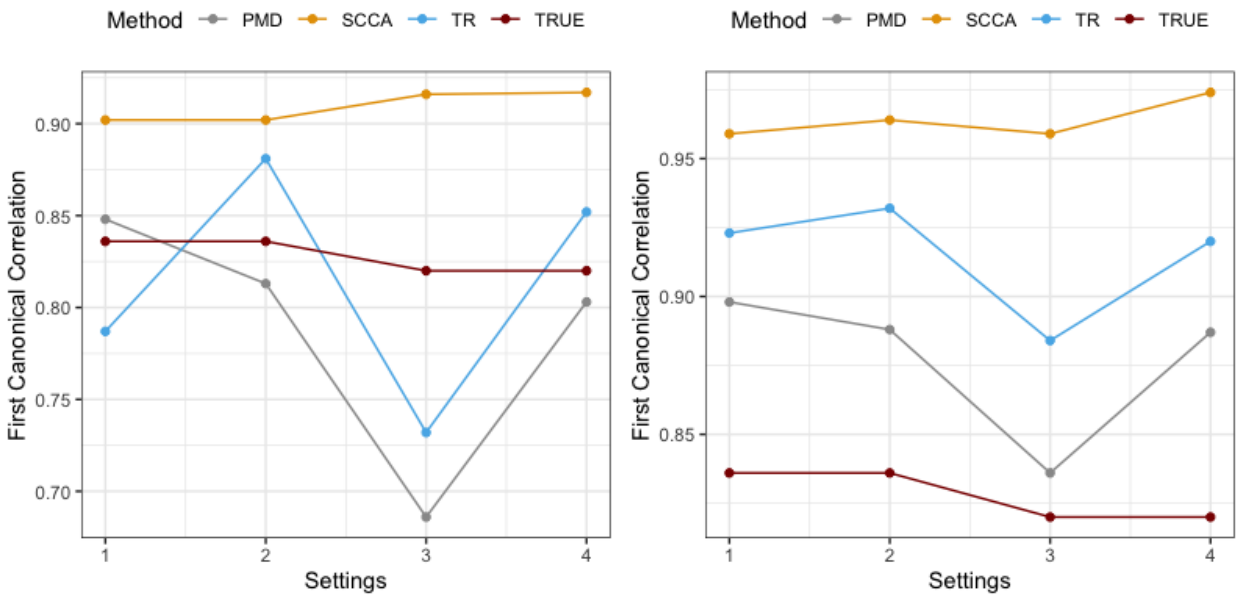


Figure 3.3: Results for Simulation II. The average estimation error, number of selected variables, and MCC over 100 simulated data sets.



(a) Simulation I

(b) Simulation II

Figure 3.4: The average first estimated canonical correlations in Simulation I and Simulation II. True canonical correlations are marked as red lines.

Table 3.1: Results of Simulation III. The metrics are averaged over 100 repetitions. Standard deviations are given inside of the parenthesis. Best metrics are bolded.

Setting 1						
Methods	Err($\hat{A}$)	Err($\hat{B}$)	NZ($\hat{A}$)	NZ($\hat{B}$)	$MCC(\hat{A})$	$MCC(\hat{B})$
PMD	**1.047**	**1.047**	69.620	69.510	0.283	0.281
	(0.331)	(0.330)	(33.389)	(33.533)	(0.184)	(0.172)
SCCA	1.632	1.640	**33.222**	**33.283**	0.305	0.294
	(0.077)	(0.075)	(6.960)	(6.136)	(0.069)	(0.069)
TR	1.190	1.177	45.050	44.970	**0.665**	**0.664**
	(0.353)	(0.362)	(19.651)	(19.798)	(0.201)	(0.215)
Setting 2						
Methods	Err($\hat{A}$)	Err($\hat{B}$)	NZ($\hat{A}$)	NZ($\hat{B}$)	$MCC(\hat{A})$	$MCC(\hat{B})$
PMD	1.954	1.958	44.432	44.500	0.005	−0.007
	(0.038)	(0.035)	(34.017)	(33.824)	(0.107)	(0.097)
SCCA	1.974	1.971	**24.323**	**24.061**	−0.004	0.018
	(0.022)	(0.020)	(8.600)	(8.252)	(0.074)	(0.067)
TR	**1.942**	**1.943**	51.200	49.950	**0.050**	**0.038**
	(0.039)	(0.039)	(24.921)	(24.242)	(0.109)	(0.109)

Above all, TR can achieve a promising performance under certain circumstances, in terms of achieving a higher value of MCC. We find that under the setting 1, 2 and 4 in Simulation I, in which the underlying distribution is multivariate normal distribution and the noise correlations are small, TR performs much better than the other two methods. When the noise correlations increase, methods will generally perform worse. In addition, we also find that the performance of all the three methods are worse when the data are t distributed compared with normal distributed. In Simulation III when there are two canonical vectors, TR still works better than the other two methods in terms of MCC. Notice that TR generally does not achieve the most sparse solution but obtains a higher MCC, indicating that TR does well in variable selectivity.

3.7 Discussions and Conclusions

In this chapter, we introduce the trace ratio optimization problem involved in dimension reduction techniques, which is always solved by a conventional sub-optimal ratio trace problem that becomes degenerate when the dimensions are high. Although there are many efforts that optimize the trace ratio problem directly, these work generally lack statistical interpretations. Comparably, the trace regularization method proposed by Ahn et al. (2020) is more emphasized on the interpretations and at the same time optimizing the trace ratio. However, the trace regularization method was proposed under the context of discriminating problems. Given the promising performance of it, we apply the trace regularization method on the non-discriminating context, i.e., the canonical correlation analysis (CCA), a well-known dimension reduction technique. We modify the trace regularization method by adding low rank approximations such that it can be applied to the high-dimensional CCA. A sparse version of the solution is obtained by the group-wise soft-thresholding.

The results from different simulation settings suggest that the trace regularization method does a better job in variable selectivity (measured by MCC) compared with some well-known sparse CCA methods, when the correlations among noise variables are small and the underlying distribution of the data is normal

distribution. However, there could be other settings that the trace regularization might not be that strong, which needs future investigations.

In addition, we use the lower-rank approximation by SVD to separate the column space of M and Q, along with a reasonable choice of the target ranks of the two matrices. The choice of the ranks works well in practice, however, there still remain further investigations on the theoretical justifications. Also, there might be other methods that can achieve a better approximation of the matrices that would worth a try. What is more, it is reasonable to assume that adding sparsity in the low-rank approximations might increase the estimation accuracy of the true canonical vectors. In terms of the sparse estimation, we use the group-wise soft-thresholding on the row vectors which is very straightforward and implementable-friendly. There might be other algorithms that could also be applied on the trace regularization to achieve a sparse solution (Ma et al., 2013). In the future, the trace regularization method could also be studied to be extended to the problem of multi-set canonical correlation analysis (Parra, 2018).

Bibliography

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Ahn, J., Chung, H. C., & Jeon, Y. (2020). Trace ratio optimization for high-dimensional multi-class discrimination. *Journal of Computational and Graphical Statistics*, 1–12.

Aliferis, C. F., Statnikov, A., & Tsamardinos, I. (2006). Challenges in the analysis of mass-throughput data: A technical commentary from the statistical machine learning perspective. *Cancer Informatics*, *2*, 117693510600200004.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*(6), 1323–1333.

Archer, K. J., Hou, J., Zhou, Q., Ferber, K., Layne, J. G., & Gentry, A. E. (2014). Ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Informatics*, *13*, CIN–S20806.

Archer, K. J., & Williams, A. A. (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, *31*(14), 1464–1474.

Bellman, R. E. (2015). *Adaptive control processes: A guided tour*. Princeton university press.

Bickel, P. J., Levina, E. et al. (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.

BladÉ, J., Samson, D., Reece, D., Apperley, J., BJÖrkstrand, B., Gahrton, G., Gertz, M., Giralt, S., Jagannath, S., & Vesole, D. (1998). Criteria for evaluating disease response and progression in patients with multiple myeloma treated by high-dose therapy and haemopoietic stem cell transplantation.

Myeloma Subcommittee of the EBMT. European Group for Blood and Marrow Transplant. *British Journal of Haematology*, *102*(5), 1115.

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

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Cardie, C., & Nowe, N. (1997). Improving minority class prediction using case-specific feature weights. *Proceedings of the Fourteenth International Conference on Machine Learning*, 57–65.

Cardoso, J. S., & Costa, J. F. (2007). Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, *8*(Jul), 1393–1429.

Cardoso, J. S., Sousa, R., & Domingues, I. (2012). Ordinal data classification using kernel discriminant analysis: A comparison of three approaches. *2012 11th International Conference on Machine Learning and Applications*, *1*, 473–477.

Chalise, P., & Fridley, B. L. (2012). Comparison of penalty functions for sparse canonical correlation analysis. *Computational Statistics & Data Analysis*, *56*(2), 245–254.

Chen, M., Gao, C., Ren, Z., & Zhou, H. H. (2013). Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., & Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, *103*(7), 2771–2778.

Chu, W., & Keerthi, S. S. (2005). New approaches to support vector ordinal regression. *Proceedings of the 22nd international conference on Machine learning*, 145–152.

Clemmensen, L., Hastie, T., Witten, D., & Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, *53*(4), 406–413.

Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, *2*(Dec), 265–292.

de La Torre, J., Puig, D., & Valls, A. (2018). Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, *105*, 144–154.

Dietterich, T. G., & Bakiri, G. (1994). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, *2*, 263–286.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. (1997). Support vector regression machines. *Advances in neural information processing systems*, *9*, 155–161.

Dudoit, S., Fridlyand, J., & Speed, T. 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.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economic and Statistics*, 92–107.

Fienberg, S. E. (1980). The analysis of cross-classified categorical data. *Massachusetts Institute of Technology Press, Cambridge and London*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Fix, E. (1985). *Discriminatory analysis: Nonparametric discrimination, consistency properties* (Vol. 1). USAF school of Aviation Medicine.

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. *European Conference on Machine Learning*, 145–156.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*(405), 165–175.

Gao, C., Ma, Z., & Zhou, H. H. (2014). An efficient and optimal method for sparse canonical correlation analysis. *ArXiV e-prints*, *1409*.

Goodman, L. A., & Kruskal, W. H. (1959). Measures of association for cross classifications. ii: Further discussion and references. *Journal of the American Statistical Association*, *54*(285), 123–163.

Goodman, L. A., & Kruskal, W. H. (1963). Measures of association for cross classifications iii: Approximate sampling theory. *Journal of the American Statistical Association*, *58*(302), 310–364.

Goodman, L. A., & Kruskal, W. H. (1972). Measures of association for cross classifications, iv: Simplification of asymptotic variances. *Journal of the American Statistical Association*, *67*(338), 415–421.

Goodman, L. A., & Kruskal, W. H. (1979). Measures of association for cross classifications. *Measures of association for cross classifications*, 2–34.

Guo, Y.-F., Li, S.-J., Yang, J.-Y., Shu, T.-T., & Wu, L.-D. (2003). A generalized foley–sammon transform based on generalized fisher discriminant criterion and its application to face recognition. *Pattern Recognition Letters*, *24*(1-3), 147–158.

Gutierrez, P. A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., & Hervas-Martinez, C. (2015). Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 127–146.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, *62*(4), 426–433.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 73–102.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, *89*(428), 1255–1270.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28.

Herbrich, R., Graepel, T., & Obermayer, K. (1999). Support vector learning for ordinal regression.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Horel, A. (1962). Applications of ridge analysis toregression problems. *Chem. Eng. Progress.*, *58*, 54–59.

Hotelling, H. (1992). Relations between two sets of variates. *Breakthroughs in statistics* (pp. 162–190). Springer.

Huhn, J. C., & Hüllermeier, E. (2008). Is an ordinal class structure useful in classifier learning? *IJDMMM*, *1*(1), 45–67.

Hunger, S. P., & Mullighan, C. G. (2015). Acute lymphoblastic leukemia in children. *New England Journal of Medicine*, *373*(16), 1541–1552.

Izenman, A. J. (2008). Modern multivariate statistical techniques. *Regression, Classification and Manifold Learning*, *10*, 978–.

Jia, Y., Nie, F., & Zhang, C. (2009). Trace ratio problem revisited. *IEEE Transactions on Neural Networks*, *20*(4), 729–735.

Jung, S., Ahn, J., & Jeon, Y. (2019). Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem. *Journal of Computational and Graphical Statistics*, *28*(3), 710–721.

Kokiopoulou, E., Chen, J., & Saad, Y. (2011). Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, *18*(3), 565–602.

Kotsiantis, S. B., & Pintelas, P. E. (2004). A cost sensitive technique for ordinal classification problems. *Hellenic Conference on Artificial Intelligence*, 220–229.

Lee, M. H., Ahn, J., & Jeon, Y. (2013). Hdlss discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics*, *22*(2), 433–451.

Lee, Y., Lin, Y., & 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*(465), 67–81.

Leha, A., Jung, K., & Beißbarth, T. (2013). Utilization of ordinal response structures in classification with high-dimensional expression data. *German Conference on Bioinformatics 2013*.

Li, X. (2019). *All: A data package* [R package version 1.28.0].

Liu, X., O'Connell, A. A., & Koirala, H. (2011). Ordinal regression analysis: Predicting mathematics proficiency using the continuation ratio model. *Journal of Modern Applied Statistical Methods*, *10*(2), 11.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, *41*(2), 772–801.

Mai, Q., & Zhang, X. (2019). An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics*, *75*(3), 734–744.

Mai, Q., & Zou, H. (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, *135*, 175–188.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

Mulligan, G., Mitsiades, C., Bryant, B., Zhan, F., Chng, W. J., Roels, S., Koenig, E., Fergus, A., Huang, Y., Richardson, P., et al. (2007). Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood*, *109*(8), 3177–3188.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, *8*(1).

Parra, L. C. (2018). Multi-set canonical correlation analysis simply explained. *arXiv preprint arXiv:1802.03759*.

Qiao, X. (2015). Noncrossing ordinal classification. *arXiv preprint arXiv:1505.03442*.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Safo, S. E. (2014). *Design and analysis issues in high dimension, low sample size problems* (Doctoral dissertation). University of Georgia.

Shashua, A., & Levin, A. (2003). Ranking with large margin principle: Two approaches. *Advances in neural information processing systems*, 961–968.

Sun, B.-Y., Li, J., Wu, D. D., Zhang, X.-M., & Li, W.-B. (2009). Kernel discriminant learning for ordinal regression. *IEEE Transactions on Knowledge and Data Engineering*, *22*(6), 906–910.

Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., et al. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, *9*(4), 287–300.

Suo, X., Minden, V., Nelson, B., Tibshirani, R., & Saunders, M. (2017). Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*.

Terragna, C., Remondini, D., Martello, M., Zamagni, E., Pantani, L., Patriarca, F., Pezzi, A., Levi, G., Offidani, M., Proserpio, I., et al. (2016). The genetic and genomic background of multiple myeloma patients achieving complete response after induction therapy with bortezomib, thalidomide and dexamethasone (vtd). *Oncotarget*, *7*(9), 9666.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*(2), 147–166.

Waaijenborg, S., de Witt Hamer, P. V., Zwinderman, A. H., et al. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, *7*(1), 3.

Wang, H., Yan, S., Xu, D., Tang, X., & Huang, T. (2007). Trace ratio vs. ratio trace for dimensionality reduction. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Wang, Y., Miller, D., & Clarke, R. (2008). Approaches to working in high-dimensional data spaces: Gene expression microarrays. *British Journal of Cancer*, *98*(6), 1023–1028.

Weiss, D. J. (1972). Canonical correlation analysis in counseling psychology research. *Journal of Counseling Psychology*, *19*(3), 241.

Witten, D. (2015). *Penalizedlda: Penalized classification using fisher's linear discriminant* [R package version 1.1]. https://CRAN.R-project.org/package=penalizedLDA

Witten, D., Tibshirani, R., Gross, S., Narasimhan, B., & Witten, M. D. (2020). Package 'pma'. *Genetics and Molecular Biology*, *8*(1), 28.

Witten, D. M., & 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., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*(3), 515–534.

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, *8*(1).

Yi, N. (2019). *Bhglm: Bayesian hierarchical glms and survival models, with applications to genomics and epidemiology* [R package version 1.1.0].

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, X., Li, B., Han, H., Song, S., Xu, H., Hong, Y., Yi, N., & Zhuang, W. (2018). Predicting multilevel drug response with gene expression profile in multiple myeloma using hierarchical ordinal regression. *BMC cancer*, *18*(1), 551–551.

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

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