LINEAR DISCRIMINANT ANALYSIS WITH HIGH DIMENSIONAL MIXED VARIABLES

YANG ZHONGQING

PhD

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University

Department of Applied Mathematics

Linear Discriminant Analysis with High Dimensional Mixed Variables

Yang Zhongqing

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
April 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: ____________________

Name of Student: Yang Zhongqing

Abstract

With the rapid development of modern measurement technologies, datasets containing both discrete and continuous variables are more and more commonly seen in different areas. In particular, the dimensions of the discrete and continuous variables can oftentimes be very high. Discriminant analysis for mixed variables under the traditional fixed dimension setting has been well studied. Despite the recent progress made in modelling high-dimensional data for continuous variables, there is a scarcity of methods that can deal with a mixed set of variables. To fill this gap, this thesis develops a novel approach for classifying high-dimensional observations with mixed variables. So in this thesis, we aim to develop a simple yet useful classification rule that addresses both the high dimensionality and the mixing structure of the variables simultaneously.

In Chapter 2-3 we introduce our framework building on a location model, in which the distributions of the continuous variables conditional on categorical ones are assumed Gaussian. We overcome the challenge of having to split data into exponentially many cells, or combinations of the categorical variables, by kernel smoothing. And provide new perspectives for its bandwidth choice to ensure an analogue of Bochner's Lemma, which is different to the usual bias-variance tradeoff. We show that the two sets of parameters in our model can be separately estimated and provide a penalized likelihood method for their estimation.

In Chapter 4, some theoretical results are shown. Efficient direct estimation schemes are developed to obtain consistent estimators of the discriminant components.

In Chapter 5, we conduct simulation studies to investigate the performance of proposed semiparametric location model. Results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by extensive simulation and real data studies.

Acknowledgements

During the last six years, I studied at The Hong Kong Polytechnic University as a PhD student. Actually, I did not have any training in statistics at my undergraduate level although I was really interested in this area. I feel so lucky to have a chance to learn a lot of things in statistics. Over the past few years, I have a deeper understanding to this field by taking some professional courses and did some research projects. I really enjoy exploring in the world of mathematics and solving some practical problems through mathematical analysis. Throughout the years, I would not be able to complete this work without the help and support from many people.

Firstly, I would like to express my deepest gratitude to my Chief Supervisor Dr. Jiang Binyan. He is an enthusiastic and knowledgeable researcher. Throughout the years, Dr. Jiang always gives recommendation to my research projects in detail that help me understand the problem clearly after each meeting we have. He is a really gracious and supportive supervisor and I am grateful to him forever.

Furthermore, I would like to express my heartfelt thanks to my co-supervisor Prof. Zhao Xingqiu. I feel a constant source encouragement and help in the guided study. It is a great honour to have her recommendation and guidance.

I would like to show my special thanks to Yu Xinyang for his help in computing area.

Finally, I am much obliged to my parents for their kind understanding. They have devoted all their love to me for past twenty-seven years and I hope to pay them back in the future.

Contents

List of Figures

List of Tables

List of Notations

$\mathbf{Z}$	p-dimensional continuous variable
$\mathbf{U}$	d-dimensional discrete variable
$\mathbf{u}$	the given d-dimensional location
$\beta(\mathbf{U})$	the location-dependent direction of semiparametric location model
$\eta(\mathbf{U})$	the location-dependent intercept of semiparametric location model
$\Phi(\cdot)$	the cumulative distribution function of the standard normal distribution
$\mathcal{D}$	the set of all functions from $\{0,1\}^d$ to $\mathbb{R}$
$\mathcal{C}_{\mathbf{u},s}$	defines the contour with radius s from the center $\mathbf{u}$
$B_{\mathbf{v}}(r)$	the ball centered at $\mathbf{v} \in \{0,1\}^d$ with radius r

Chapter 1

Introduction

1.1 Bayes' LDA

As a traditional classification technique, linear discriminant analysis(LDA) is commonly used in modern statistical research. It works well when data are in continuous type with low dimension.

Consider X as a set of random variables under the assumption of Gaussian distribution. L is the class label of X with the prior probabilities $\pi_i = Pr(L = i), i = 1, 2, ...$ The classical Bayes rule classifies a new observation into class i if $\pi_i f_i(X)$ is the maximum among all classes. The linear form can be shown as:

$$l(X) = \arg\max_i \left\{ \mu_i^T \Sigma^{-1} X - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log \pi_i \right\}$$

When $L = 2$, it would be reduced to a binary problem. In that case, the new observation would be classified into class 1 if $\pi_1 f_1 > \pi_2 f_2$. Intuitively,

$$\left(X - \left(\frac{\mu_1 + \mu_2}{2} \right) \right)^T \Sigma^{-1} (\mu_1 - \mu_2) + \log \frac{\pi_1}{\pi_2} > 0$$

For simplicity of representation, we will focus on the binary problem with only two classes.

1.2 Motivation

High dimensional data sets containing both discrete and continuous variables arise frequently in practice in the past decades. For example, many diagnosis studies in medical research entail the collection of both high dimensional continuous gene expressions and categorical features such as gender, indicator of medical history and indicators of the presence of certain symptoms. As highlighted by Fan et al. (2017), such a structure poses new challenges in statistical modeling. For discriminant analysis, several approaches have been developed in the early literature for mixed variables under the traditional fixed dimensional setting, and recent research focuses mainly on data sets with high dimensional continuous variables. Promising approaches taking into account both the high dimensionality and the mixing nature of the data sets are still missing. In this thesis, we aim to develop a simple yet useful classification rule that addresses both high dimensionality and the mixing nature of the variables, with sound theoretical justifications. In particular, we are interested in a more challenging case where the dimensions of the discrete variables and the continuous variables can both be very high.

Now we consider the problem of classifying very high-dimensional observations into categories. In a great many cases, the datasets can contain a mixed set of variables including discrete and continuous ones, both of which can be high-dimensional while the sample size is small. There are some examples below:

- Example 1. In clinical practice, it is common to collect data that come with continuous variables and discrete variables. The dimension of these features can be relatively high, while the number of patients is relatively small, especially for serious or rare diseases. For example, in the Hepatocellular Carcinoma dataset that we considered in the real data study, there are 165 patients with 22 continuous variables which are mainly from patients' medical test results, and 118 binary variables which are mainly indicators of related symptoms and medical histories.

- Example 2. In integrative analysis, the main objective is to combine different datasets for a comprehensive study. One of the possible possibilities is to integrate continuous-type data with discrete-type data. For example, the Breast Cancer Gene Expression Profiles data that we considered in our analysis consists of 489 mRNA Z-Scores (which are measurements of the relative expression of patients' genes to the reference population), and a set of indicators of mutation for 173 genes. A strong motivation for combining these two datasets is that together they may provide more information about the mortality risk, the main quantity of interest.

- Example 3. In addition to medical field, datasets with a mixed set of variables can be collected in other fields. For example, the Australian Credit Card Approval dataset we considered in our real data analysis concerns credit card applications. The dataset contains person information which consists of different types of data.

In order to deal with these mixed variables, a simple strategy is to treat the categorical variables as continuous ones and apply existing classification methods developed for continuous variables. Such a treatment ignores the nature of categorical variables and intuitively incurs loss of information. Here we present a simple example. Consider a two-class classification problem where there are one continuous variable X and one binary categorical variable U. Assume that the prior for the class label L is balanced, i.e., $P(L = 1) = P(L = 2) = 0.5$, and that $P(U = 0) = P(U = 1) = 0.5$ for both classes. For Class 1, suppose the conditional distribution of X given U

satisfies

$$X|U = 0 \sim N(-1, 1), \quad X|U = 1 \sim N(1, 1).$$

Likewise, for Class 2, assume that

$$X|U = 0 \sim N(1, 1), \quad X|U = 1 \sim N(-1, 1).$$

If we simply treat U as a continuous variable that takes value 0 or 1, and seek for the best linear classifier, the misclassification rate for the optimal linear classifier will be easily seen as $\frac{1}{2}\left[\frac{1}{2} + \Phi(-1)\right]$, which is more than twice of the optimal Bayes misclassification rate $\Phi(-1)$. Details are as follows:

(1). With some simple calculations it can be shown that the optimal Bayes classifier is: classify (X, U) into class 1 if $XU > 0$, and into class 2 otherwise. The corresponding misclassification rate can be easily established. (2). Without loss of generality, consider a classifier that classify (X, U) into class 1 if and only if $X + aU > b$ for some a and b. Let Z be a random variable following the standard normal distribution. The misclassification rate of the linear classifier can be computed as:

$$\begin{aligned} R(a, b) &= \frac{1}{4}P(X + aU > b|L = 2, U = 0) + \frac{1}{4}P(X + aU > b|L = 2, U = 1) \\ &\quad + \frac{1}{4}P(X + aU < b|L = 1, U = 0) + \frac{1}{4}P(X + aU < b|L = 1, U = 1). \end{aligned}$$

Note that $P(X + aU > b|L = 2, U = 0) = P(X > b|X \sim N(1, 1)) = P(Z > b - 1)$. Similarly, we have $P(X + aU > b|L = 2, U = 1) = P(Z > b - a + 1)$, $P(X + aU < b|L = 1, U = 0) = P(Z < b+1)$ and $P(X+aU < b|L = 1, U = 1) = P(Z < b-a-1)$. Consequently, we have

$$\begin{aligned} R(a, b) &= \frac{1}{2} + \frac{1}{4}P(b - 1 < Z < b + 1) - \frac{1}{4}P(b - a - 1 < Z < b - a + 1) \\ &\geq \frac{1}{2} - \frac{1}{4}P(-1 < Z < 1) \end{aligned}$$

$$= \frac{1}{4} + \frac{1}{2}\Phi(-1),$$

where the equal sign in the second inequality is obtained when $a = b \to \infty$. An immediate message from this simple example is that, in order to obtain a sound classifier, we may have to handle the effects of categorical variables and continuous variable differently, and seek for ways of capturing their interactions. In our setting, the challenge on the need to handle mixed variables is further exasperated by the high-dimensionality of the problem.

1.3 Literature review

For discriminant analysis, while there are early works in investigating how to model mixed variables under the traditional fixed dimensional setting, recent research has focused almost exclusively on datasets with high dimensional continuous variables. Methodologies that can effectively take into account both the high-dimensionality and the mixing nature of the datasets are scarce.

When the dimensionality is fixed, discriminant analysis with discrete and continuous data has been well studied. Some simple summaries are listed in Hamid (2010).

- A simple strategy is transforming the two different types of variables into one type. For example, Cochran and Hopkins (1961) transformed the discrete variables into continuous type.

- Another approach is to establish different discriminant rules for different types of variables and combine them to obtain a final classifier, see Xu et al. (1992) for example.

- Krzanowski (1975, 1980) first proposed linear discriminant rules based on the location model (2.1).

Later, more comprehensive discriminant rules were proposed based on logistic discrimination, kernel estimation and the location model; see for example Aitchison and Aitken (1976), Knoke (1982), Asparoukhov and Krzanowski (2000), Kokonendji and Ibrahim (2016) and the references therein. Variable selection for location model based discriminant rules and further extension to quadratic rules and a nonparametric smoothing version can be found in Daudin (1986), Krzanowski (1993), Krzanowski (1994), Krzanowski (1995), Asparoukhov and Krzanowski (2000) and Mahat et al. (2007). Among these approaches, location model based on discriminant rules have received most attentions and were shown to be comparable or better than other procedures under suitable Conditions (Asparoukhov and Krzanowski, 2000; De Leon et al., 2011). However, these approaches are not applicable to the case that the dimension of the discrete variables and the dimension of the continuous variables are both large. In particular, in terms of theoretical analysis, these discriminant rules are either algorithmic without theoretical justification, or statistically justified under the fixed dimensional assumption. For better discussion of existing literatures, we introduce some notations related to the location model and the optimal Bayes rule first.

We focus on binary problems where observations are from two classes. Let $(\mathbf{Z}, \mathbf{U})$ be a random observation and denote its class label as L, where $L \in \{1, 2\}$. Here $\mathbf{Z}$ is a p dimensional continuous variable and $\mathbf{U}$ is a d dimensional discrete variable. Both p and d are very large. We shall assume that all variables in $\mathbf{U}$ are binary since for variables that have more than two categories, we can further reduce it to the binary case by introducing a set of dummy binary variables (Krzanowski, 1993). Denote the probability function of $(\mathbf{Z}, \mathbf{U})$ in class i as $f_i(\mathbf{Z}, \mathbf{U})$ for $i = 1, 2$. It is well known that Bayes' rule is optimal in that it achieves the smallest misclassification rate among all discriminant rules; see for example Anderson. The Bayes rule classifies a data point to the first class if and only if

$$\pi_1 f_1(\mathbf{Z}, \mathbf{U}) > \pi_2 f_2(\mathbf{Z}, \mathbf{U}),$$

where $P(L = i) = \pi_i, i = 1, 2$ is the prior probability of an observation coming from class i. As an application of the Bayes rule, consider the case where there are no discrete variables. Classical linear discriminant analysis (LDA) assume that observations from class i follow $N(\mu_i, \Sigma)$, a multivariate Gaussian distribution with class-specific mean $\mu_i \in \mathbb{R}^p$, and a common covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, an application of the Bayes rule gives rise to the familiar linear discriminant analysis (LDA) rule which assigns observation to class 1 if and only if

$$(\mu_1 - \mu_2)^T \Sigma^{-1} \left[\mathbf{Z} - \frac{\mu_1 + \mu_2}{2}\right] + \log \frac{\pi_1}{\pi_2} > 0. \tag{1.1}$$

The Gaussianity assumption of the observations with a common covariance matrix is particularly appealing since the resulting Bayes rule is a simple linear function of the variable with an index $\Sigma^{-1}(\mu_1 - \mu_2)$ and an intercept $\log\{\pi_1/\pi_2\}$. Indeed, a widely studied and popular approach in high-dimensional classification, as discussed in previous work below, is to assume a sparse index that gives rise to the so-called sparse LDA (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012; Mai and Zou, 2013).

Under the high dimensional setting with discrete variables absent, there is a large growing literature devoted to the study of classification. Bickel and Levina (2004) first showed that using estimates developed under the fixed-dimensionality scenario for high-dimensional problems gives a classifier equivalent to random guessing in the worst case scenario. For the LDA in (1.1), there are many approaches proposed to deal with the high-dimensionality. In recent years, a plethora of methods built on suitable sparsity Conditions have been proposed to estimate (1.1). Under suitable sparsity conditions on μ_1, μ_2 and Σ, Shao et al. (2011) proposed to shrink the entries of their empirical estimates. By assuming that $\Sigma^{-1}(\mu_1 - \mu_2)$ is sparse, several papers proposed to estimate this quantity by minimizing a penalized loss function that constrains its ℓ_1 norm (Cai and Liu, 2011; Fan et al., 2012; Mai et al., 2012; Mai and Zou, 2013). Mai et al. (2019) further studied a multiclass extension of the LDA. Jiang et al. (2018) proposed a sparse quadratic discriminant analysis method that allows the within-class covariance matrices to differ. Jiang et al. (2020) investigated a scenario where the covariance matrices are varying with a fixed-dimensional continuous variable. Despite these new developments in high dimensional LDA for model (1.1), it is challenging to develop proper estimation procedures for (2.1), when the dimensions of the continuous and discrete variables, namely p and d, are both large. On one hand, for a given d, there will be 2^d locations and hence there will be a lot of empty cells unless the sample size grows exponentially in d (Hall, 1981). On the other hand, the means and covariance matrix in (2.1) are now functions of the location, i.e., the high dimensional discrete variable. Of all the papers reviewed, (Jiang et al., 2020) is the closest to this thesis. However, unlike the dynamic LDA where the index variable is defined on a compact, continuous and finite dimensional space (Jiang et al., 2020), the space $\{0, 1\}^d$ is much irregular and as d tends to infinity, one would encounter a small ball probability issue analogous to that occurs in infinite dimensional spaces (Hong and Linton, 2016), bringing an extra layer of complexity in establishing theoretical properties.

To tackle these challenges, we propose a new semiparametric model based on the location model in which the classification rule relies on a functional classification direction $\beta(\mathbf{u})$ and a parametric intercept $\eta(\mathbf{u})$. We show that the classification direction $\beta(\mathbf{u})$ and the intercept $\eta(\mathbf{u})$ can be independently estimated, and overcome the curse of dimensionality of estimating these two high-dimensional functions by penalized likelihood. Crucially for estimating $\beta(\mathbf{u})$, we are the first to leverage traditional kernel smoothing with modern development on small ball probability under high dimensionality which can be of independent interest. More specifically, although asymptotic properties of kernel smoothing estimators for regression functions of infinite order have been rigorously derived in Hong and Linton (2016), the case considered in this paper is very different and more challenging in that high dimensionality appears in both the regression function and the binary covariates. In particular, we have identified that unlike classical kernel smoothing where the bandwidth is usually chosen to balance the bias and variance of a kernel smoothing estimator, the bandwidth here must be chosen to be large enough to ensure an analogue of Bochner's Lemma to hold for high dimensional discrete variables. Built on this, we have further established concentration inequalities for the kernel smoothing estimators, which are essential for obtaining consistency under high dimensionality. To the best of our knowledge, this is the first attempt to establish a theoretical framework to evaluate the concentration behavior of kernel smoothing estimators for high dimensional regression functions with high dimensional independent covariates. Lastly, we have integrated all the estimation errors and derived the asymptotic misclassification rate of the proposed semiparametric classifier, from which one gains useful insights on how the estimation errors of the classification direction and the intercept affect the classification accuracy.

The remainder of this thesis is organized as follows. In Chapter 2, we introduce the separability property of the semiparametric location model, and provide more details on the estimation of its parameters. In Chapter 3, we first provide a motivating proposition which claims that the classification direction $\beta(\mathbf{u})$ and the intercept $\eta(\mathbf{u})$ can be independently estimated, followed by efficient estimation schemes for estimating $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$. In Chapter 4, we provide consistency results for the estimation of $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$, and evaluate the asymptotic misclassification rate of the estimated classifier. Also, technical proofs are deferred to this chapter. In Chapter

5, we conduct extensive numeric studies on simulated data and seven real datasets to illustrate the competitive performance of our method, with comparison to some modern approaches. Some discussion about this work is in Chapter 6.

Chapter 2

A Semiparametric Location Model

The location model treats the discrete random vector $\mathbf{U}$ as a location or cell and assumes that conditional on $\mathbf{U}$, the continuous random vector $\mathbf{Z}$ follows a location-dependent multivariate normal distribution. More specifically, following Olkin and Tate (1961), we assume that the probability an individual selected at random from class i falls in cell $\mathbf{U} = \mathbf{u}$ is $p_i(\mathbf{u})$, and conditional on $\mathbf{U}$, $\mathbf{Z}|\mathbf{U} \sim N(\mu_i(\mathbf{U}), \Sigma(\mathbf{U})), i = 1, 2$. The common discrete variable dependent covariance $\Sigma(\mathbf{U})$ is inspired by the assumption in LDA. Denote the posterior probability of $(\mathbf{Z}, \mathbf{U})$ from class i as $P(i|\mathbf{Z}, \mathbf{U}), i = 1, 2$. Under the location model, the optimal Bayes' rule reduces to a functional linear classifier: classifies $(\mathbf{Z}, \mathbf{U}_{\mathbf{Z}})$ into class 1 if

$$\log \frac{P(1|\mathbf{Z}, \mathbf{U})}{P(2|\mathbf{Z}, \mathbf{U})} = \log \frac{\pi_1 f_1(\mathbf{Z}, \mathbf{U})}{\pi_2 f_2(\mathbf{Z}, \mathbf{U})} \tag{2.1}$$

$$= [\mu_1(\mathbf{U}) - \mu_2(\mathbf{U})]^T \Sigma^{-1}(\mathbf{U}) \left[\mathbf{Z} - \frac{\mu_1(\mathbf{U}) + \mu_2(\mathbf{U})}{2}\right] + \log \frac{\pi_1 p_1(\mathbf{U})}{\pi_2 p_2(\mathbf{U})} > 0,$$

and into class 2 otherwise. Denote $\beta(\mathbf{U}) := \Sigma^{-1}(\mathbf{U})[\mu_1(\mathbf{U}) - \mu_2(\mathbf{U})]$, and $\eta(\mathbf{U}) := \log \frac{\pi_1 p_1(\mathbf{U})}{\pi_2 p_2(\mathbf{U})}$. The classifier in (2.1) can be written as

$$\beta(\mathbf{U})^T \left[\mathbf{Z} - \frac{\mu_1(\mathbf{U}) + \mu_2(\mathbf{U})}{2}\right] + \eta(\mathbf{U}) > 0, \tag{2.2}$$

which is a functional linear classifier in $\mathbf{Z}$, with a location-dependent direction $\beta(\mathbf{U})$ and a location-dependent intercept $\eta(\mathbf{U})$. In our setting, $\beta(\mathbf{u})$ is a p-dimensional vector for each fixed $\mathbf{u}$, and $\mathbf{u}$ itself is also high-dimensional. As a result, the dimensionality of the model is extremely high. In the setting of this thesis, we considered $\mathbf{U}$ as a vector of binary variables. And we have

- For $\beta(\mathbf{U})$, we have 2^d p-dimensional vectors to estimate;
- For $\eta(\mathbf{U})$, we have 2^d different values to estimate.

Thus, the estimation problem even after assuming the location model is much more challenging than that of the LDA in which only a scalar intercept and a p-dimensional vector need estimating. Given a relative small sample size, there is no hope that either $\beta(\mathbf{U})$ or $\eta(\mathbf{U})$ can be reasonably estimated unless some kind of structures are assumed. In this thesis, we focus on a general scenario where $\beta(\mathbf{U})$ is treated as a function which varies smoothly over the locations $\mathbf{U}$, and $\eta(\mathbf{U})$ is modelled by a parametric first order approximation. Some of the properties regarding the location model (e.g., Proposition 3.1) and our theoretical framework could be adapted to cases where more stringent assumptions are imposed to further simplify the model complexity. Because of the dependence of $\beta(\mathbf{U})$ on U, we shall refer to our model as the semiparametric location model.

Denote the sample in population 1 as $(\mathbf{U}_1, \mathbf{X}_1), \ldots, (\mathbf{U}_{n_1}, \mathbf{X}_{n_1})$ and in population 2 as $(\mathbf{V}_1, \mathbf{Y}_1), \ldots, (\mathbf{V}_{n_2}, \mathbf{Y}_{n_2})$. We now summarize the main steps for estimating the parameters based on this sample. First, we observe that the estimation of $\beta(\mathbf{U})$ and that of $\eta(\mathbf{U})$ can be made separate, thanks to a key result provided in Proposition 3.1. Note that from Proposition 3.1 we have:

$$\eta(\mathbf{u}) = \log \frac{P(L = 1|\mathbf{U} = \mathbf{u})}{P(L = 2|\mathbf{U} = \mathbf{u})},$$

i.e., $\eta(\mathbf{u})$ is the logit transformation of the probability $P(L = 1|\mathbf{U} = \mathbf{u})$. Consequently, $\eta(\mathbf{U})$ can be simply obtained by fitting a logistic regression with $\mathbf{U}$ as covariates, $I(L = 1)$ as the response, and $\eta(\mathbf{u})$ as the discriminant function. Here $I(\cdot)$ is the indicator function. Note that for any injective function $G(\mathbf{u}) : \{0,1\}^d \mapsto R$, it can be easily shown by induction that $G(\mathbf{u})$ can be written as a polynomial of $\mathbf{u} = (u_1, \dots, u_d) \in \{0,1\}^d$. Specifically, there exist constants $a_0; a_1, \dots, a_p; a_{1,2}, a_{1,3}, \dots, a_{p-1,p}; \dots; a_{1,2\dots,d}$, such that

$$G(\mathbf{u}) = a_0 + \sum_{i=1}^{d} a_i u_i + \sum_{1\le i<j\le d} a_{i,j} u_i u_j + \cdots + a_{1,2\dots,d} u_1 u_2 \cdots u_d.$$

Suppose $\log p_1(\mathbf{u})$ and $\log p_2(\mathbf{u})$ are injective functions such that

$$\begin{aligned} &\log \pi_j p_j(\mathbf{u}) && (2.3)\\ = \ & a_0^{(j)} + \sum_{i=1}^{d} a_i^{(j)} u_i + \sum_{1\le i<j\le d} a_{i,j} u_i u_j + \cdots + a_{1,2\dots,d} u_1 u_2 \cdots u_d, \quad j = 1, 2, \end{aligned}$$

where the intercepts and first order coefficients between the two classes, $a_0^{(1)}, a_1^{(1)}, \dots, a_p^{(1)}$ and $a_0^{(2)}, a_1^{(2)}, \dots, a_p^{(2)}$, are potentially different, while the higher order coefficients $a_{1,2}, \dots, a_{1,2\dots,d}$ are assumed to be common between the two classes.

In the context we consider in this thesis where d (i.e., the dimension of $\mathbf{U}$) grows to infinity, one popular approach to fit this logistic regression is to focus on the lower order terms in $\eta(\mathbf{U})$. That is, only the main effects or lower-order interaction terms of the variables in $\mathbf{U}$ are considered. In this thesis, we focus on the main effects model by modelling $\eta(\mathbf{U})$ as

$$\eta(\mathbf{u}) = A_0 + \sum_{i=1}^{d} A_i u_i. \tag{2.4}$$

with $A_i = a_i^{(1)} - a_i^{(2)}, i = 0, \ldots, d$ and $\mathbf{A} = (A_1, ..., A_d)$. The assumption of common higher order coefficients between the two classes in (2.3) is analogous to the common covariance assumption in the classical LDA setting. Importantly, since all the u_i's are binary variables, the probability of a higher order term $u_{i_1}u_{i_2}\cdots u_{i_k}$ taking the value 1 tends to zero in a geometric rate. Hence the expansion (2.3) can in general be well approximated by the lower order terms in practice. This is in contrast to the unusual lower order approximation widely applied in linear models where the rational is to achieve certain amount of parsimony. In particular, (2.4) is true if we only consider model (2.3) with the first two terms only. We note that in addition to achieve parsimony of modelling, the approximation using low order terms for binary variables is more meaningful than that in the usual linear model.

The estimation of $\beta(\mathbf{U}) = \Sigma^{-1}(\mathbf{U})[\mu_1(\mathbf{U}) - \mu_2(\mathbf{U})]$ is more challenging. Our strategy is to estimate $\mu_i(\mathbf{U})$ and $\Sigma(\mathbf{U})$ first via kernel smoothing. Since $\mathbf{U}$ is discrete, we employ Hamming distance to measure the closeness of two vectors of discrete values. Theoretically, we handle the diverging dimension of $\mathbf{U}$ by analyzing normalized versions of the kernel weights, and analyze the interplay between the bandwidth and the dimensionality in the kernel smoothing via a novel analysis on the small ball probability to ensure suitable convergence. Different from classical kernel smoothing theory where the bandwidth only affects the estimation bias and variance, the bandwidth here is crucial for the integrated small ball probability that is the normalizing constant in our case. Interestingly, as indicated by Lemma 4.2 , the bandwidth should be large enough to guarantee an analogue of Bochner's Lemma (Bosq, 2012) to hold. As a result, we establish concentration inequalities for the normalized terms in the kernel estimators, and further show that the misclassification rate of our proposed classifier converges to the optimal Bayes risk under appropriate assumptions.

We remark that the semiparametric location model is much more complicated than the well-known semiparametric varying coefficient model in that (i) the varying coefficient $\beta(\mathbf{U})$ is a function of $\mathbf{U}$, which is a high dimensional binary variable, while in the literature, the dynamic variable is usually univariate and continuous. Existing estimation approaches such as spline methods for semiparametric varying coefficient models (Wei et al., 2011) are no longer valid. (ii) the covariate $\mathbf{U}$ is the same as the dynamic factor, introducing possible confounding in the estimation of $\beta(\mathbf{u})$ and $\eta(\mathbf{u}) = A_0 + \mathbf{A}^T\mathbf{u}$. Fortunately, in the next chapter, we show that the direction $\beta(\mathbf{u})$ and the intercept $\eta(\mathbf{u})$ can be independently estimated, and hence direct estimations can be efficiently constructed.

Chapter 3

Estimation

In our semiparametric location classifier in (2.2), $\eta(\mathbf{u})$ is a linear function of $\mathbf{u}$ and $\beta(\mathbf{u})$ is a function of the location $\mathbf{u}$. In this chapter we will first look at the classification problem from the perspective of minimizing the expected misclassification rate, from which it is found that although $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$ both appear in the Bayes rule (2.2), they can be independently estimated. We will then discuss how to estimate $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$.

3.1 Separability of $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$

For a given location $\mathbf{U} = \mathbf{u}$, consider a general discriminant rule

$$D(\mathbf{b}(\mathbf{u}), b_0(\mathbf{u})) := \mathbf{b}(\mathbf{u})^T[\mathbf{Z} - \frac{\mu_1(\mathbf{u}) + \mu_2(\mathbf{u})}{2}] + b_0(\mathbf{u}), \tag{3.1}$$

which classifies a new observation $(\mathbf{Z}, \mathbf{u})$ to class 1 if and only if $D(\mathbf{b}(\mathbf{u}), b_0(\mathbf{u})) > 0$. Under the location model, the misclassification rate of the classifier $D(\mathbf{b}(\mathbf{u}), b_0(\mathbf{u}))$ over all locations can be seen as

$$R(D(\mathbf{b}(\mathbf{u}), b_0(\mathbf{u})) = \sum_{\mathbf{u}\in\{0,1\}^d} [\pi_1 p_1(\mathbf{u}) P_{\mathbf{u}}\,(2|1) + \pi_2 p_2(\mathbf{u}) P_{\mathbf{u}}\,(1|2)], \tag{3.2}$$

where $P_{\mathbf{u}}(i|j)$ is the conditional misclassification probability of classifying $\mathbf{Z}$ from class i to class j given the location $\mathbf{u}$. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. Under the Gaussianity assumption we have

$$P_{\mathbf{u}}(i|j) = \Phi\left(\frac{b_0(\mathbf{u}) - \mathbf{b}(\mathbf{u})^T[\mu_i(\mathbf{u}) - \mu_j(\mathbf{u})]/2}{\sqrt{\mathbf{b}(\mathbf{u})^T\Sigma(\mathbf{u})\mathbf{b}(\mathbf{u})}}\right). \tag{3.3}$$

The purpose of classification is to seek a classification direction $\mathbf{b}(\mathbf{u})$ and the corresponding intercept $b_0(\mathbf{u})$ such that the expected misclassification rate in (3.2) is minimized. Although the estimation of these two arguments are interrelated, we show in the following proposition that their estimation can be separated.

Proposition 3.1. *Assume the location model hold, and let $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$ be the optimal classification direction and intercept in the Bayes classifier* (2.2)*, respectively. Consider $D(b(\mathbf{U}), 0)$, a special case of the general discriminant rule* (3.1) *with a zero intercept: $b_0(\mathbf{u}) = 0$. We have*

$$\begin{aligned}\beta(\mathbf{u}) &= \Sigma^{-1}(\mathbf{u})[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})] \\ &= \underset{\mathbf{b}(\mathbf{u})\in\mathcal{D}}{\arg\min} \quad E\{R(\mathbf{b}(\mathbf{u}), 0)\},\end{aligned}$$

where $\mathcal{D}$ is the set of all functions from $\{0,1\}^d$ to $\mathbb{R}$, and E is the expectation. On the other hand, for the optimal intercept $\eta(\mathbf{u})$ we have: $\eta(\mathbf{u}) = \log\frac{P(L=1|\mathbf{U}=\mathbf{u})}{P(L=2|\mathbf{U}=\mathbf{u})}$.

Proof. Given the classification direction $\mathbf{b}(\mathbf{u})$, to minimize (3.2), it can be shown after some simple calculation that the optimal intercept is given as

$$\tilde{b}_0(\mathbf{u}) = \frac{\mathbf{b}(\mathbf{u})^T\Sigma(\mathbf{u})\mathbf{b}(\mathbf{u})}{[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})]^T\mathbf{b}(\mathbf{u})}\log\left(\frac{\pi_1 p_1(\mathbf{u})}{\pi_2 p_2(\mathbf{u})}\right).$$

Plugging the above equation into (3.2), we obtain that, to minimize the expected misclassification rate, it is equivalent to find $\mathbf{b}(\mathbf{u})$ that minimizes

$$\begin{aligned} & R(D(\mathbf{b}(\mathbf{u}), \tilde{b}_0(\mathbf{u})) \qquad (3.4) \\ = & \sum_{\mathbf{u}\in\{0,1\}^d} \left[\pi_1 p_1(\mathbf{u})\Phi\left(-\frac{\Delta}{2} - \frac{\eta(\mathbf{u})}{\Delta}\right) + \pi_2 p_2(\mathbf{u})\Phi\left(-\frac{\Delta}{2} + \frac{\eta(\mathbf{u})}{\Delta}\right)\right], \end{aligned}$$

where $\eta(\mathbf{u}) = \log\left(\frac{\pi_1 p_1(\mathbf{u})}{\pi_2 p_2(\mathbf{u})}\right)$, and $\Delta = \frac{[\mu_1(\mathbf{u})-\mu_2(\mathbf{u})]^T\mathbf{b}(\mathbf{u})}{\sqrt{\mathbf{b}(\mathbf{u})^T\Sigma(\mathbf{u})\mathbf{b}(\mathbf{u})}}$. For a given $u \in \{0,1\}^d$, denote

$$R_{\mathbf{u}}(\Delta, \eta) = \pi_1 p_1(\mathbf{u})\Phi\left(-\frac{\Delta}{2} - \frac{\eta(\mathbf{u})}{\Delta}\right) + \pi_2 p_2(\mathbf{u})\Phi\left(-\frac{\Delta}{2} + \frac{\eta(\mathbf{u})}{\Delta}\right).$$

By taking the partial derivation with the respect to Δ, it can be shown that

$$\frac{\partial R_{\mathbf{u}}(\Delta, c)}{\partial \Delta} = -\sqrt{\pi_1 p_1(\mathbf{u})\pi_2(\mathbf{u})p_2(\mathbf{u})}\exp\left\{-\left(\frac{c^2(\mathbf{u})}{\Delta^2} + \frac{1}{4}\Delta^2\right)/2\right\} < 0,$$

which implies that to minimize the expected misclassification rate, it is equivalent to look for $\mathbf{b}(\mathbf{u})$ that maximizes $\Delta(\mathbf{u})$. Now let's consider the set of zero-intercept linear classifiers defined as in (3.1) with $b_0(\mathbf{u}) = 0$. Consequently (3.3) can be written as $R(1|2) = R(2|1) = \Phi\left(-\frac{\Delta(\mathbf{u})}{2}\right)$, and the expected misclassification rate (3.2) reduces to:

$$R(\mathbf{b}(\mathbf{u})) = \sum_{\mathbf{u}\in\{0,1\}^d} [\pi_1 p_1(\mathbf{u}) + \pi_2 p_2(\mathbf{u})]\Phi\left(-\frac{\Delta(\mathbf{u})}{2}\right) = \Phi\left(-\frac{\Delta(\mathbf{u})}{2}\right),$$

which is also minimized when $\Delta(\mathbf{u})$ is maximized. Consequently the optimal zero-intercept classification direction is also the minimizer of (3.4).

On the other hand, by the definition of $\eta(\mathbf{u})$, we have:

$$\eta(\mathbf{u}) = \log\frac{\pi_1 p_1(\mathbf{u})}{\pi_2 p_2(\mathbf{u})}$$

$$
\begin{aligned}
&= \log \frac{P(\mathbf{U} = \mathbf{u}, L = 1)}{P(\mathbf{U} = \mathbf{u}, L = 2)} \\
&= \log \frac{P(L = 1 | \mathbf{U} = \mathbf{u})}{P(L = 2 | \mathbf{U} = \mathbf{u})}
\end{aligned}
$$

This proves the claim on $\eta(\mathbf{u})$. □

This proposition ensures that the estimation $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$ can be conducted separately. In particular, it indicates that the estimation of $\beta(\mathbf{u})$ can be conducted by simply setting $\eta(\mathbf{u}) = 0$.

3.2 Estimation of $\beta(\mathbf{u})$

Denote the estimate of $\mu_1(\mathbf{u}), \mu_2(\mathbf{u})$ and $\Sigma(\mathbf{u})$ as $\widehat{\mu}_1(\mathbf{u})$, $\widehat{\mu}_2(\mathbf{u})$ and $\widehat{\Sigma}(\mathbf{u})$ respectively. To estimate $\beta(\mathbf{u})$ at any cell $\mathbf{u}$, we will resort to kernel smoothing. Recall that our sample $n = n_1 + n_2$ consists of $(\mathbf{U}_1, \mathbf{X}_1), \ldots, (\mathbf{U}_{n_1}, \mathbf{X}_{n_1})$ from population 1 and sample consists of $(\mathbf{V}_1, \mathbf{Y}_1), \ldots, (\mathbf{V}_{n_2}, \mathbf{Y}_{n_2})$ from population 2. For any $\mathbf{u}_1, \mathbf{u}_2 \in \{0, 1\}^d$, define the normalized Hamming distance between $\mathbf{u}_1$ and $\mathbf{u}_2$ as

$$
< \mathbf{u}_1, \mathbf{u}_2 >= \frac{|\mathbf{u}_1 - \mathbf{u}_2|_1}{d},
$$

where $|\cdot|_1$ is the ℓ_1 norm. Our estimation of the mean and covariance matrix is based on the following Nadaraya-Watson type local smoothing. For given bandwidths h_x and h_y, we estimate μ_1 and μ_2 as

$$
\widehat{\mu}_1(\mathbf{u}) = \sum_{j=1}^{n_1} \frac{\exp\{-(dh_x)^{-1} |\mathbf{U}_j - \mathbf{u}|_1\} \mathbf{X}_j}{\sum_{j=1}^{n_1} \exp\{-(dh_x)^{-1} |\mathbf{U}_j - \mathbf{u}|_1\}},
$$

$$
\widehat{\mu}_2(\mathbf{u}) = \sum_{j=1}^{n_2} \frac{\exp\{-(dh_y)^{-1} |\mathbf{V}_j - \mathbf{u}|_1\} \mathbf{Y}_j}{\sum_{j=1}^{n_2} \exp\{-(dh_y)^{-1} |\mathbf{V}_j - \mathbf{u}|_1\}}.
$$

Based on these, we shall establish the theory for the kernel smoothing estimators. Note that $\Sigma(\mathbf{u}) = E[\mathbf{X}(\mathbf{u})\mathbf{X}(\mathbf{u})^T] - E\mathbf{X}(\mathbf{u})(E\mathbf{X}(\mathbf{u}))^T = E[\mathbf{Y}(\mathbf{u})\mathbf{Y}(\mathbf{u})^T] - E\mathbf{Y}(\mathbf{u})(E\mathbf{Y}(\mathbf{u}))^T$. We estimate $\Sigma(\mathbf{u})$ as

$$\widehat{\Sigma}(\mathbf{u}) = \frac{n_1}{n}\widehat{\Sigma}_1(\mathbf{u}) + \frac{n_2}{n}\widehat{\Sigma}_2(\mathbf{u}),$$

where

$$\widehat{\Sigma}_1(\mathbf{u}) = \frac{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}\mathbf{X}_j\mathbf{X}_j^T}{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}} - \widehat{\mu}_1(\mathbf{u})\widehat{\mu}_1(\mathbf{u})^T,$$

$$\widehat{\Sigma}_2(\mathbf{u}) = \frac{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}\mathbf{Y}_j\mathbf{Y}_j^T}{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}} - \widehat{\mu}_2(\mathbf{u})\widehat{\mu}_2(\mathbf{u})^T,$$

with the bandwidth parameters h_{xx} and h_{yy} controlling the smoothness of the estimators. By noting that given $\mathbf{u}$, $\beta(\mathbf{u})$ is the minimizer of the convex loss function $R(\beta(\mathbf{u})) := \beta(\mathbf{u})^T\Sigma(\mathbf{u})\beta(\mathbf{u}) - 2\beta(\mathbf{u})^T[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})]$, we estimate $\beta(\mathbf{u})$ by minimizing the following penalized loss function

$$\widehat{\beta}(\mathbf{u}) := \arg\min_{b\in\mathbb{R}^p} \quad b^T\widehat{\Sigma}(\mathbf{u})b - 2b^T(\widehat{\mu}_1(\mathbf{u}) - \widehat{\mu}_2(\mathbf{u})) + \lambda_\beta\left|\beta(\mathbf{u})\right|_1, \tag{3.5}$$

at each $\mathbf{u}$, where $\lambda_\beta > 0$ is a tuning parameter, and $\widehat{\mu}_1, \widehat{\mu}_2$ and $\widehat{\Sigma}$ are kernel estimators of μ_1, μ_2 and Σ discussed above and defined in Chapter 3 Section 3.2. Because $\beta(\mathbf{U})$ and $\eta(\mathbf{U})$ can be estimated separately, λ_β in (3.6) can be independently chosen without referencing to the estimation of $\eta(\mathbf{u})$. In particular, Proposition 3.1 implies that the choice of λ_β can be determined by minimizing the misclassification rate when the intercept $\eta(\mathbf{U})$ is set to be zero.

3.3 Estimation of $\eta(\mathbf{u})$

Note that $\eta(\mathbf{u}) = \log \frac{\pi_1 p_1(\mathbf{u})}{\pi_2 p_2(\mathbf{u})} = \log \frac{P(\mathbf{U}=\mathbf{u}, L=1)}{P(\mathbf{U}=\mathbf{u}, L=2)} = \log \frac{P(L=1|\mathbf{U}=\mathbf{u})}{P(L=2|\mathbf{U}=\mathbf{u})}$. Under the semiparametric location model we have,

$$\log \frac{P(L=1|\mathbf{U}=\mathbf{u})}{P(L=2|\mathbf{U}=\mathbf{u})} = A_0 + \mathbf{A}^T \mathbf{u},$$

which corresponds to the well known logistic regression model. Therefore, given the samples $\mathbf{U}_1, \ldots, \mathbf{U}_{n_1}$; and $\mathbf{V}_1, \ldots, \mathbf{V}_{n_2}$, we propose to estimate $(A_0, \mathbf{A})$ by minimizing the following penalized entropy loss:

$$(\widehat{A}_0, \widehat{\mathbf{A}}) = \underset{A_0 \in \mathbb{R}, \mathbf{A} \in \mathbb{R}^d}{\arg\min} \tfrac{1}{n} \Big\{ \textstyle\sum_{i=1}^{n_1} \left[-(A_0 + \mathbf{A}^T \mathbf{U}_i) + \log(1 + \exp\{A_0 + \mathbf{A}^T \mathbf{U}_i\}) \right] \quad (3.6)$$

$$+ \textstyle\sum_{j=1}^{n_2} \log(1 + \exp\{A_0 + \mathbf{A}^T \mathbf{V}_i\}) \Big\} + \lambda_\eta |\mathbf{A}|_1,$$

where $\mathbb{R} = (-\infty, \infty)$, $|\mathbf{A}|_1$ denotes the ℓ_1 norm of $\mathbf{A}$, and $\lambda_\eta > 0$ is a tuning parameter.

Chapter 4

Theoretical Results

Notations here: For a $k \times p$ matrix $\mathbf{M} = (M_{ij})_{k\times p}$ we denote the vector l_∞ norm induced matrix norm as $\|M\|_L := \max_{1\le i\le k}\sum_{j=1}^p |M_{ij}|$, and denote $\|\mathbf{M}\|_\infty = \max_{1\le i\le k, 1\le j\le p}|M_{ij}|$. Write the jth component of $\mathbf{X}_i$ as X_{ij} and denote Y_{ij} likewise. With some abuse of notations, for a given $\mathbf{u} \in \{0,1\}^d$, we denote $\Delta_{\mathbf{u}} = \mathbf{U} - \mathbf{u}$ and $N_{\mathbf{u}} = |\mathbf{U} - \mathbf{u}|_1$, where $\mathbf{U}$ is a random variable with probability mass function $p_1(\mathbf{U})$ or $p_2(\mathbf{U})$, depending on whether $\mathbf{U}$ is from class 1 or 2. We use $m(\mathbf{u})$ as a generic notation to denote any of the following conditional mean functions: $E[X_{1i}^k|\mathbf{u}]$, $E[Y_{1i}^k|\mathbf{u}]$, $E[(X_{1i}X_{1j})^k|\mathbf{u}]$, and $E[(Y_{1i}Y_{1j})^k|\mathbf{u}]$ with $k = 1$ or 2, $1 \le i, j \le p$. We denote $a \succ b$ if $a/b \to \infty$ and $a \asymp b$ if $a = O(b)$ and $b = O(a)$. Throughout this thesis, $c, C, C_0, C_1, C_2, \dots$ refer to some generic constants that may take different values in different places.

4.1 Concentration inequalities

To study the property of the proposed estimators, we make the following assumptions.

(C1). There exists a constant $B \in (0, 0.5]$ such that $B \le d^{-1}EN_{\mathbf{u}} \le 1 - B$ holds for any $\mathbf{u} \in \{0,1\}^d$. There exist constants $C > 0$ and $0 \le \alpha < \frac{1}{2}$ such that for any $\mathbf{u} \in \{0,1\}^d$, $E(\sum_{i=1}^d W_{\mathbf{u}}^{(i)})^k \leq Ck!d^{1+\alpha(k-2)}$ for any $k \geq 2$, where $\mathbf{W}_{\mathbf{u}} = (W_{\mathbf{u}}^{(1)}, \ldots, W_{\mathbf{u}}^{(d)})^T = \Delta_{\mathbf{u}} - E\Delta_{\mathbf{u}}$.

(C2). Write $n = n_1 + n_2$. We assume that $n_1 \asymp n_2$, $\frac{\log(p+d)}{n} + \frac{\log(p+d)}{d\log^2 n} \to 0$, and for $h = h_x, h_y, h_{xx}$ or h_{yy}, there exists a positive constant $C > 0$ such that $\frac{\log(d+n)}{h^2 d} \leq C$.

(C3). For any $\mathbf{u} \in \{0,1\}^d$ and $t > 0$, denote

$$\kappa_{\mathbf{u}}(t) := \frac{E[m(\mathbf{U}) - m(\mathbf{u})]\exp\{-(dt)^{-1}N_{\mathbf{u}}\}}{E\exp\{-(dt)^{-1}N_{\mathbf{u}}\}},$$

and let $\kappa(t) = \sup_{\mathbf{u}\in\{0,1\}^d} \kappa_{\mathbf{u}}(t)$. We assume that $\kappa(t) \to 0$ as $t \to 0$ and $d \to \infty$.

(C4). There exists a positive constant M such that $\sup_{1\leq i\leq p, \mathbf{u}\in\{0,1\}^d} EX_{1i}^4(\mathbf{u}) \leq M < \infty$ and $\sup_{1\leq i\leq p, \mathbf{u}\in\{0,1\}^d} EY_{1i}^4(\mathbf{u}) \leq M < \infty$. There exists a constant $M_\Sigma > 1$ such that

$$M_\Sigma^{-1} \leq \inf_{\mathbf{u}\in\{0,1\}^d} \lambda_1(\Sigma(\mathbf{u})) \leq \sup_{\mathbf{u}\in\{0,1\}^d} \lambda_p(\Sigma(\mathbf{u})) \leq M_\Sigma,$$

where $\lambda_1(\Sigma(\mathbf{u}))$ and $\lambda_p(\Sigma(\mathbf{u}))$ are the smallest and largest eigenvalues of $\Sigma(\mathbf{u})$, respectively.

Condition (C1) is a regularization condition on the discrete variable $\mathbf{U}$, and is generally true when the success probability of each element in $\mathbf{U}$ is bounded away from zero and one.

Condition (C2) specifies the order of the bandwidth h. Unlike classical results in kernel smoothing where h is chosen to balance the bias and variance, our h here has to be large enough to ensure the small ball probability $E\exp\{-(hd)^{-1}|\mathbf{U}-\mathbf{u}|_1\}$ is large enough. Specifically, as a form of Bochner's Lemma, a commonly applied result in classical kernel smoothing estimation is that $\frac{1}{h}EK\left(\frac{|\mathbf{U}-\mathbf{u}|_1}{dh}\right)$ would tend to the density function of $\mathbf{u}$ for a properly chosen kernel function under some regular conditions (Bosq, 2012). However, such a conclusion fails in our case. More specifically, suppose we use a continuous density to approximate the discrete probability mass function of $\mathbf{u}$ as $d \to \infty$. With some abuse of notations, let $p_d(\mathbf{u})$ be the point mass probability of $\mathbf{u}$. The approximated density at location $\mathbf{u}$, following traditional arguments, is given as $f(\mathbf{u}) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon} \sum_{\mathbf{u}_\epsilon : d^{-1}|\mathbf{u}_\epsilon - \mathbf{u}|_1 \le \epsilon} p_d(\mathbf{u}_\epsilon)$. Similar to the small ball probability issue in the Hong and Linton (2016), such a density relies on the choice of ϵ and hence is not well defined. On the one hand, as d grows, the point mass function at $\mathbf{u}$ converges to zero in an exponential rate. Such a point mass probability can't be well estimated unless the sample size also grows exponentially in d. To tackle this issue, other than evaluating the denominator and numerator in the Nadaraya-Watson type estimators directly, we establish concentration inequalities for their normalized versions such as $\frac{1}{n_1} \sum_{j=1}^{n_1} \frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}}{E \exp\{-(dh_x)^{-1}|\mathbf{U}_1 - \mathbf{u}|_1\}}$ and $\frac{1}{n_1} \sum_{j=1}^{n_1} \frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j - \mathbf{u}|_1\} X_{ji}}{E \exp\{-(dh_x)^{-1}|\mathbf{U}_1 - \mathbf{u}|_1\}}$. Similar to equation (6) of Hong and Linton (2016), the normalizing coefficient $E \exp\{-(hd)^{-1}|\mathbf{U} - \mathbf{u}|_1\}$ can be viewed as an integrated small ball probability. Condition (C3) quantifies the smoothness of $m(\mathbf{u})$. For better understanding, we provide some examples where (C3) is satisfied.

Example 1. Smoothness on the expected different function on the contour Note that for any integer s, $\mathcal{C}_{\mathbf{u},s} := \{\mathbf{U} : N_{\mathbf{u}} = s\}$ defines a contour with radius s from the center $\mathbf{u}$. Let $G_{\mathbf{u}}(s) := E[m(\mathbf{U}) - m(\mathbf{u})|N_{\mathbf{u}} = s]$ be the expected difference of $m(\cdot)$ over all the $\mathbf{U}$'s on the contour $\mathcal{C}_{\mathbf{u},s}$. (C3) is satisfied if $G_{\mathbf{u}}(s)$ is smooth in the sense that $\frac{E[G_{\mathbf{u}}(s) \exp\{-(dt)^{-1} N_{\mathbf{u}}\}]}{E \exp\{-(dt)^{-1} N_{\mathbf{u}}\}} = \kappa_{\mathbf{u}}(t) \to 0$.

As an example, suppose for any $\mathbf{u}_i \in \mathcal{C}_{\mathbf{u},1}$, we have $m(\mathbf{u}_i) = m(\mathbf{u}) + s_{\mathbf{u},\mathbf{u}_i}$, where $s_{\mathbf{u},\mathbf{u}_i}$ is the a "signal" generated from $s_{\mathbf{u},\mathbf{u}_i} = I\{\varepsilon_{\mathbf{u}} = 0\} N(0, d^{-1}\sigma^2)$, for some constant $\sigma^2 > 0$ and some Bernoulli random innovation $\varepsilon_{\mathbf{u}}$ such that $P(\varepsilon_{\mathbf{u}} = 1) = 1 - \varepsilon_{\mathbf{u}} = 0 = \pi_{\mathbf{u}}$. The point mass probability $\pi_{\mathbf{u}}$ controls the proportion of neighbors that take the same value as $m(\mathbf{u})$. Condition (C3) is satisfied if $\pi_{\mathbf{u}} \to 1$ fast enough such that $G_{\mathbf{u}}(s) = o(1)$ for all $s = 1, \ldots, d$.

Example 2. Lipschitz with exponential order

For any $\mathbf{u}_1, \mathbf{u} \in \{0,1\}^d$, suppose there exists a constant $0 \leq c \leq 1$ such that,

$$|m(\mathbf{u}_1) - m(\mathbf{u})| \leq \kappa_1 \exp\left\{ \frac{|\mathbf{u}_1 - \mathbf{u}|_1^c - EN_{\mathbf{u}}^c}{(d \log d)^{\frac{c}{2}}} \right\},$$

for some $\kappa_1 \to 0$. Note that by Jensen inequality we have for any $c \in [0,1]$, $EN_{\mathbf{u}}^c \leq (EN_{\mathbf{u}})^c$. By Kimball's inequality, Lemma 4.1, Conditions (C1) and (C4), there exists a large enough constant $C_1 > 0$ such that

$$\begin{aligned}
&\frac{E[m(\mathbf{U}) - m(\mathbf{u})]\exp\{-(dt)^{-1}N_{\mathbf{u}}\}}{E\exp\{-(dt)^{-1}N_{\mathbf{u}}\}} \\
\leq\ & \kappa_1 E \exp\left\{ \frac{N_{\mathbf{u}}^c - EN_{\mathbf{u}}^c}{(d \log d)^{\frac{c}{2}}} \right\} \\
=\ & \kappa_1 E \exp\left\{ \frac{EN_{\mathbf{u}}^c}{(d \log d)^{\frac{c}{2}}} \cdot \left(\frac{N_{\mathbf{u}}^c}{EN_{\mathbf{u}}^r} - 1 \right) \right\} \\
=\ & O\left(\kappa_1 \exp\left\{ C_1 \left(\frac{d}{\log d} \right)^{\frac{c}{2}} \cdot \left(\frac{\log d}{d} \right)^{\frac{c}{2}} \right\} \right) \\
=\ & O(\kappa_1).
\end{aligned}$$

Example 3. Centered Lipschitz with exponential order

Existing literature for estimating the distribution of high dimensional discrete variables sometimes models the probability mass as a function of the centered variable instead (Grund and Hall, 1993). We hence consider the following Lipschitz condition centered at the mean of $N_{\mathbf{u}}$: For any $\mathbf{u}_1, \mathbf{u} \in \{0,1\}^d$, there exist constants $c \geq 0$,

$C > 0$ such that,

$$|m(\mathbf{u}_1) - m(\mathbf{u})| \leq C\kappa_1 \exp\left\{ \frac{(|\mathbf{u}_1 - \mathbf{u}|_1 - EN_{\mathbf{u}})^c}{(d\log d)^{\frac{c}{2}}} \right\},$$

for some $\kappa_1 \to 0$. By Kimball's inequality and Lemma 4.1 we have,

$$\begin{aligned} & \frac{E[m(\mathbf{U}) - m(\mathbf{u})]\exp\{-(dt)^{-1}N_{\mathbf{u}}\}}{E\exp\{-(dt)^{-1}N_{\mathbf{u}}\}} \\ = \; & \frac{E[m(\mathbf{U}) - m(\mathbf{u})]\exp\{-(dt)^{-1}(N_{\mathbf{u}} - EN_{\mathbf{u}})\}}{E\exp\{-(dt)^{-1}(N_{\mathbf{u}} - EN_{\mathbf{u}})\}} \\ \leq \; & C\kappa_1 E\exp\left\{ \frac{(N_{\mathbf{u}} - EN_{\mathbf{u}})^r}{(d\log d)^{\frac{r}{2}}} \right\} \\ = \; & O(\kappa_1). \end{aligned}$$

Next we establish concentration inequalities for the weighted estimators of the mean and covariance matrix functions. Note that in practice, the sample size in either the training dataset or the testing dataset can be much smaller than the total locations 2^d. For simplicity, we shall assume that the region of interest is the ball centered at $\mathbf{v} \in \{0,1\}^d$ with radius r: $B_{\mathbf{v}}(r) := \{\mathbf{u} \in \{0,1\}^d : |\mathbf{u} - \mathbf{v}|_1 \leq r\}$. Let $\kappa(\cdot)$ be defined as in Condition (C3).

Before going to the main theorems and proofs of them, we firstly introduce some technical lemmas and establish concentration inequalities for the weighted estimators of the mean and matrix functions.

Lemma 4.1. *Let $\mathbf{N} = (N_1, \ldots, N_d)^T \in \{0,1\}^d$ be a random vector with mean $E\mathbf{N} = (p_1, \ldots, p_d)^T$. Denote $W_i = N_i - p_i$, and assume that there exist constants $C > 0$ and $0 \leq \alpha < \frac{1}{2}$ such that $E(\sum_{i=1}^d W_i)^k \leq Ck!d^{1+\alpha(k-2)}$ for any $k \geq 2$. For any $\epsilon > 0$*

such that $d^\alpha \epsilon \to 0$, we have, when d is large enough,

$$P\left(\frac{1}{d}\left|\sum_{i=1}^{d} W_i\right| \geq \epsilon\right) \leq 2\exp\{-C_1 d\epsilon^2\}.$$

Proof. Note that for any $t \leq cd^{1-\alpha}$ for some constant $c < 1$, we have,

$$\begin{aligned}
E\exp\left\{\frac{t}{d}\sum_{i=1}^{d} W_i\right\} &= 1 + \frac{t^2}{2!d^2}E\Big(\sum_{i=1}^{d} W_i\Big)^2 + \frac{t^3}{3!d^3}E\Big(\sum_{i=1}^{d} W_i\Big)^3 + \cdots \\
&\leq 1 + \frac{Ct^2}{d} + \frac{Ct^3}{d^{2-\alpha}} + \cdots \\
&\leq 1 + \frac{Ct^2}{d}\left[1 + \frac{t}{d^{1-\alpha}} + \cdots\right] \\
&\leq 1 + \frac{Ct^2}{d(1-c)} \\
&\leq \exp\left\{\frac{Ct^2}{d(1-c)}\right\}.
\end{aligned}$$

Consequently, by the general form of the Chebyshev-Markov inequalities (c.f. 6.1.a of Lin and Bai (2011)), we have, for any $\epsilon > 0$ such that $d^\alpha \epsilon \to 0$,

$$P\left(\frac{1}{d}\sum_{i=1}^{d} W_i \geq \epsilon\right) \leq E\exp\left\{\frac{t}{d}\sum_{i=1}^{d} W_i - t\epsilon\right\} \leq \exp\left\{\frac{Ct^2}{d(1-c)} - t\epsilon\right\}.$$

Notice that the last term in the above inequality is minimized at $t_0 = (2C)^{-1}d(1-c)\epsilon$. On the other hand, when $t = t_0$ we have $t = o(d^{1-\alpha})$. Consequently, when d is large enough such that $t \leq cd^{1-\alpha}$, we have

$$P\left(\frac{1}{d}\sum_{i=1}^{d} W_i \geq \epsilon\right) \leq \exp\left\{-\frac{d(1-c)\epsilon^2}{4C}\right\}.$$

The lemma is proved by setting $C_1 = \frac{1-c}{4C}$. □

We say $W_1, \ldots, W_d$ are m-dependent if for any $1 \le i \le d$, W_i is dependent to at most m variables in $\{W_j : j \neq i, 1 \le j \le d\}$. When $W_1, \ldots, W_d$ are m-dependent for a constant m, we have $E(\sum_{i=1}^d W_i)^k \le d(m+1)^{k-1}$. Hence Lemma 4.1 holds for m-dependent sequences with $m = o(d^\alpha)$.

Lemma 4.2. *Under Conditions (C1) and (C2) we have, when d is large enough, for any constant $C_2 > B$ and $\mathbf{u} \in \{0,1\}^d$, there exists a constant $C_1 > 0$ such that,*

$$\begin{aligned} E\exp\{-(hd)^{-1}N_{\mathbf{u}}\} &\ge \exp\{-(dh)^{-1}EN_{\mathbf{u}}\}; && (4.1) \\ E\exp\{-(hd)^{-1}N_{\mathbf{u}}\} &< 2(n+d)^{-C_2} + \exp\left\{-(dh)^{-1}EN_{\mathbf{u}} + h^{-1}\sqrt{\frac{C_1\log(d+n)}{d}}\right\} \\ &= \exp\{-(dh)^{-1}EN_{\mathbf{u}}\}\left(1 + O\left(\sqrt{\frac{\log(d+n)}{h^2 d}}\right)\right). && (4.2) \end{aligned}$$

where $B \in [0.5, 1)$ is defined as in Condition 1.

Proof. (4.1) is a direct result of Jensen's inequality.
Recall that $N_{\mathbf{u}} = |\mathbf{U} - \mathbf{u}|_1$, $B \le d^{-1}EN_{\mathbf{u}} \le 1 - B$,

$$\begin{aligned} E\exp\{-(hd)^{-1}N_{\mathbf{u}}\} &= E\exp\{-(hd)^{-1}|\mathbf{U}-\mathbf{u}|_1\} \\ &< 2\exp\{-C_1 d\epsilon^2\} + \exp\{-(dh)^{-1}EN_{\mathbf{u}} + h^{-1}\epsilon\} \\ &= \exp\{-(dh)^{-1}EN_{\mathbf{u}}\}\left(1 + O\left(\sqrt{\frac{\log(d+n)}{h^2 d}}\right)\right) \end{aligned}$$

then (4.2) can be obtained from Lemma 4.1 with $\epsilon = \sqrt{\frac{C_1\log(d+n)}{d}}$. □

Similar to Lemma 1 and Corollary 1 in Hong and Linton (2016), Lemma 4.2 above provides an analogue of Bochner's Lemma for infinite-dimensional discrete variables. However, the convergence will only be obtained when $\frac{\log(d+n)}{h^2 d} \to 0$, which in return indicates that h should not converge to zero with a rate faster than

$O\Big(\sqrt{\frac{\log(d+n)}{d}}\Big)$.

Lemma 4.3. *Under Conditions (C1)-(C3), we have, when n_1 and n_2 are large enough, for any $\mathbf{v} \in \{0,1\}^d$ and a bounded radius r, and any small enough $\epsilon_n > 0$, there exist constants $C_1 > 0$ and $C_2 > 0$ such that*

$$P\left(\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)}\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-1\right|\geq\epsilon_n\right)\leq C_1 d^r\exp\left\{-C_2 n_1\epsilon_n^2\right\}.$$

and

$$P\left(\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)}\left|\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{\exp\{-(dh_y)^{-1}|\mathbf{V}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-1\right|\geq\epsilon_n\right)\leq C_1 d^r\exp\left\{-C_2 n_2\epsilon_n^2\right\}.$$

Proof. Denote $W_j(\mathbf{u}) := n_1^{-1}\big(\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-1\big)$. From Condition 2 and Lemma 4.2, we have, $EW_j(\mathbf{u})^2 \leq n_1^{-2}C$ for some constant $C > 0$. By Doob's submartingale inequality we have, for any $\mathbf{u} \in \{0,1\}^d$, and any $t > 0$ such that $n_1^{-1}t$ is small enough,

$$\begin{aligned}
&P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-1\right|>\epsilon_n\right) \qquad (4.3)\\
\leq\ & 2\exp\{-t\epsilon_n\}\Pi_{j=1}^{n_1}E\exp\{tW_j(\mathbf{u})\}\\
\leq\ & 2\exp\{-t\epsilon_n\}\Pi_{j=1}^{n_1}\{1+t^2EW_j(\mathbf{u})^2\}\\
\leq\ & 2\exp\left\{-t\epsilon_n+\sum_{j=1}^{n_1}t^2EW_j(\mathbf{u})^2\right\}\\
\leq\ & 2\exp\left\{-t\epsilon_n+\frac{t^2C}{n_1}\right\}.
\end{aligned}$$

Here in the second inequality we have used the fact that $EW_j(\mathbf{u}) = 0$ and $e^x \leq$

$1+x+x^2$ when $x>0$ is small enough. By setting $t=(2C)^{-1}n_1\epsilon_n$, we have:

$$P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-1\right|>\epsilon_n\right)\leq 2\exp\left\{-\frac{n_1\epsilon_n^2}{4C}\right\}.$$

By noticing that the cardinality of $B_{\mathbf{v}}(r)$ is less than $d^r/r!$, we have

$$P\left(\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)}\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-1\right|\geq\epsilon_n\right)\leq\frac{2d^r}{r!}\exp\left\{-\frac{n_1\epsilon_n^2}{4C}\right\}.$$

This proves the first statement of the lemma. The second statement can be obtained similarly. □

The following theorems establish uniform consistency for any $\mathbf{u}$ in the ball $B_{\mathbf{v}}(r)$.

Theorem 4.1. *Under Conditions (C1)-(C4), when n_1 and n_2 are large enough, there exist constant $C_1>0$, $C_2>0$ such that for any $\epsilon_n>\kappa(h)$,*

$$P\left(\sup_{1\leq i\leq p,\mathbf{u}\in B_{\mathbf{v}}(r)}|\widehat{\mu}_{1i}(\mathbf{u})-\mu_{1i}(\mathbf{u})|>\epsilon_n\right)\leq C_1pd^r\exp\left\{-C_2n_1(\epsilon_n-\kappa(h_x))^2\right\},$$

and

$$P\left(\sup_{1\leq i\leq p,\mathbf{u}\in B_{\mathbf{v}}(r)}|\widehat{\mu}_{2i}(\mathbf{u})-\mu_{2i}(\mathbf{u})|>\epsilon_n\right)\leq C_1pd^r\exp\left\{-C_2n_2(\epsilon_n-\kappa(h_y))^2\right\}.$$

Proof. Denote

$$W_{ji}(\mathbf{u}):={n_1}^{-1}\left[\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\mu_{1i}(\mathbf{u})\right],$$

and

$$B_i(\mathbf{u}):=\frac{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}X_{1i}(\mathbf{u})}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\mu_{1i}(\mathbf{u}).$$

Similar to (4.3), for any $0 < t < T$, where T is a small enough constant, we have,

$$P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\mu_{1i}(\mathbf{u})\right|>\epsilon_n\right)$$

$$\leq 2\exp\left\{-t\epsilon_n+tB_i(\mathbf{u})+t^2\sum_{j=1}^{n_1}E(W_{ji}(\mathbf{u}))^2\right\}.$$

From Condition (C3) we have $B_i(\mathbf{u}) = \kappa_{\mathbf{u}}(h_x)$. On the other hand, by Lemma 4.2 and Condition (C4), we have,

$$\begin{aligned} E(n_1W_{ji}(\mathbf{u}))^2 &\leq 2E\left|\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\frac{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}\mu_{1i}(\mathbf{U}_1)}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}\right|^2 \\ &\quad +2\left|\frac{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}\mu_{1i}(\mathbf{U}_1)}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\mu_{1i}(\mathbf{u})\right|^2 \\ &\leq C, \end{aligned} \tag{4.4}$$

for some large enough constant $C > 0$. Consequently, we have

$$\begin{aligned} &P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_x)^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}}{E\exp\{-(dh_x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}-\mu_{1i}(\mathbf{u})\right|>\epsilon_n\right) \\ \leq\ & 2\exp\left\{-t\epsilon_n+t\kappa(h_x)+\frac{Ct^2}{n_1}\right\} \\ \leq\ & 2\exp\left\{-\frac{n_1(\epsilon_n-\kappa(h_x))^2}{4C}\right\}. \end{aligned}$$

Together with Lemma 4.3, we have

$$P\left(\sup_{1\leq i\leq p,\mathbf{u}\in B_{\mathbf{v}}(r)}|\widehat{\mu}_{1i}(\mathbf{u})-\mu_{1i}(\mathbf{u})|>\epsilon_n\right)\leq C_1pd^r\exp\left\{-C_2n_1(\epsilon_n-\kappa(h_x))^2\right\}.$$

This proves the first statement of the theorem. Similarly changing

$$W_{ji}(\mathbf{u}) := {n_2}^{-1}\left[\frac{\exp\{-(dh_y)^{-1}|\mathbf{V}_j-\mathbf{u}|_1\}Y_{ji}}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\mu_{2i}(\mathbf{u})\right],$$

and

$$B_i(\mathbf{u}) := \frac{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}Y_{1i}(\mathbf{u})}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\mu_{2i}(\mathbf{u}).$$

$$\begin{aligned}&P\left(\left|\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{\exp\{-(dh_y)^{-1}|\mathbf{V}_j-\mathbf{u}|_1\}Y_{ji}}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\mu_{2i}(\mathbf{u})\right|>\epsilon_n\right)\\&\le 2\exp\left\{-t\epsilon_n+tB_i(\mathbf{u})+t^2\sum_{j=1}^{n_2}E(W_{ji}(\mathbf{u}))^2\right\}.\end{aligned}$$

$$\begin{aligned}E(n_2W_{ji}(\mathbf{u}))^2 &\le 2E\left|\frac{\exp\{-(dh_y)^{-1}|\mathbf{V}_j-\mathbf{u}|_1\}Y_{ji}}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\frac{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}\mu_{2i}(\mathbf{V}_1)}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}\right|^2\\&\quad+2\left|\frac{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}\mu_{2i}(\mathbf{V}_1)}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\mu_{2i}(\mathbf{u})\right|^2\\&\le C,\end{aligned}$$

$$\begin{aligned}&P\left(\left|\frac{1}{n_2}\sum_{j=1}^{n_2}\frac{\exp\{-(dh_y)^{-1}|\mathbf{V}_j-\mathbf{u}|_1\}Y_{ji}}{E\exp\{-(dh_y)^{-1}|\mathbf{V}_1-\mathbf{u}|_1\}}-\mu_{2i}(\mathbf{u})\right|>\epsilon_n\right)\\&\le 2\exp\left\{-t\epsilon_n+t\kappa(h_y)+\frac{Ct^2}{n_2}\right\}\\&\le 2\exp\left\{-\frac{n_2(\epsilon_n-\kappa(h_y))^2}{4C}\right\}.\end{aligned}$$

$$P\left(\sup_{1\le i\le p,\mathbf{u}\in B_{\mathbf{V}}(r)}|\widehat{\mu}_{2i}(\mathbf{u})-\mu_{2i}(\mathbf{u})|>\epsilon_n\right)\le C_1pd^r\exp\left\{-C_2n_2(\epsilon_n-\kappa(h_y))^2\right\}$$

□

Similarly, we can show that:

Theorem 4.2. *Under Conditions (C1)-(C4), when* n_1 *and* n_2 *are large enough, there exist constant* $C_1 > 0$, $C_2 > 0$ *such that for any* $\epsilon_n > \kappa(h)$,

$$P\left(\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \left\|\widehat{\Sigma}(\mathbf{u}) - \Sigma(\mathbf{u})\right\|_\infty > \epsilon_n\right) \le C_1 p^2 d^r \exp\left\{-C_2 n(\epsilon_n - \kappa(h))^2\right\}.$$

Proof. Note that $\left\|\widehat{\Sigma}(\mathbf{u}) - \Sigma(\mathbf{u})\right\|_\infty = \max_{1\le i,j\le p} |\widehat{\sigma}_{i,j}(\mathbf{u}) - \sigma_{i,j}(\mathbf{u})|$, and $\Sigma(\mathbf{u}) = \Sigma_{\mathbf{X}}(\mathbf{u}) = \Sigma_{\mathbf{Y}}(\mathbf{u})$, where

$$\widehat{\Sigma}_1(\mathbf{u}) = \frac{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}\mathbf{X}_j\mathbf{X}_j^T}{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}} - \widehat{\mu}_1(\mathbf{u})\widehat{\mu}_1(\mathbf{u})^T, \tag{4.5}$$

and

$$\widehat{\Sigma}_2(\mathbf{u}) = \frac{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}\mathbf{Y}_j\mathbf{Y}_j^T}{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}} - \widehat{\mu}_2(\mathbf{u})\widehat{\mu}_2(\mathbf{u})^T.$$

From Theorem 4.1 and Condition (C4), and with some abuse of notations, in this proof we shall simply set

$$\widehat{\Sigma}_1(\mathbf{u}) = \frac{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}\mathbf{X}_j\mathbf{X}_j^T}{\sum_{j=1}^{n_1} \exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}} - \mu_1(\mathbf{u})\mu_1(\mathbf{u})^T,$$

and

$$\widehat{\Sigma}_2(\mathbf{u}) = \frac{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}\mathbf{Y}_j\mathbf{Y}_j^T}{\sum_{j=1}^{n_2} \exp\{-(dh_{yy})^{-1}|\mathbf{V}_j - \mathbf{u}|_1\}} - \mu_2(\mathbf{u})\mu_2(\mathbf{u})^T.$$

Denote

$$W_{jik}(\mathbf{u}) := {n_1}^{-1}\left[\frac{\exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}X_{ji}X_{jk}^T}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_j - \mathbf{u}|_1\}} - \mu_{1i}(\mathbf{u})\mu_{1k}(\mathbf{u})^T\right],$$

and

$$B_{ik}(\mathbf{u}) := \frac{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}X_{1i}(\mathbf{u})X_{1k}(\mathbf{u})}{E\exp\{-(dh_x x)^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}} - \mu_{1i}(\mathbf{u})\mu_{1k}(\mathbf{u}).$$

Similar to (4.3), for any $0 < t < T$, where T is a small enough constant, we have,

$$P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_{xx})^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}X_{jk}}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}} - \mu_{1i}(\mathbf{u})\mu_{1j}(\mathbf{u})\right| > \epsilon_n\right)$$
$$\leq 2\exp\left\{-t\epsilon_n + tB_{ik}(\mathbf{u}) + t^2\sum_{j=1}^{n_1}E(W_{jik}(\mathbf{u}))^2\right\}.$$

From Condition (C3) we have $B_{ik}(\mathbf{u}) = \kappa_{\mathbf{u}}(h_{xx})$. On the other hand, by Lemma 4.2 and Condition (C4), we have,

$$\begin{aligned}
& E(n_1W_{jik}(\mathbf{u}))^2 \\
\leq\ & 2E\left|\frac{\exp\{-(dh_{xx})^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}X_{jk}}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}} - \frac{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}\mu_{1i}(\mathbf{U}_1)\mu_{1k}(\mathbf{U}_1)}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}}\right|^2 \\
& +2\left|\frac{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}\mu_{1i}(\mathbf{U}_1)\mu_{1k}(\mathbf{U}_1)}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}} - \mu_{1i}(\mathbf{u})\mu_{1k}(\mathbf{u})\right|^2 \\
\leq\ & C, \qquad (4.6)
\end{aligned}$$

for some large enough constant $C > 0$. Consequently, we have

$$\begin{aligned}
& P\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\frac{\exp\{-(dh_{xx})^{-1}|\mathbf{U}_j-\mathbf{u}|_1\}X_{ji}X_{jk}}{E\exp\{-(dh_{xx})^{-1}|\mathbf{U}_1-\mathbf{u}|_1\}} - \mu_{1i}(\mathbf{u})\mu_{1k}(\mathbf{u})\right| > \epsilon_n\right) \\
\leq\ & 2\exp\left\{-t\epsilon_n + t\kappa(h_{xx}) + \frac{Ct^2}{n_1}\right\} \\
\leq\ & 2\exp\left\{-\frac{n_1(\epsilon_n-\kappa(h_x))^2}{4C}\right\}.
\end{aligned}$$

We can establish similar results for $\widehat{\Sigma}_2(\mathbf{u})$. Since

$$\widehat{\Sigma}(\mathbf{u}) = \frac{n_1}{n}\widehat{\Sigma}_1(\mathbf{u}) + \frac{n_2}{n}\widehat{\Sigma}_2(\mathbf{u}),$$

we immediately have

$$P\left(\sup_{1\le i,k\le p,\mathbf{u}\in B_{\mathbf{v}}(r)} |\widehat{\sigma}_{i,k}(\mathbf{u}) - \sigma_{i,k}(\mathbf{u})| > \epsilon_n\right) \le C_1 p^2 d^r \exp\left\{-C_2 n(\epsilon_n - \kappa(h))^2\right\}.$$

This proves this theorem. □

Theorems 4.1 and 4.2 above provide concentration results for $\widehat{\mu}_1(\mathbf{u})$ and $\widehat{\mu}_2(\mathbf{u})$. In particular, the right hand side of the concentration inequalities will tend to 0 when we set $\epsilon_n = C\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)$ for some large enough constant $C > 0$. The rate $\sqrt{\frac{\log(p+d)}{n}}$ echoes a classical rate that quantifies its dependence on the dimension and the sample size, while the term $\kappa(h)$ is a bias caused by local smoothing. It is generally hard to evaluate $\kappa(h)$ unless some strong structural assumptions are imposed for $m(\mathbf{u})$. Under the classical context, the bandwidth h is usually chosen to obtain a trade-off between the bias and the variance, and hence it is theoretically crucial to know the rate of the bias. However, under the setting that $m(\mathbf{u})$ is high dimensional, the h that provides the best bias-variance trade-off may not necessarily provide any guarantee for the uniform convergence of the estimator, which is an essential requirement for establishing consistency under high dimensionality. Alternatively, we suggest setting h to be large enough (as in Condition (C3)) to ensure an analogue of Bochner's Lemma (i.e., Lemma 4.2) to hold, and as a result, concentration results in the above two theorems can be appropriately established. Practically, although it is common to select the bandwidth by minimizing the mean integrated squared error via cross validation, for classification with high dimensional mixed variables, we have found that it works better to choose h to minimize the misclassification rate.

4.2 Consistency of $\widehat{\beta}(\mathbf{u})$

4.2.1 ℓ_1-penalized estimation

Before we introduce the main results for the estimation of $\beta(\mathbf{u})$, we study the theoretical properties of the solution of the following generic penalized quadratic loss. These results will be used to prove the consistency of $\widehat{\beta}(\mathbf{u})$ later. For a given $\mathbf{u} \in \{0,1\}^d$, let $\widehat{\mathbf{\Omega}}(\mathbf{u})$ and $\widehat{\mathbf{a}}$ be consistent estimators of a $p \times p$ parameter matrix $\mathbf{\Omega}(\mathbf{u})$ and a p dimensional parameter $\mathbf{a}$, respectively. For a given tuning parameter λ, we define the l_1 penalized estimator of $\mathbf{b}^*(\mathbf{u}) := \mathbf{\Omega}(\mathbf{u})^{-1}\mathbf{a}(\mathbf{u})$ as:

$$\widehat{\mathbf{b}}(\mathbf{u}) = \underset{\mathbf{b}\in R^p}{\arg\min} \frac{1}{2}\mathbf{b}^T\widehat{\mathbf{\Omega}}(\mathbf{u})\mathbf{b} - \widehat{\mathbf{a}}^T(\mathbf{u})\mathbf{b} + \lambda\|\mathbf{b}\|_1. \tag{4.7}$$

Denote the support of $\mathbf{b}^*(\mathbf{u})$ as $\mathcal{S}_{\mathbf{u}} = \{i : b_i^*(\mathbf{u}) \neq 0\}$ where $b_i^*(\mathbf{u})$ is the i-th element of $\mathbf{b}^*(\mathbf{u})$. When there is no ambiguity we shall use $\mathcal{S}$ instead of $\mathcal{S}_{\mathbf{u}}$ in some occasions. For example, we shall use $\mathbf{b}_{\mathcal{S}}(\mathbf{u})$ instead of $\mathbf{b}_{\mathcal{S}_{\mathbf{u}}}(\mathbf{u})$ to denote the nonzero subset of $\mathbf{b}(\mathbf{u})$. The following proposition establishes an upper bound for the estimation error of $\widehat{\mathbf{b}}(\mathbf{u})$ in terms of the estimation accuracy of $\widehat{\mathbf{a}}(\mathbf{u})$ and $\widehat{\mathbf{\Omega}}(\mathbf{u})$.

Proposition 4.1. *Denote* $\epsilon_{\mathbf{u}} = \|\widehat{\mathbf{a}}(\mathbf{u}) - \mathbf{a}(\mathbf{u})\|_\infty + \|[\widehat{\mathbf{\Omega}}(\mathbf{u}) - \mathbf{\Omega}(\mathbf{u})]\mathbf{b}^*(\mathbf{u})\|_\infty$, *and* $e_{\mathbf{u}} = \|\mathbf{b}^*(\mathbf{u})\|_0\|\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\mathbf{\Omega}}(\mathbf{u}) - \mathbf{\Omega}(\mathbf{u})\|_\infty$. *Assume the following inequalities hold:*

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \{\|\mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L + e_{\mathbf{u}}\} < 1,$$

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} 2[1 - \|\mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L - 2e_{\mathbf{u}}]^{-1}\epsilon_{\mathbf{u}} < \lambda.$$

We have, for any $\mathbf{u} \in B_{\mathbf{v}}(r)$,

(i) $\widehat{\mathbf{b}}_{\mathcal{S}^c}(\mathbf{u}) = \mathbf{0}$*;*

(ii) $\|\widehat{\mathbf{b}}(\mathbf{u}) - \mathbf{b}^*(\mathbf{u})\|_\infty \le 2(1-e_{\mathbf{u}})^{-1}\|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\lambda$.

Proof. Given the true support $\mathcal{S}$, we consider the estimation

$$\begin{aligned}\widehat{\mathbf{b}}^0(\mathbf{u}) &= \mathop{\arg\min}_{\mathbf{b}\in R^q,\ \mathbf{b}_{\mathcal{S}^c}=0} \frac{1}{2}\mathbf{b}^T\widehat{\boldsymbol{\Omega}}(\mathbf{u})\mathbf{b} - \widehat{\mathbf{a}}(\mathbf{u})^T\mathbf{b} + \lambda\|\mathbf{b}\|_1 \\ &= \mathop{\arg\min}_{\mathbf{b}\in R^q,\ \mathbf{b}_{\mathcal{S}^c}=0} \frac{1}{2}\mathbf{b}_{\mathcal{S}}^T\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u})\mathbf{b}_{\mathcal{S}} - \widehat{\mathbf{a}}_{\mathcal{S}}(\mathbf{u})^T\mathbf{b}_{\mathcal{S}} + \lambda\|\mathbf{b}_{\mathcal{S}}\|_1.\end{aligned}$$

By the Karush-Kuhn-Tucker (KKT) condition, we have

$$\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u})\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \widehat{\mathbf{a}}_{\mathcal{S}}(\mathbf{u}) = -\lambda\mathbf{Z}, \tag{4.8}$$

where $\mathbf{Z}$ is the sub-gradient of $\|\mathbf{b}_{\mathcal{S}}\|_1$. By the definition of $\mathbf{b}^*(\mathbf{u}) := \boldsymbol{\Omega}(\mathbf{u})^{-1}\mathbf{a}(\mathbf{u})$, we have

$$\begin{pmatrix}\mathbf{a}_{\mathcal{S}}(\mathbf{u})\\ \mathbf{a}_{\mathcal{S}^c}(\mathbf{u})\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}) & \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}^c}(\mathbf{u})\\ \boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u}) & \boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}^c}(\mathbf{u})\end{pmatrix}\begin{pmatrix}\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\\ \boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\end{pmatrix},$$

and hence we have $\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u})\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})\mathbf{b}^*_{\mathcal{S}}(\mathbf{u}) + \mathbf{a}_{\mathcal{S}}(\mathbf{u}) - \widehat{\mathbf{a}}_{\mathcal{S}}(\mathbf{u}) = -\lambda\mathbf{Z}$. Consequently, we have,

$$\begin{aligned}&\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \mathbf{b}^*_{\mathcal{S}}(\mathbf{u}) \qquad (4.9)\\ =\ & -\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\left\{\lambda\mathbf{Z} + [\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u}) - \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})]\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) + (\mathbf{a}_{\mathcal{S}}(\mathbf{u}) - \widehat{\mathbf{a}}_{\mathcal{S}}(\mathbf{u}))\right\}.\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}&\|\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\|_\infty\\ \le\ & \|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\Big\{\lambda\|\mathbf{Z}\|_\infty + \|[\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u}) - \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})](\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \mathbf{b}^*_{\mathcal{S}}(\mathbf{u}))\|_\infty\\ & + \|[\widehat{\boldsymbol{\Omega}}_{\mathcal{S},\mathcal{S}}(\mathbf{u}) - \boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})]\mathbf{b}^*_{\mathcal{S}}(\mathbf{u}) + \mathbf{a}_{\mathcal{S}}(\mathbf{u}) - \widehat{\mathbf{a}}_{\mathcal{S}}(\mathbf{u})\|_\infty\Big\}\\ \le\ & \|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\Big\{\lambda + \|\mathbf{b}^*(\mathbf{u})\|_0\|\widehat{\boldsymbol{\Omega}}(\mathbf{u}) - \boldsymbol{\Omega}(\mathbf{u})\|_\infty\|\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u}) - \mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\|_\infty\end{aligned}$$

$$+\|(\widehat{\mathbf{\Omega}}(\mathbf{u})-\mathbf{\Omega}(\mathbf{u}))\mathbf{b}^*(\mathbf{u})+\mathbf{a}(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u})\|_\infty\Big\},$$

which implies that

$$\begin{aligned}
& \|\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\|_\infty \\
\leq\ & [1-\|\mathbf{b}^*(u)\|_0\|\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\mathbf{\Omega}}(\mathbf{u})-\mathbf{\Omega}(\mathbf{u})\|_\infty]^{-1}\|\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L(\lambda+\epsilon_{\mathbf{u}}) \\
\leq\ & 2[1-\|\mathbf{b}^*(u)\|_0\|\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\mathbf{\Omega}}(\mathbf{u})-\mathbf{\Omega}(\mathbf{u})\|_\infty]^{-1}\|\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\lambda, \qquad (4.10)
\end{aligned}$$

where in the last inequality we have used the fact (from the Conditions) that, $\epsilon_{\mathbf{u}} < \lambda$. Next, we complete the proof by showing that $\widehat{\mathbf{b}}^0(\mathbf{u})$ is exactly the minimizer of (4.7). By the KKT condition, it is sufficient to show

$$\|(\widehat{\mathbf{\Omega}}(\mathbf{u})\widehat{\mathbf{b}}^0(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u}))_{\mathcal{S}}\|_\infty \leq \lambda, \text{ and} \qquad (4.11)$$

$$\|(\widehat{\mathbf{\Omega}}(\mathbf{u})\widehat{\mathbf{b}}^0(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u}))_{\mathcal{S}^c}\|_\infty < \lambda. \qquad (4.12)$$

(4.11) is a direct result of (4.8). For (4.12), we have

$$\begin{aligned}
& (\widehat{\mathbf{\Omega}}(\mathbf{u})\widehat{\mathbf{b}}^0(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u}))_{\mathcal{S}^c} \\
=\ & \widehat{\mathbf{\Omega}}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\widehat{\mathbf{a}}_{\mathcal{S}^c}(\mathbf{u}) \\
=\ & \widehat{\mathbf{\Omega}}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})+\mathbf{a}_{\mathcal{S}^c}(\mathbf{u})-\widehat{\mathbf{a}}_{\mathcal{S}^c}(\mathbf{u}) \\
=\ & \widehat{\mathbf{\Omega}}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})(\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u}))+[\widehat{\mathbf{\Omega}}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})-\mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})]\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})+\mathbf{a}_{\mathcal{S}^c}(\mathbf{u})-\widehat{\mathbf{a}}_{\mathcal{S}^c}(\mathbf{u}) \\
=\ & [\widehat{\mathbf{\Omega}}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})-\mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})](\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})) \\
+\ & \mathbf{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\{\mathbf{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})(\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u}))\} \\
& +\{(\widehat{\mathbf{\Omega}}(\mathbf{u})-\mathbf{\Omega}(\mathbf{u}))\mathbf{b}^*(\mathbf{u})+\mathbf{a}(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u})\}_{\mathcal{S}^c}.
\end{aligned}$$

Together with (4.9) and (4.10), we have

$$\|(\widehat{\mathbf{\Omega}}(\mathbf{u})\widehat{\mathbf{b}}^0(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u}))_{\mathcal{S}^c}\|_\infty$$

$$
\begin{aligned}
\leq\ & \|\mathbf{b}^*(\mathbf{u})\|_0\|\widehat{\boldsymbol{\Omega}}(\mathbf{u})-\boldsymbol{\Omega}(\mathbf{u})\|_\infty\|\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\|_\infty \\
& +\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L(\lambda+\epsilon_{\mathbf{u}}+\|\mathbf{b}^*(\mathbf{u})\|_0\|\widehat{\boldsymbol{\Omega}}(\mathbf{u})-\boldsymbol{\Omega}(\mathbf{u})\|_\infty\|\widehat{\mathbf{b}}^0_{\mathcal{S}}(\mathbf{u})-\mathbf{b}^*_{\mathcal{S}}(\mathbf{u})\|_\infty) \\
& +\epsilon_{\mathbf{u}} \\
\leq\ & \frac{(1+\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u}))\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}))^{-1}\|_L)(\lambda+\epsilon_{\mathbf{u}})}{1-\|\mathbf{b}^*(\mathbf{u}))\|_0\|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}))^{-1}\|_L\|\widehat{\boldsymbol{\Omega}}(\mathbf{u}))-\boldsymbol{\Omega}(\mathbf{u}))\|_\infty}-\lambda \\
=\ & \left\{\epsilon_{\mathbf{u}}-\frac{1-\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u}))\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}))^{-1}\|_L-2\|\mathbf{b}^*(\mathbf{u}))\|_0\|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}))^{-1}\|_L\|\widehat{\boldsymbol{\Omega}}(\mathbf{u}))-\boldsymbol{\Omega}(\mathbf{u}))\|_\infty}{1+\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u}))\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u}))^{-1}\|_L}\lambda\right\} \\
& \quad\cdot\left\{\frac{1+\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L}{1-\|\mathbf{b}^*(\mathbf{u})\|_0\|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\boldsymbol{\Omega}}(\mathbf{u})-\boldsymbol{\Omega}(\mathbf{u})\|_\infty}\right\} \\
& +\lambda.
\end{aligned}
$$

Under the Condition that

$$
2[1-\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L-2\|\mathbf{b}^*(\mathbf{u})\|_0\|\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\boldsymbol{\Omega}}(\mathbf{u})-\boldsymbol{\Omega}(\mathbf{u})\|_\infty]^{-1}\epsilon_{\mathbf{u}}<\lambda,
$$

we have $\|(\widehat{\boldsymbol{\Omega}}(\mathbf{u})\widehat{\mathbf{b}}^0(\mathbf{u})-\widehat{\mathbf{a}}(\mathbf{u}))_{\mathcal{S}^c}\|_\infty<\lambda$. Consequently, $\widehat{\mathbf{b}}(\mathbf{u})=\widehat{\mathbf{b}}^0(\mathbf{u})$. Lastly, note that the inequality conditions in this proposition hold for all $\mathbf{u}\in B_{\mathbf{v}}(r)$, we conclude that the model selection consistency and the bound (4.10) hold for all $\mathbf{u}\in B_{\mathbf{v}}(r)$. □

Proposition 4.1 provides a general result for the oracle property of an estimator defined by minimizing the estimated quadratic loss (4.7). We remark that Proposition 4.1 can be applied to many different statistical problems where quadratic loss taking the form (4.7) is adopted. In particular, we do not impose any assumption on the signal strength of the nonzero parameters. Conditions in the above proposition rely on the magnitude of $\|\boldsymbol{\Omega}_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\boldsymbol{\Omega}_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L$, which is related to the well known irrepresentable condition (Zhao and Yu, 2006; Zou, 2006), and the uniform estimation error bounds of $\|\widehat{\mathbf{a}}(\mathbf{u})-\mathbf{a}(\mathbf{u})\|_\infty$ and $\|\widehat{\boldsymbol{\Omega}}(\mathbf{u})-\boldsymbol{\Omega}(\mathbf{u})\|_\infty$, which require specific evaluation for different applications.

4.2.2 Oracle properties of $\widehat{\beta}(\mathbf{u})$

Next we establish the oracle properties of $\widehat{\beta}(\mathbf{u})$ by using the results obtained in the previous part. With some abuse of notations, denote the support of $\beta(\mathbf{u})$ as $\mathcal{S}_{\mathbf{u}} = \{i : \beta_i(\mathbf{u}) \neq 0\}$ where $\beta_i(\mathbf{u})$ is the ith element of $\beta(\mathbf{u})$. Similarly, we use $\widehat{\mathcal{S}}_{\mathbf{u}}$ to denote the support of the estimator $\widehat{\beta}(\mathbf{u})$ based on (3.5). From Proposition 4.1 and the uniform bounds established in Section 4.1, we have:

Theorem 4.3. *Assume that Conditions (C1)-(C4) hold. In addition, assume that* $\left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \|\beta(\mathbf{u})\|_0 \|\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L \to 0$*, and there exists a constant* $0 < \kappa_0 < 1$ *such that* $\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \|\Sigma_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L < 1 - \kappa_0$*. By choosing* $\lambda_\beta = C_2 \left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)}(\|\beta(\mathbf{u})\|_1 + 1)$ *for some large enough constant* C_2*, and denoting* $M_r = \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \|\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L$*, we have, for any given* $r = O(\log_d(p+d))$*,*

(i) $P\left(\bigcap_{\mathbf{u}\in B_{\mathbf{v}}(r)} \{\widehat{\mathcal{S}}_{\mathbf{u}} = \mathcal{S}(\mathbf{u})\}\right) = 1 - O((p+d)^{-1})$*;*

(ii) $P\left(\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \|\widehat{\beta}(\mathbf{u}) - \beta(\mathbf{u})\|_\infty \leq 2\lambda_\beta M_r\right) = 1 - O((p+d)^{-1})$.

Proof. Set $\epsilon_n = C\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)$ for some large enough constant $C > 0$. From Theorems 4.1 and 4.2 we have when C is large enough, with probability larger than $1-O((p+d)^{-1})$, $\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \|\widehat{\mu}_1(\mathbf{u}) - \mu_1(\mathbf{u})\|_\infty = O(\epsilon_n)$ and $\|\widehat{\Sigma}(\mathbf{u}) - \Sigma(\mathbf{u})\|_\infty = O(\epsilon_n)$. Consequently, when C is large enough and $\epsilon_n \to 0$, we have

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \{\|\Sigma_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L + 2\|\beta(\mathbf{u})\|_0 \|\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L \|\widehat{\Sigma}(\mathbf{u}) - \Sigma(\mathbf{u})\|_\infty\} < 1,$$

and

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} 2[1-\|\Sigma_{\mathcal{S}^c,\mathcal{S}}(\mathbf{u})\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L - 2\|\beta(\mathbf{u})\|_0\|\Sigma_{\mathcal{S},\mathcal{S}}(\mathbf{u})^{-1}\|_L\|\widehat{\Sigma}(\mathbf{u})-\Sigma(\mathbf{u})\|_\infty]^{-1}\epsilon_{\mathbf{u}}$$
$$< \lambda_\beta.$$

The theorem then follows from Proposition 4.1. □

Note that we do allow the support of $\beta(\mathbf{u})$ to be different across different $\mathbf{u} \in B_{\mathbf{v}}(r)$. Part (i) in Theorem 4.3 states that the common support of non-informative features can be consistently identified. Part (ii) of Theorem 4.3 indicates that the estimation error for the nonzero elements is of order $O_p(\lambda_\beta M_r)$. This is similar to the error bound obtained in Theorem 2 of Mai et al. (2012). However, instead of imposing any assumption for the minimal signal $|\beta(\mathbf{u})|_{\min}$ as in Condition 2 of Mai et al. (2012), our error bound relies on the total signal strength $\|\beta(\mathbf{u})\|_1$ through the choice of λ_β.

4.3 Consistency of $\widehat{\eta}(\mathbf{u})$

Theoretical properties of penalized logistic regression have been well explored in the literature; see for example Meier et al. (2008) and Rocha et al. (2009). Other than studying the excess risk or the global error of the estimator $\widehat{\eta}$ as in Meier et al. (2008) or establishing consistency for the coefficient estimators with requires additional assumptions on the signal strength for the nonzero parameters, here we directly explore the estimation accuracy of $\widehat{\eta}(\mathbf{u}) = \widehat{A}_0 + \widehat{\mathbf{A}}^T\mathbf{u}$ in estimating $\eta(\mathbf{u}) = A_0 + \mathbf{A}^T\mathbf{u}$. We first assume the following conditions hold:

(C5). Let $\mathbf{W} = \mathbf{U}I\{L=1\} + \mathbf{V}I\{L=2\}$ where $\mathbf{U}$ and $\mathbf{V}$ are independent binary vectors with density functions $p_1(\cdot)$ and $p_2(\cdot)$, respectively. We assume that the covariance matrix $\mathbf{H} = \mathrm{Var}(\mathbf{W})$ is positive definite with eigenvalues bounded away from zero.

(C6). Let $\mathcal{M} = \{j : A_j \neq 0\}$ and $M = |\mathbf{A}_1|_0$. We assume that $1 - \max_{e \in \mathcal{M}^c} |\mathbf{H}_{e,\mathcal{M}} \mathbf{H}^{-1}_{\mathcal{M},\mathcal{M}}| > 0$, and $M^2 e^M \sqrt{\frac{\log d}{n}} \to 0$.

Condition (C6) is an irrepresentability assumption to guarantee model selection consistency. The following theorem provides the uniform estimation error for $\widehat{\eta}(\mathbf{u})$.

Theorem 4.4. *Let Conditions (C5) and (C6) hold. We have*

$$\sup_{\mathbf{u} \in \{0,1\}^d} |\widehat{\eta}(\mathbf{u}) - \eta(\mathbf{u})| = O_p\left(M^2 e^M \sqrt{\frac{\log d}{n}} \right).$$

Proof. Denote the objective function on the right hand side of (3.6) as $L_n(\eta)$. The true linear function $\eta(\mathbf{W}) = A_0 + \mathbf{A}^T\mathbf{W}$ is the minimizer of the expected risk:

$$El(\tilde{\eta}; \mathbf{W}, L)) = E\{-(2 - L)\tilde{\eta}(\mathbf{W}) + \log(1 + \exp\{\tilde{\eta}(\mathbf{W})\})\}.$$

Let $\widehat{\mathcal{M}} := \{j : \widehat{A}_j \neq 0\}$. Using similar arguments as in the proof of Proposition 4.1, we can first show that under Conditions (C5) and (C6), $P(\widehat{\mathcal{M}} = \mathcal{M}) \to 1$. For convenience we shall assume hereafter in this proof that $\widehat{\mathcal{M}} = \mathcal{M}$.

Let $\mathcal{B}_\eta(\delta)$ be the set of linear functions $\bar{\eta}(\mathbf{w}) = \bar{A}_0 + \bar{\mathbf{A}}_1^T \mathbf{w}$, such that $\sum_{i=0}^d |\bar{A}_i - A_i| = \delta$ for some $\delta > 0$. Here $\bar{A}_i$ is the i-th element of $\bar{\mathbf{A}}_1$, and $\bar{A}_i = 0$ for $i \in \mathcal{M}^c$. With some abuse of notations, let $W_i = \mathbf{U}_i, \tilde{W}_i = \binom{1}{W_i}$ for $i = 1 \ldots, n_1$, $W_{n_1+i} = \mathbf{V}_i, \tilde{W}_{n_1+i} = \binom{1}{W_{n_1+i}}$ for $i = 1 \ldots, n_2$, and denote $\mathbf{W}_n = (\tilde{W}_1, \ldots, \tilde{W}_n)$. The Hessian matrix of L_n is then given as:

$$H_n(\eta) = \sum_{i=1}^n \tilde{W}_i \tilde{W}_i^T \frac{\exp\{\eta(W_i)\}}{(1 + \exp\{\eta(W_i)\})^2}.$$

By Taylor expansion we have, there exists an $\eta^*(\mathbf{w}) = A_0^* + \mathbf{w}^T \mathbf{A}_1^*$, such that $A_0^* \in$

$[\bar{A}_0, A_0]$, $A_i^* \in [\bar{A}_i, A_i]$ for $i \in \mathcal{M}$, $A_i^* = 0$ for $i \in \mathcal{M}^c$, and

$$L_n(\bar{\eta}) = L_n(\eta) + L_n'(\eta)\begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix} + \begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix}^T H_n(\eta^*)\begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix}.$$

Note that from Bernstein's inequality (Lin and Bai, 2011) and the fact that $EL_n'(\eta) = \mathbf{0}$, we have, there exists a constant $C > 0$ such that with probability larger than $1 - O(d^{-1})$,

$$\left| L_n'(\eta)\begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix} \right| \le C\delta M \sqrt{\frac{\log d}{n}},$$

holds for all $\bar{\eta} \in \mathcal{B}_\eta(\delta)$. On the other hand, using Condition (C5) and the fact that $(\bar{A}_0 - A_0)^2 + \|\bar{\mathbf{A}}_1 - \mathbf{A}_1\|_2^2 \ge M^{-1}\delta^2$, we have with probability larger than $1 - O(d^{-1})$,

$$\begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix}^T H_n(\eta^*)\begin{pmatrix} \bar{A}_0 - A_0 \\ \bar{\mathbf{A}}_1 - \mathbf{A}_1 \end{pmatrix} \ge \frac{\delta^2}{Me^M},$$

holds for all $\bar{\eta} \in \mathcal{B}_\eta(\delta)$. Consequently, by choosing $\delta = C_1 M^2 e^M \sqrt{\frac{\log d}{n}}$ for some large enough constant $C_1 > 0$, we have with probability larger than $1 - O(d^{-1})$, $L_n(\bar{\eta}) < L_n(\eta)$ holds for all $\bar{\eta} \in \mathcal{B}_\eta(\delta)$. Consequently, by continuity and convexity of the objective function L_n, we conclude that with probability tending to 1, the minimizer $\widehat{\eta}$ of L_n must lie inside the L_1-ball with radius δ, i.e., $|\widehat{\eta}(\mathbf{w}) - \eta(\mathbf{w})| \le \sum_{i=0}^d |\widehat{A}_i - A_i| < \delta$. This proves the theorem. □

From Theorem 4.4, the uniform error bound is reduced to $O_p\left(\sqrt{\frac{\log d}{n}}\right)$ when the number of nonzero parameters M is bounded.

4.4 Misclassification rate

Given $\mathbf{U} = \mathbf{u}$, denote $D(\mathbf{Z}; \mathbf{u}) = \beta(\mathbf{u})^T \left[\mathbf{Z} - \frac{\mu_1(\mathbf{u})+\mu_2(\mathbf{u})}{2}\right] + \eta(\mathbf{u})$ and $\widehat{D}(\mathbf{Z}; \mathbf{u}) = \widehat{\beta}(\mathbf{u})^T \left[\mathbf{Z} - \frac{\widehat{\mu}_1(\mathbf{u})+\widehat{\mu}_2(\mathbf{u})}{2}\right] + \widehat{\eta}(\mathbf{u})$. The optimal Bayes' risk is given as

$$R_{\mathbf{u}} = \pi_1 R_{\mathbf{u}}(2|1) + \pi_2 R_{\mathbf{u}}(1|2),$$

where $R_{\mathbf{u}}(2|1) = P(D(\mathbf{Z}; \mathbf{u}) \leq 0|\mathbf{Z} \in N(\mu_1(\mathbf{u}), \Sigma(\mathbf{u})))$ and $R_{\mathbf{u}}(1|2) = P(D(\mathbf{Z}; \mathbf{u}) > 0|\mathbf{Z} \in N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u})))$. Correspondingly, the misclassification rate of our proposed method is given as

$$\widehat{R}_{\mathbf{u}} = \pi_1 \widehat{R}_{\mathbf{u}}(2|1) + \pi_2 \widehat{R}_{\mathbf{u}}(1|2),$$

where $\widehat{R}_{\mathbf{u}}(2|1) = P(\widehat{D}(\mathbf{Z}; \mathbf{u}) \leq 0|\mathbf{Z} \in N(\mu_1(\mathbf{u}), \Sigma(\mathbf{u})))$ and $\widehat{R}_{\mathbf{u}}(1|2) = P(\widehat{D}(\mathbf{Z}, \mathbf{u}) > 0|\mathbf{Z} \in N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u})))$. Note that when $\mathbf{Z} \in N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u}))$, we have,

$$D(\mathbf{Z}; \mathbf{u}) \sim N\left(\beta(\mathbf{u})^T[\mu_2(\mathbf{u}) - \mu_1(\mathbf{u})]/2 + \eta(\mathbf{u}), \beta(\mathbf{u})^T[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})]\right).$$

Denote $\Delta(\mathbf{u}) := \beta(\mathbf{u})^T[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})] = [\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})]^T \Sigma^{-1}(\mathbf{u})[\mu_1(\mathbf{u}) - \mu_2(\mathbf{u})]$. We assume that:

(C7). $\sup_{\mathbf{u} \in \{0,1\}^d} \Delta(\mathbf{u}) \geq \delta$ for some constant $\delta > 0$.

We remark that $\Delta(\mathbf{u})$ captures how far are the (normalized) centers of the two classes away from each other. Condition (C7) ensures that the center of the two class are separable. The following theorem indicates that the estimated semiparametric classification rule $I\{\widehat{D}(\mathbf{Z}; \mathbf{u}) > 0\}$ is asymptotically optimal.

Theorem 4.5. *Let Conditions (C1)-(C7) hold. For a given $\mathbf{u} \in \{0,1\}^d$ of interest, let $s = \sup_{\mathbf{u} \in B_{\mathbf{v}}(r)} \|\beta(\mathbf{u})\|_0$. We have:*

$$\sup_{\mathbf{u} \in B_{\mathbf{v}}(r)} |\widehat{R}_{\mathbf{u}} - R_{\mathbf{u}}| \tag{4.13}$$

$$= O_p\left(\Big(s + \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\beta(\mathbf{u})|_1\Big) M_r \left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) + M^2 e^M \sqrt{\frac{\log d}{n}}\right).$$

Proof. By definition we have:

$$\widehat{D}(\mathbf{Z},\mathbf{u}) - D(\mathbf{Z},\mathbf{u}) = \Bigg\{\mathbf{Z}^T[\widehat{\beta}(\mathbf{u}) - \beta(\mathbf{u})] - \frac{1}{2}\widehat{\beta}(\mathbf{u})^T[\widehat{\mu}_1(\mathbf{u}) + \widehat{\mu}_2(\mathbf{u})]$$
$$+\frac{1}{2}\beta(\mathbf{u})^T[\mu_1(\mathbf{u}) + \mu_2(\mathbf{u})] + [\widehat{\eta}(\mathbf{u}) - \eta(\mathbf{u})]\Bigg\}.$$

Write $\mathbf{Z} = (Z_1, \cdots, Z_p)^T$. From Theorem 4.3, we have: $\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\mathbf{Z}^T[\widehat{\beta}(\mathbf{u}) - \beta(\mathbf{u})]| \leq \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} \sum_{i=1}^p |Z_i[\widehat{\beta}(\mathbf{u}) - \beta(\mathbf{u})]_i| = O_p\left(s_{\mathbf{u}} M_r \left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right)\right)$. Similarly, we have

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\widehat{\beta}(\mathbf{u})^T[\widehat{\mu}_1(\mathbf{u}) + \widehat{\mu}_2(\mathbf{u})] - \beta(\mathbf{u})^T[\mu_1(\mathbf{u}) + \mu_2(\mathbf{u})]|$$
$$= O_p\left(\Big(s + \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\beta(\mathbf{u})|_1\Big) M_r \left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right)\right).$$

Together with Theorem 4.4, we have:

$$\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\widehat{D}(\mathbf{Z};\mathbf{u}) - D(\mathbf{Z};\mathbf{u})| \tag{4.14}$$
$$= O_p\left(\Big(s + \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\beta(\mathbf{u})|_1\Big) M_r \left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) + M^2 e^M \sqrt{\frac{\log d}{n}}\right).$$

Note that

$$\begin{aligned}\widehat{R}(1|2) &= P(\widehat{D}(\mathbf{Z};\mathbf{u}) > 0 | \mathbf{Z} \in N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u})))\\ &= P(D(\mathbf{Z};\mathbf{u}) > D(\mathbf{Z};\mathbf{u}) - \widehat{D}(\mathbf{Z};\mathbf{u}) | \mathbf{Z} \in N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u}))).\end{aligned}$$

and

$$\begin{aligned}\widehat{R}(2|1) &= P(\widehat{D}(\mathbf{Z};\mathbf{u}) < 0|\mathbf{Z} \in N(\mu_1(\mathbf{u}), \Sigma(\mathbf{u})))\\ &= P(D(\mathbf{Z};\mathbf{u}) < D(\mathbf{Z};\mathbf{u}) - \widehat{D}(\mathbf{Z};\mathbf{u})|\mathbf{Z} \in N(\mu_1(\mathbf{u}), \Sigma(\mathbf{u}))).\end{aligned}$$

On the other hand, for a given $\mathbf{u}$, denote the density of $D(\mathbf{Z};\mathbf{u})$ as $F_i(\cdot;\mathbf{u})$ for $\mathbf{Z}$ in class $i = 1, 2$. Under Conditions (C5) and (C6), we have $F_i(\cdot;\mathbf{u})$ is bounded from above. Together with (4.14), we conclude that

$$\begin{aligned}&\widehat{R}(1|2) - R(1|2)\\ =\ & O_p\left(\int_0^{\left(s+\sup_{\mathbf{u}\in B_{\mathbf{v}}(r)}|\beta(\mathbf{u})|_1\right)M_r\left(\sqrt{\frac{\log(p+d)}{n}}+\kappa(h)\right)+M^2e^M\sqrt{\frac{\log d}{n}}} F_2(z;\mathbf{u})dz\right)\\ =\ & O_p\left(\left(s + \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\beta(\mathbf{u})|_1\right) M_r\left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) + M^2e^M\sqrt{\frac{\log d}{n}}\right).\end{aligned}$$

Similarly,

$$\begin{aligned}&\widehat{R}(2|1) - R(2|1)\\ =\ & O_p\left(\left(s + \sup_{\mathbf{u}\in B_{\mathbf{v}}(r)} |\beta(\mathbf{u})|_1\right) M_r\left(\sqrt{\frac{\log(p+d)}{n}} + \kappa(h)\right) + M^2e^M\sqrt{\frac{\log d}{n}}\right).\end{aligned}$$

Then the theorem can be proved. □

There are two terms on the right hand side of (4.13). The first term is similar to those for other high dimensional LDA classifiers with continuous data only (Cai and Liu, 2011; Jiang et al., 2020), and is mainly introduced by the estimation of $\beta(\mathbf{u})$. The second term is introduced by the estimation of $\eta(\mathbf{u})$. Although both p and d are allowed to grow exponentially in n, the sparsity requirement for the discrete variables seems to be more restricted as we require M^2e^M to be $o\big(\sqrt{\frac{\log d}{n}}\big)$.

Chapter 5

Numerical Study

5.1 Tuning parameters for simulation

In what follows we will introduce the cross validation procedures for determining the bandwidths and the tuning parameters λ_β and λ_η. We remark that owing to Proposition 3.1, the determination of λ_β and λ_η can be conducted separately, results in additive computation cost other than multiplicative computation cost.

In this section, we will use the following two methods for bandwidth selection. Correspondingly, we conduct two simulation studies using two methods under six models.

5.1.1 Bandwidth selection I

One approach to choose the bandwidths h_x, h_y, h_{xx} and h_{yy} is to use leave-one-out cross validation as is usually done in kernel smoothing. However, this bandwidth selection I approach does not recognize the constraint in our theory that the bandwidths should be large enough (see Condition C2) to guarantee an analogue of Bochner's Lemma (i.e., Lemma 4.2) to hold. As an alternative, we propose to select the bandwidths and the tuning parameter λ_β together by minimizing classification error.

For simplicity, we shall use a common bandwidth parameter h for all the bandwidths h_x, h_y, h_{xx} and h_{yy}. Suppose we reformulate the weights by introducing $\theta = \frac{\exp\{-(dh)^{-1}\}}{1+\exp\{-(dh)^{-1}\}}$. That is, by writing $\exp\{-(dh)^{-1}t\} = \left(\frac{\theta}{1-\theta}\right)^t$, we have

$$\widehat{\mu}_1(\mathbf{u}) = \sum_{j=1}^{n_1} \frac{(\frac{\theta}{1-\theta})^{|\mathbf{U}_j-\mathbf{u}|_1}\mathbf{X}_j}{\sum_{j=1}^{n_1}(\frac{\theta}{1-\theta})^{|\mathbf{U}_j-\mathbf{u}|_1}}, \qquad \widehat{\mu}_2(\mathbf{u}) = \sum_{j=1}^{n_2} \frac{(\frac{\theta}{1-\theta})^{|\mathbf{V}_j-\mathbf{u}|_1}\mathbf{Y}_j}{\sum_{j=1}^{n_2}(\frac{\theta}{1-\theta})^{|\mathbf{V}_j-\mathbf{u}|_1}}.$$

Clearly $\theta \in [0, 0.5]$. When $\theta \to \frac{1}{2}$, $\widehat{\mu}_{1i}(\mathbf{u})$ and $\widehat{\mu}_{2i}(\mathbf{u})$ reduce to the means of all the samples, and when $\theta \to 0$, $\widehat{\mu}_{1i}(\mathbf{u})$ and $\widehat{\mu}_{2i}(\mathbf{u})$ reduce to the sample means in the cells only. Coincidentally, under this formulation, we found that subject to a normalizing term $(1-\theta_x)^d$, the denominator is the same as the smoothing estimator for the distribution of a high dimensional and binary random vector in Aitchison and Aitken (1976) , Grund and Hall (1993). We shall be adopting this new formulation for tuning selection, as the parameter θ is now bounded, which is practically more convenient for tuning.

5.1.2 Bandwidth selection I: λ_β for the estimation of $\beta(\mathbf{u})$

Note that Proposition 3.1 implies that the estimation of $\beta(\mathbf{u})$ can be independently conducted by minimizes the expected misclassification rate over the class of zero-intercept classifiers. More specifically, For given (θ, λ_β), let $\widehat{\beta}_{-i}(\mathbf{U}_i)$ and $\widehat{\beta}_{-i}(\mathbf{V}_i)$ be the estimators obtained using (3.5) by leaving $(\mathbf{X}_i, \mathbf{U}_i)$ and $(\mathbf{Y}_i, \mathbf{V}_i)$ out, respectively. We choose (θ, λ_β) such that the following misclassification rate is minimized:

$$\begin{aligned} R_0(\lambda_\beta) &= \sum_{i=1}^{n_1} I\left\{\widehat{\beta}_{-i}(\mathbf{U}_i)^T\left(\mathbf{X}_i - \frac{\widehat{\mu}_{1,-i}(\mathbf{U}_i)+\widehat{\mu}_2(\mathbf{U}_i)}{2}\right) \le 0\right\} \\ &\quad + \sum_{j=1}^{n_2} I\left\{\widehat{\beta}_{-j}(\mathbf{V}_j)^T\left(\mathbf{Y}_j - \frac{\widehat{\mu}_1(\mathbf{V}_j)+\widehat{\mu}_{2,-j}(\mathbf{V}_j)}{2}\right) \ge 0\right\}. \end{aligned}$$

5.1.3 Bandwidth selection I : λ_η for the estimation of $\eta(\mathbf{u})$

Given the chosen (θ, λ_β), we denote $\zeta_i := \widehat{\beta}_{-i}(\mathbf{U}_i)^T \left(\mathbf{X}_i - \frac{\widehat{\mu}_{1,-i}(\mathbf{U}_i)+\widehat{\mu}_2(\mathbf{U}_i)}{2}\right)$ and $\zeta_{n_1+j} = \widehat{\beta}_{-j}(\mathbf{V}_j)^T \left(\mathbf{Y}_j - \frac{\widehat{\mu}_1(\mathbf{V}_j)+\widehat{\mu}_{2,-j}(\mathbf{V}_j)}{2}\right)$, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Note that these values have been computed when determining λ_β and hence it requires no extra computation burden. Let $(\widehat{A}_{0,-i}, \widehat{\mathbf{A}}_{-i})$ and $(\widehat{A}_{0,-(n_1+j)}, \widehat{\mathbf{A}}_{-(n_1+j)})$ be the estimator based on (3.6) by leaving $\mathbf{U}_i$ and $\mathbf{V}_j$ out, respectively. We then choose λ_η by minimizing the following misclassification rate:

$$\begin{aligned} R(\lambda_\eta) &= \sum_{i=1}^{n_1} I\left\{\zeta_i + \widehat{A}_{0,-i} + \mathbf{U}_i^T \widehat{\mathbf{A}}_{-i} \le 0\right\} \\ &\quad + \sum_{j=1}^{n_2} I\left\{\zeta_{n_1+j} + \widehat{A}_{0,-(n_1+j)} + \mathbf{V}_j^T \widehat{\mathbf{A}}_{-(n_1+j)} \ge 0\right\}. \end{aligned}$$

5.1.4 Bandwidth selection II

We also introduce the way to choose the bandwidths h_x, h_y, h_{xx} and h_{yy} is to use leave-one-out cross validation as is usually done in kernel smoothing. More specifically, h_x is obtained by minimizing the following cross-validation score:

$$CV(h_x) = \frac{1}{n_1} \sum_{i=1}^{n_1} \|\mathbf{X}_i - \widehat{\mu}_{1,-i}(\mathbf{U}_i)\|_2^2,$$

where $\widehat{\mu}_{1,-i}(\cdot)$ is the weighted estimator obtained by leaving the ith sample $\mathbf{X}_i$ out. Given h_x and the leave-one-out estimators $\widehat{\mu}_{1,-i}(\cdot), i = 1, \ldots, n_1$, we then choose h_{xx} such that the following cross-validation score is minimized:

$$CV(h_{xx}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \|(\mathbf{X}_i - \widehat{\mu}_{1,-i}(\mathbf{U}_i))(\mathbf{X}_i - \widehat{\mu}_{1,-i}(\mathbf{U}_i))^T - \widehat{\Sigma}_{1,-i}(\mathbf{U}_i)\|_2^2,$$

where $\widehat{\Sigma}_{1,-i}(\mathbf{U}_i)$ is the weighted estimator obtained by leaving the ith sample $\mathbf{X}_i$ out in (4.5). Bandwidths h_y and h_{yy} in Class 2 are chosen in a similar way.

5.1.5 Bandwidth selection II : λ_β for the estimation of $\beta(\mathbf{u})$

Recall that $\beta(\mathbf{u})$ can be estimated independently in Bandwidth selection I. For a given λ_β, let $\widehat{\beta}_{-i}(\mathbf{U}_i)$ and $\widehat{\beta}_{-i}(\mathbf{V}_i)$ be the estimators obtained using (3.5) by leaving $(\mathbf{X}_i, \mathbf{U}_i)$ and $(\mathbf{Y}_i, \mathbf{V}_i)$ out, respectively. We choose λ_β such that the following misclassification rate is minimized:

$$\begin{aligned} R_0(\lambda_\beta) &= \sum_{i=1}^{n_1} I\left\{\widehat{\beta}_{-i}(\mathbf{U}_i)^T \left(\mathbf{X}_i - \frac{\widehat{\mu}_{1,-i}(\mathbf{U}_i) + \widehat{\mu}_2(\mathbf{U}_i)}{2}\right) \le 0\right\} \\ &\quad + \sum_{j=1}^{n_2} I\left\{\widehat{\beta}_{-j}(\mathbf{V}_j)^T \left(\mathbf{Y}_j - \frac{\widehat{\mu}_1(\mathbf{V}_j) + \widehat{\mu}_{2,-j}(\mathbf{V}_j)}{2}\right) \ge 0\right\}. \end{aligned}$$

We use warm start for the above cross validation procedure to further accelerate the computation speed.

5.1.6 Bandwidth selection II : λ_η for the estimation of $\eta(\mathbf{u})$

Given the chosen λ_β, we denote ζ_i and ζ_{n_1+j} same as those in Bandwidth selection I. Leaving $\mathbf{U}_i$ and $\mathbf{V}_j$ out, we also denote $(\widehat{A}_{0,-i}, \widehat{\mathbf{A}}_{-i})$ and $(\widehat{A}_{0,-(n_1+j)}, \widehat{\mathbf{A}}_{-(n_1+j)})$ as the estimator based on (3.6). Then λ_η can be chosen by minimizing the misclassification rate:

$$\begin{aligned} R(\lambda_\eta) &= \sum_{i=1}^{n_1} I\left\{\zeta_i + \widehat{A}_{0,-i} + \mathbf{U}_i^T \widehat{\mathbf{A}}_{-i} \le 0\right\} \\ &\quad + \sum_{j=1}^{n_2} I\left\{\zeta_{n_1+j} + \widehat{A}_{0,-(n_1+j)} + \mathbf{V}_j^T \widehat{\mathbf{A}}_{-(n_1+j)} \ge 0\right\}. \end{aligned}$$

5.2 Simulation I

Let $\mathbf{u} = (u_1, \ldots, u_d)^T$ be a generic location in $\{0,1\}^d$. Given location $\mathbf{u}$, samples from class 1 are generated from $N(\mu_1(\mathbf{u}), \Sigma(\mathbf{u}))$ and samples from class 2 are generated from $N(\mu_2(\mathbf{u}), \Sigma(\mathbf{u}))$. In this simulation, we set the mean functions $\mu_1(\mathbf{u}) = (\mu_{11}(\mathbf{u}), \cdots, \mu_{1p}(\mathbf{u}))^T$ and $\mu_2(\mathbf{u}) = (\mu_{21}(\mathbf{u}), \cdots, \mu_{2p}(\mathbf{u}))^T$ are set as $\mu_1(\mathbf{u}) = \frac{\Sigma(\mathbf{u})\beta(\mathbf{u})}{2} = -\mu_2(\mathbf{u})$. In the following we consider different settings for $\Sigma(\mathbf{u})$ and $\beta(\mathbf{u})$:

Model 1. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = \bar{u}^{|i-j|}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^d u_k$, and let $\beta_1(\mathbf{u}) = \beta_2(\mathbf{u}) = 5\big(\frac{\sum_{k=1}^d u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\big), \beta_3(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

Model 2. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = [2\bar{u}(1-\bar{u})]^{|i-j|}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^d u_k$, and let $\beta_1(\mathbf{u}) = \beta_2(\mathbf{u}) = \beta_3(\mathbf{u}) == 5\big(\frac{\sum_{k=1}^d u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\big), \beta_4(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

Model 3. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = \frac{\bar{u}^{|i-j|}}{\exp\{\bar{u}|i-j|\}}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^d u_k$, and we set

$\beta_1(\mathbf{u}) = \cdots = \beta_{15}(\mathbf{u}) = \text{sign}\Big(\frac{\sum_{k=1}^d u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\Big)\frac{1}{2}\exp\Big\{|\frac{2\sum_{k=1}^d u_k}{\sqrt{d}} - \sqrt{d}|\Big\}$, and $\beta_{16}(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

Model 4. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = \bar{u}^{\frac{|i-j|}{2}}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^d u_k$, and let $\beta_1(\mathbf{u}) = \beta_2(\mathbf{u}) = 5\big(\frac{\sum_{k=1}^d u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\big)$, and $\beta_3(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

Model 5. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = [3\bar{u}(1-\bar{u})]^{|i-j|}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^{d} u_k$, and let $\beta_1(\mathbf{u}) = \beta_2(\mathbf{u}) = \beta_3(\mathbf{u}) == 5\big(\frac{\sum_{k=1}^{d} u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\big), \beta_4(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

Model 6. We set the (i,j)th element of $\Sigma(\mathbf{u})$ as $\sigma_{ij}(\mathbf{u}) = \frac{\bar{u}^{|i-j|}}{\exp\{\bar{u}|i-j|\}}, i,j \in 1, \cdots, p$, where $\bar{u} = d^{-1}\sum_{k=1}^{d} u_k$, and we set
$\beta_1(\mathbf{u}) = \cdots = \beta_5(\mathbf{u}) = \text{sign}\Big(\frac{\sum_{k=1}^{d} u_k}{\sqrt{d}} - \frac{\sqrt{d}}{2}\Big)\frac{1}{2}\exp\Big\{|\frac{2\sum_{k=1}^{d} u_k}{\sqrt{d}} - \sqrt{d}|\Big\}$, and $\beta_6(\mathbf{u}) = \cdots = \beta_p(\mathbf{u}) = 0$.

The locations $\mathbf{V}_i = (V_{i1}, \ldots, V_{id})^T$ in Class 2 are randomly generated by $P(V_{ij} = 0) = 0.5, i = 1, 2, \cdots, n_2, j = 1, 2, \cdots, d$. Similarly, for Class 1, we generate the $\mathbf{U}_i = (U_{i1}, \ldots, U_{id})^T$ by $P(U_{ij} = 1) = 0.5 + \xi_j$ for $i = 1, 2, \cdots, n_1, j = 1, 2, 3, 4, 5$. We simply set $\xi_1 = \cdots = \xi_5 = 0.25$ and $\xi_6 = \cdots = \xi_d = 0$ for Model 1 and Model 2; $\xi_1 = \cdots = \xi_5 = 0.3$, $\xi_6 = \cdots = \xi_d = 0$ for Model3, Model 4, Model 5 and Model 6. $P(U_{ij} = 1) = 0.5$ for $i = 1, 2, \cdots, n, j = 6, 7, \cdots, d$ for class 1.

We use SLM to denote our proposed semiparametric location model. For comparison, we also consider the following classifiers:

- SLM: our proposed Semiparametric Location Model.
- PLG: l_1 Penalized Logistic Regression(Meier et al., 2008).
- RF: Random Forest(Breiman, 2001).
- DSDA: Direct Sparse Discriminant Analysis in Mai et al. (2012).

We first fix the sample size to be $n_1 = n_2 = 200$, and compare these 4 methods under different dimensions for the discrete and continuous variables: $(d,p) = (10,20), (20,50)$ and $(50,100)$. The misclassification rates of these methods on 200

Model 1			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.158	0.177	0.191
R_{SLM}	0.208(0.031)	0.231(0.031)	0.266(0.035)
R_{PLG}	0.216(0.035)	0.272(0.037)	0.315(0.041)
R_{RF}	0.233(0.029)	0.294(0.032)	0.338(0.037)
R_{DSDA}	0.213(0.032)	0.263(0.030)	0.295(0.036)

Table 5.1: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 1 with $n_1 = n_2 = 200$.

testing samples are computed over 100 replications, in the situation that more replications can cause time-consuming. The means and standard deviations of these misclassification rates are reported in Tables 5.1-5.6, from which we observe that our proposed SLM classifier outperforms other classifiers. The performance of DSDA is slightly better than that of Penalized Logistic Regression. Random Forest, is comparable to other methods when d and p are small, but the misclassification rates become larger than those of the other methods in cases where d and p are large.

Next, we fix the dimensions (d, p), and let the sample size $n = n_1 + n_2$ increase from 200 to 500 (with $n_1 = n_2$). The regret, which is defined as the misclassification rate minus the Bayes risk, was computed for each n. Figures 5.1-5.6 show that as n increases, the regret becomes smaller for all the four methods, while the curves for our semiparametric location model generally produce small regret values among all methods as the sample size n increases.

5.3 Simulation II

Considering the same models (1-6) mentioned in Simulation I, we also show some results of four methods (semiparametric location model, penalized logistic regression, random forest and direct sparse discriminant analysis) based on Bandwidth selection

Model 2			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.125	0.133	0.144
R_{SLM}	0.169(0.024)	0.228(0.035)	0.283(0.033)
R_{PLG}	0.224(0.031)	0.280(0.039)	0.327(0.046)
R_{RF}	0.208(0.028)	0.285(0.033)	0.342(0.036)
R_{DSDA}	0.221(0.028)	0.273(0.034)	0.304(0.039)

Table 5.2: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 2 with $n_1 = n_2 = 200$.

Model 3			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.041	0.093	0.041
R_{SLM}	0.170(0.030)	0.185(0.032)	0.234(0.028)
R_{PLG}	0.190(0.034)	0.222(0.040)	0.264(0.044)
R_{RF}	0.186(0.032)	0.238(0.036)	0.285(0.030)
R_{DSDA}	0.184(0.030)	0.212(0.028)	0.247(0.032)

Table 5.3: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 3 with $n_1 = n_2 = 200$.

II. Under same settings of Simulation I, Tables 5.7-5.12 show the mean and standard deviations of misclassification rates under dimension $(d, p) = (10, 20), (20, 50)$ and $(50, 100)$. And the curves of regrets for all the four methods are shown in Figures 5.7-5.12 with n increasing from 200 to 500.

5.4 Real data analysis

5.4.1 Selection of tuning parameters

We use the first method(Bandwidth selection I) for determining the bandwidths and the tuning parameters in real data case studies.

Model 4			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.147	0.172	0.189
R_{SLM}	0.161(0.028)	0.192(0.034)	0.214(0.030)
R_{PLG}	0.172(0.031)	0.213(0.033)	0.248(0.040)
R_{RF}	0.189(0.028)	0.245(0.035)	0.283(0.033)
R_{DSDA}	0.166(0.031)	0.204(0.029)	0.235(0.034)

Table 5.4: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 4 with $n_1 = n_2 = 200$.

Model 5			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.111	0.119	0.133
R_{SLM}	0.123(0.023)	0.168(0.031)	0.226(0.036)
R_{PLG}	0.174(0.031)	0.223(0.034)	0.252(0.040)
R_{RF}	0.163(0.024)	0.237(0.032)	0.286(0.034)
R_{DSDA}	0.164(0.028)	0.210(0.028)	0.241(0.032)

Table 5.5: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 5 with $n_1 = n_2 = 200$.

5.4.2 Real data cases

In this section, we investigate the performance of the proposed SLM model by analyzing seven real datasets. We compare SLM to the three other classifiers used in simulaiton. In addition, we compute the misclassification rates for Classification Tree (CT), which to some degree can be viewed as a non-ensemble version of RF.

The real cases we studied include: Hepatocellular Carcinoma data, Breast Cancer Gene Expression Profiles (METABRIC) data, Cylinder bands data, Heart-Disease data, Australian credit card application data, Hepatitis data and German Credit data. All these datasets are publicly available on the UCI Machine Learning Repos-

Model 6			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.145	0.173	0.166
R_{SLM}	0.176(0.029)	0.193(0.029)	0.235(0.032)
R_{PLG}	0.205(0.039)	0.231(0.037)	0.264(0.043)
R_{RF}	0.203(0.030)	0.232(0.031)	0.289(0.031)
R_{DSDA}	0.195(0.029)	0.220(0.032)	0.247(0.033)

Table 5.6: Simulation I : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 6 with $n_1 = n_2 = 200$.

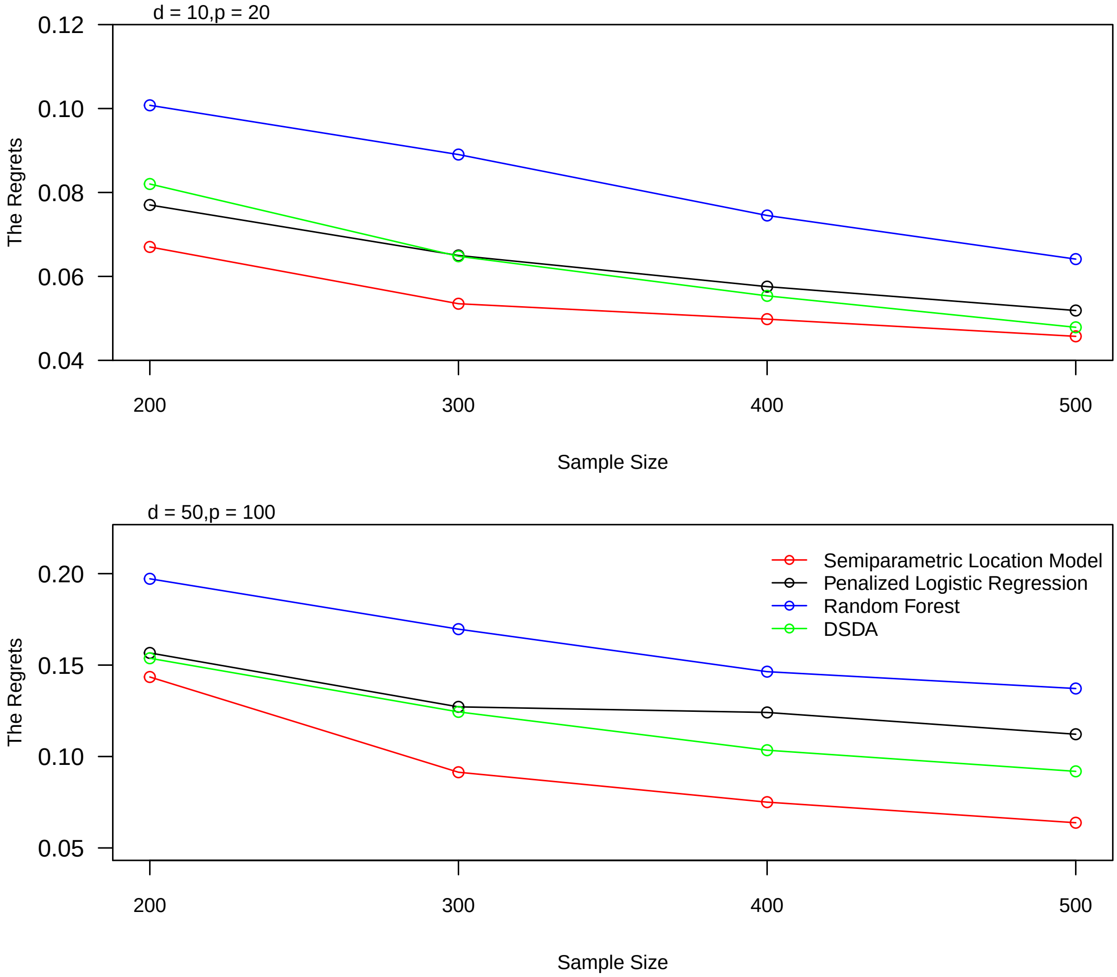

Figure 5.1: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 1.

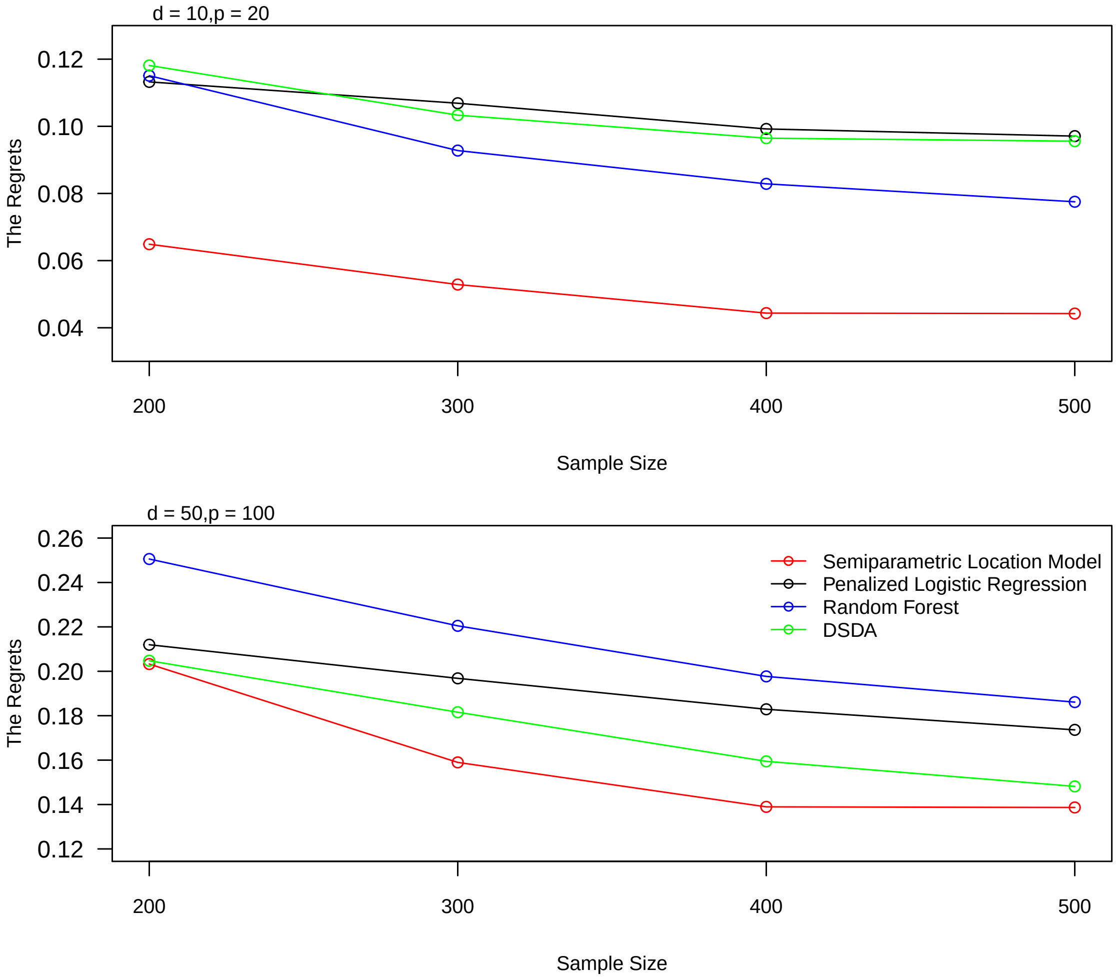

Figure 5.2: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 2

Model 1			
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayes Risk	0.159	0.175	0.188
R_{SLM}	0.208(0.030)	0.238(0.036)	0.275(0.044)
R_{PLG}	0.220(0.031)	0.268(0.033)	0.306(0.033)
R_{RF}	0.236(0.033)	0.295(0.033)	0.338(0.037)
R_{DSDA}	0.219(0.031)	0.263(0.031)	0.294(0.037)

Table 5.7: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 1 with $n_1 = n_2 = 200$.

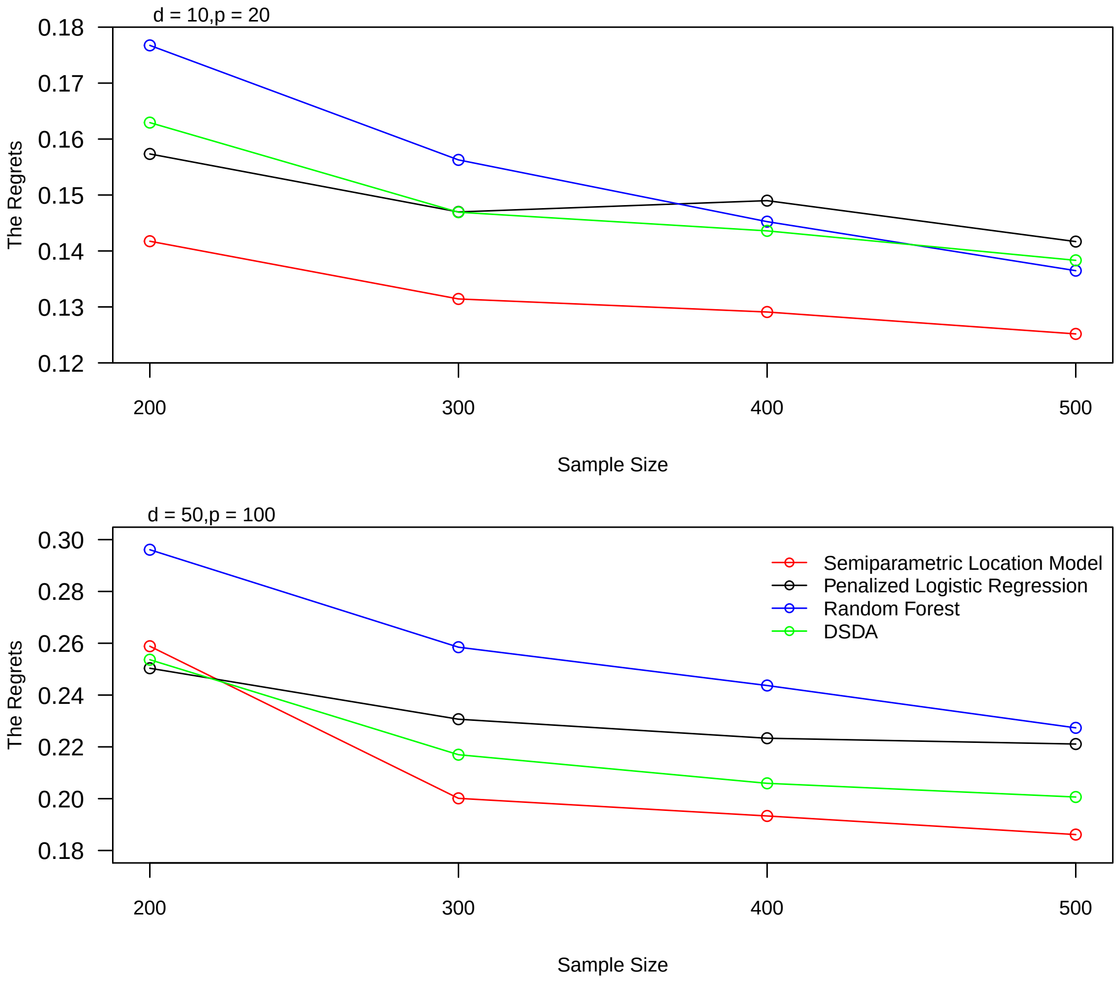

Figure 5.3: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 3

itory or the public data platform Kaggle. All categorical variables are translated into binary variables using dummy variable encoding. Missingness in the categorical variable is treated as one category, and mean imputation is used for missing values of the continuous variables. We perform a 10-fold cross-validation and the average misclassification rates are reported in Table 5.13. The datasets we considered are described as follows:

Hepatocellular Carcinoma dataset Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer in adults and the third leading cause of cancer-related death worldwide. This dataset was collected at a University Hospital in

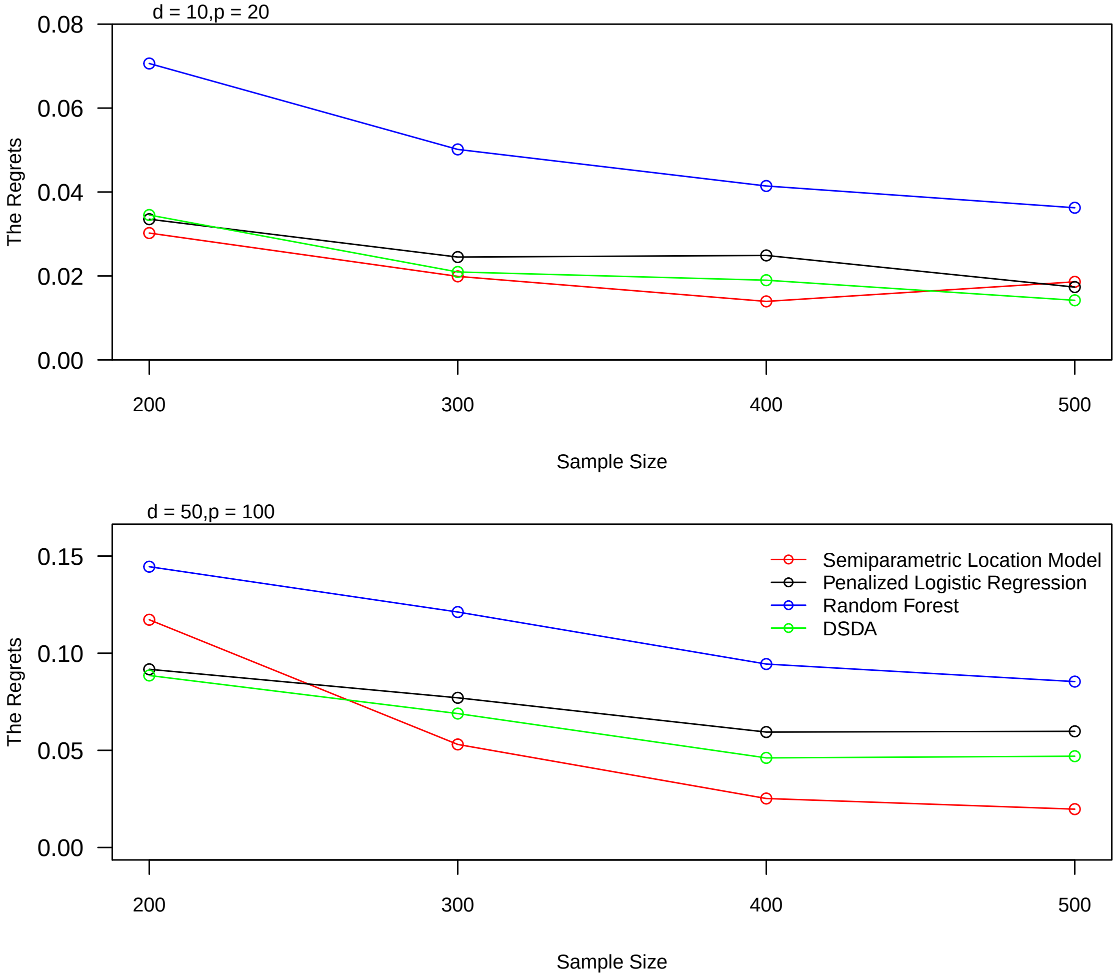

Figure 5.4: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 4

Portugal. It contains real clinical data of 165 patients diagnosed with HCC, in which only 102 patients finally survived. There are 22 continuous variables and 118 binary variables.

Breast Cancer Gene Expression Profiles (METABRIC) Breast cancer is the most frequent cancer among women, and one of the leading causes of cancer deaths in females. This dataset comes from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database. The original data was published on Nature Communications (Pereira et al., 2016). There are 1904 patients with breast cancer in this data. 489 mRNA Z-scores for 331 genes, and indicators of

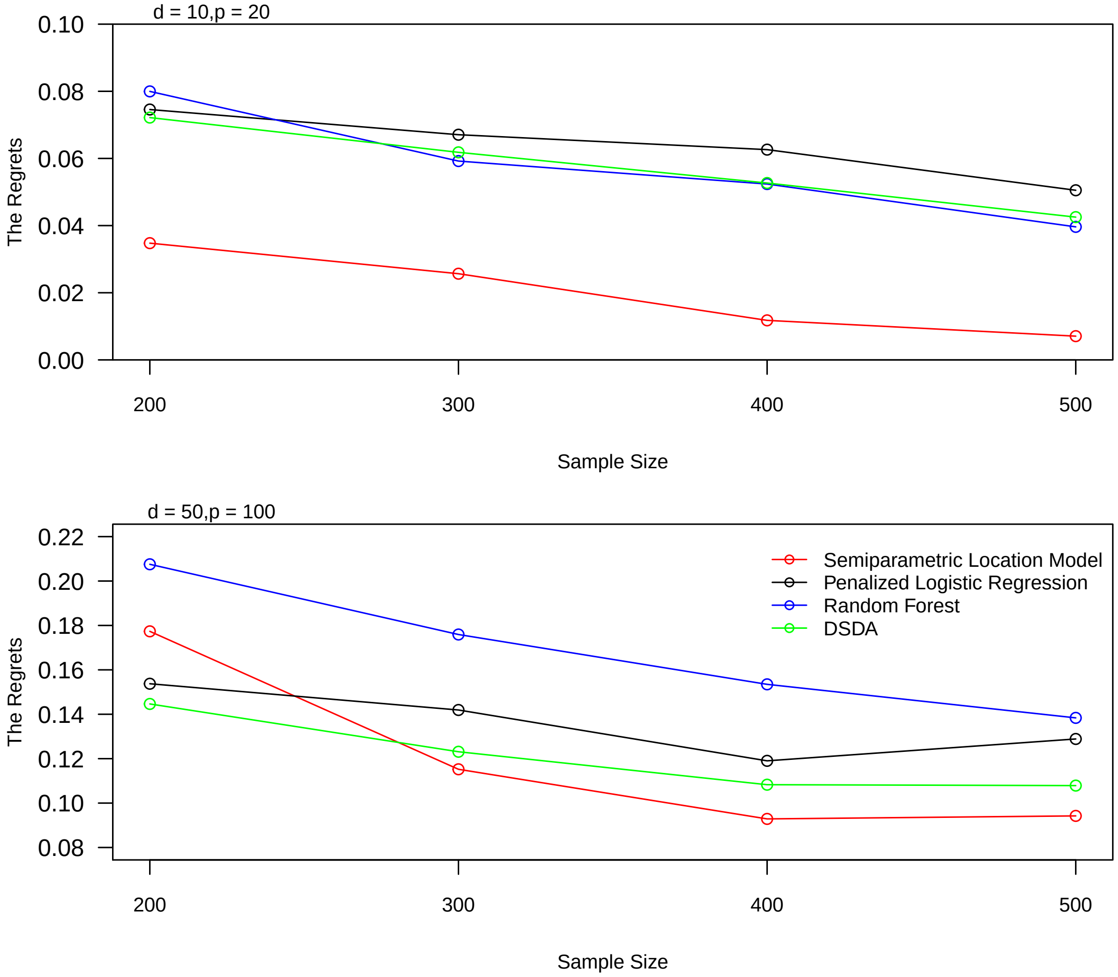

Figure 5.5: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 5

mutation for 173 genes are recorded. To reduce the computational costs, following Cai and Liu (2011), only 100 mRNA Z-scores with the largest absolute values of the two sample t statistics are used.

Cylinder bands data Cylinder bands data contains 20 categorical variables which were transformed into 482 binary variables via dummy variable encoding. There are 20 continuous variables, but we have removed variables “ESA Amperage” and “chrome content”, owing to the fact that more than 95% of the observations are taking a same value or having a missing value for these two variables. This dataset contains 277 instances.

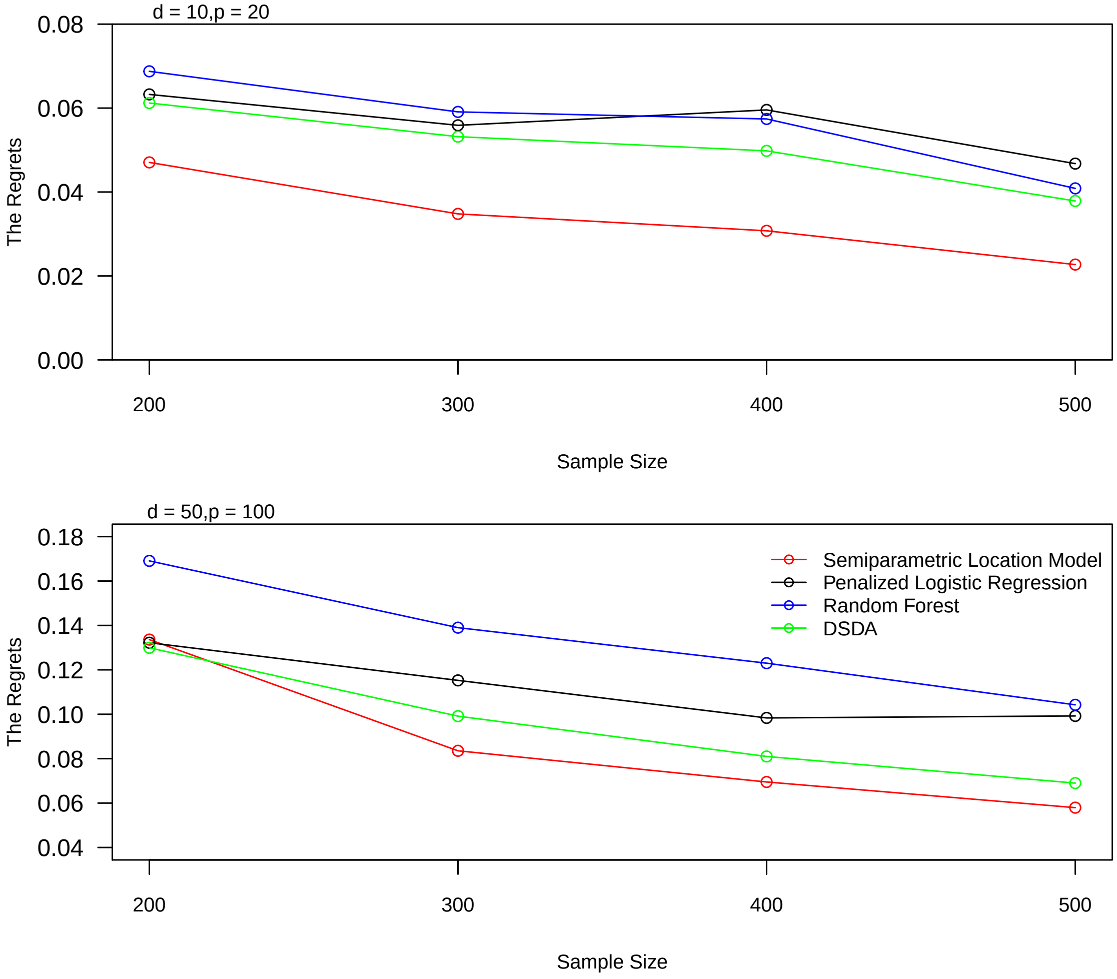

Figure 5.6: Simulation I : The regrets of SLM,PLG,RF,DSDA under Model 6

Heart-Disease data The heart-disease dataset contains 13 attributes (which have been extracted from a larger set of 75). There are 7 categorical variables which can be transformed into 19 binary variables and 6 continuous variables. This dataset contains 270 instances. And there are no missing values in this dataset.

Hepatitis data Hepatitis dataset contains 155 patients diagnosed with hepatitis, in which 123 patients are lived. Missingness in the categorical variable of this dataset is treated as one category. Overall, this dataset contains 7 continuous variables, and 12 categorical variables which were transformed into 32 binary variables.

Australian Credit Card Approval data This dataset concerns credit card appli-

	Model 2		
	n = 400		
(d, p)	(10, 20)	(20, 50)	(50, 100)
$Bayesrisk$	0.127	0.134	0.144
R_{SLM}	0.180(0.030)	0.227(0.032)	0.301(0.039)
R_{PLG}	0.227(0.037)	0.280(0.036)	0.324(0.046)
R_{RF}	0.213(0.032)	0.286(0.035)	0.348(0.038)
R_{DSDA}	0.224(0.034)	0.278(0.035)	0.308(0.035)

Table 5.8: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 2 with $n_1 = n_2 = 200$.

	Model 3		
	n = 400		
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayesrisk	0.042	0.093	0.041
R_{SLM}	0.172(0.030)	0.188(0.027)	0.242(0.038)
R_{PLG}	0.186(0.031)	0.220(0.030)	0.265(0.037)
R_{RF}	0.187(0.031)	0.235(0.034)	0.280(0.038)
R_{DSDA}	0.184(0.026)	0.214(0.028)	0.254(0.035)

Table 5.9: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 3 with $n_1 = n_2 = 200$.

cations. There are 6 numerical variables, and 8 categorical attributions which were transformed into 36 binary variables via dummy variable encoding. This dataset contains 690 instances.

German Credit data In this German Credit dataset, there are 7 continuous variables such as duration in month and credit amount, and 13 categorical variables which were transformed into 54 binary variables. The objective is to class a customer as a “good” or “bad” customer. This dataset contains 1000 instances.

	Model 4		
	n = 400		
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayesrisk	0.149	0.172	0.188
R_{SLM}	0.164(0.028)	0.189(0.035)	0.220(0.036)
R_{PLG}	0.174(0.031)	0.220(0.035)	0.255(0.039)
R_{RF}	0.192(0.036)	0.249(0.035)	0.287(0.037)
R_{DSDA}	0.166(0.026)	0.210(0.030)	0.239(0.030)

Table 5.10: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 4 with $n_1 = n_2 = 200$.

	Model 5		
	n = 400		
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayesrisk	0.113	0.119	0.134
R_{SLM}	0.128(0.024)	0.170(0.032)	0.244(0.042)
R_{PLG}	0.169(0.032)	0.223(0.043)	0.266(0.043)
R_{RF}	0.161(0.028)	0.229(0.033)	0.296(0.037)
R_{DSDA}	0.162(0.028)	0.218(0.031)	0.245(0.033)

Table 5.11: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 5 with $n_1 = n_2 = 200$.

From Table 5.13 we observe that our method is the best classifier for six out of seven datasets, and is the second best for the ACA data. Random Forest performs well too, being the best classifier for the Australian Credit Card Approval dataset and among the best two classifiers in another four datasets. The CT method seems to be the worse classifier overall.

	Model 6		
	n = 400		
(d, p)	(10, 20)	(20, 50)	(50, 100)
Bayesrisk	0.147	0.171	0.164
R_{SLM}	0.174(0.031)	0.192(0.030)	0.247(0.047)
R_{PLG}	0.202(0.036)	0.230(0.032)	0.270(0.040)
R_{RF}	0.197(0.033)	0.233(0.033)	0.282(0.029)
R_{DSDA}	0.193(0.031)	0.216(0.032)	0.251(0.031)

Table 5.12: Simulation II : The mean and sd of the misclassification rates of SLM, PLG, RF and DSDA over 100 replications under Model 6 with $n_1 = n_2 = 200$.

	SLM	PLG	RF	DSDA	CT
Hepatocellular Carcinoma	**0.231**	0.279	0.291	0.255	0.376
Breast Cancer	**0.357**	0.393	0.358	0.366	0.420
Cylinder Bands	**0.332**	0.379	0.368	0.394	0.411
Heart-Disease	**0.148**	0.163	0.178	0.159	0.296
Hepatitis	**0.135**	0.180	0.155	0.174	0.199
Australian Credit Card Approval	0.142	0.142	**0.122**	0.146	0.139
German Credit	**0.236**	0.263	0.246	0.253	0.285

Table 5.13: Classification errors for real data study under 10-fold cross-validation. The best classifier for each data set is highlighted in boldface.

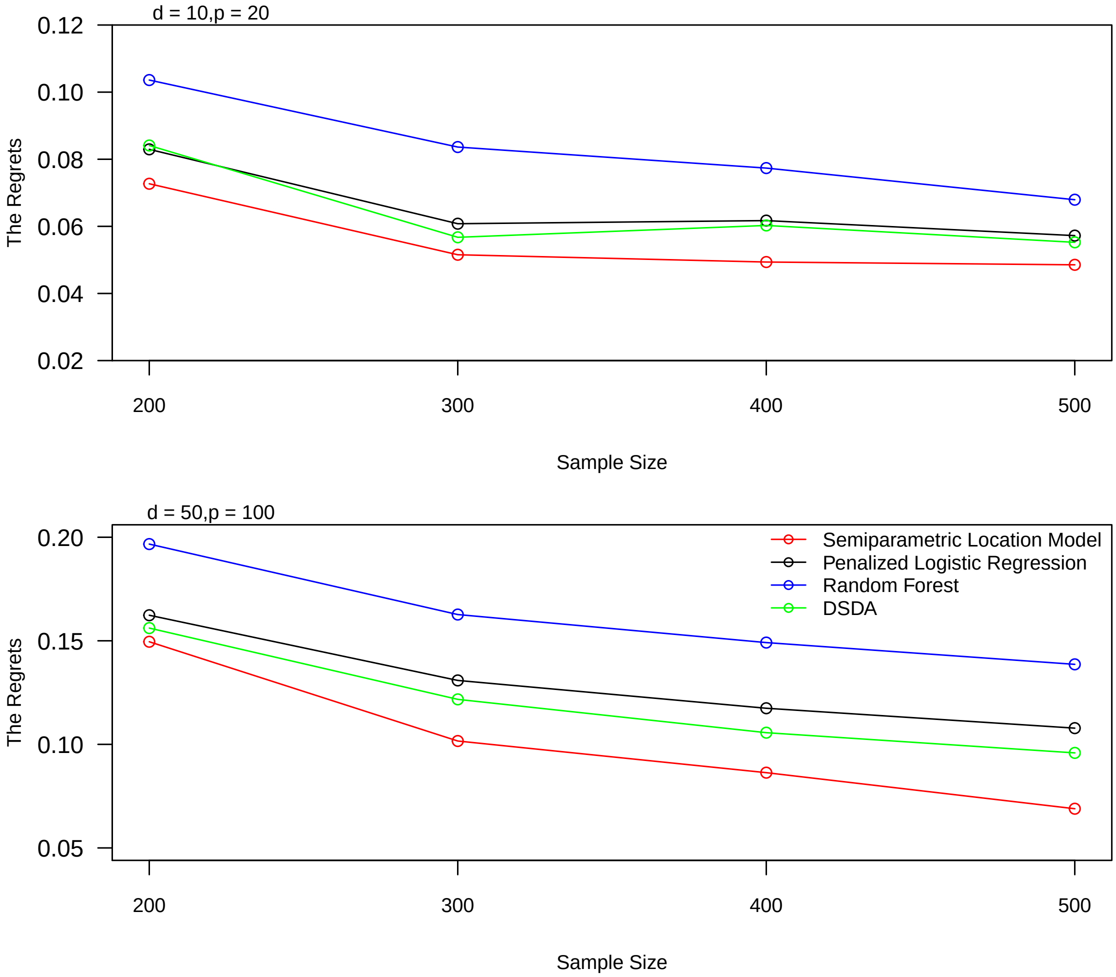

Figure 5.7: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 1.

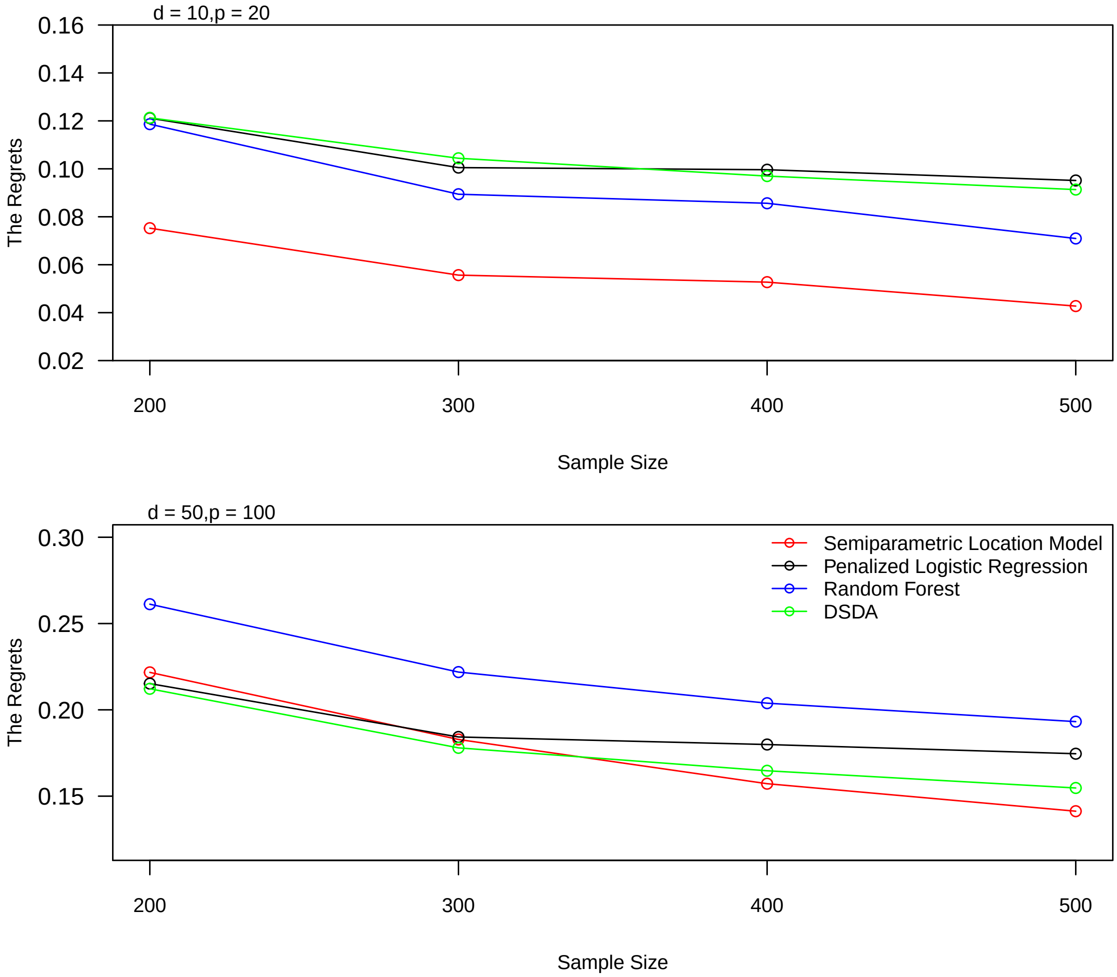

Figure 5.8: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 2

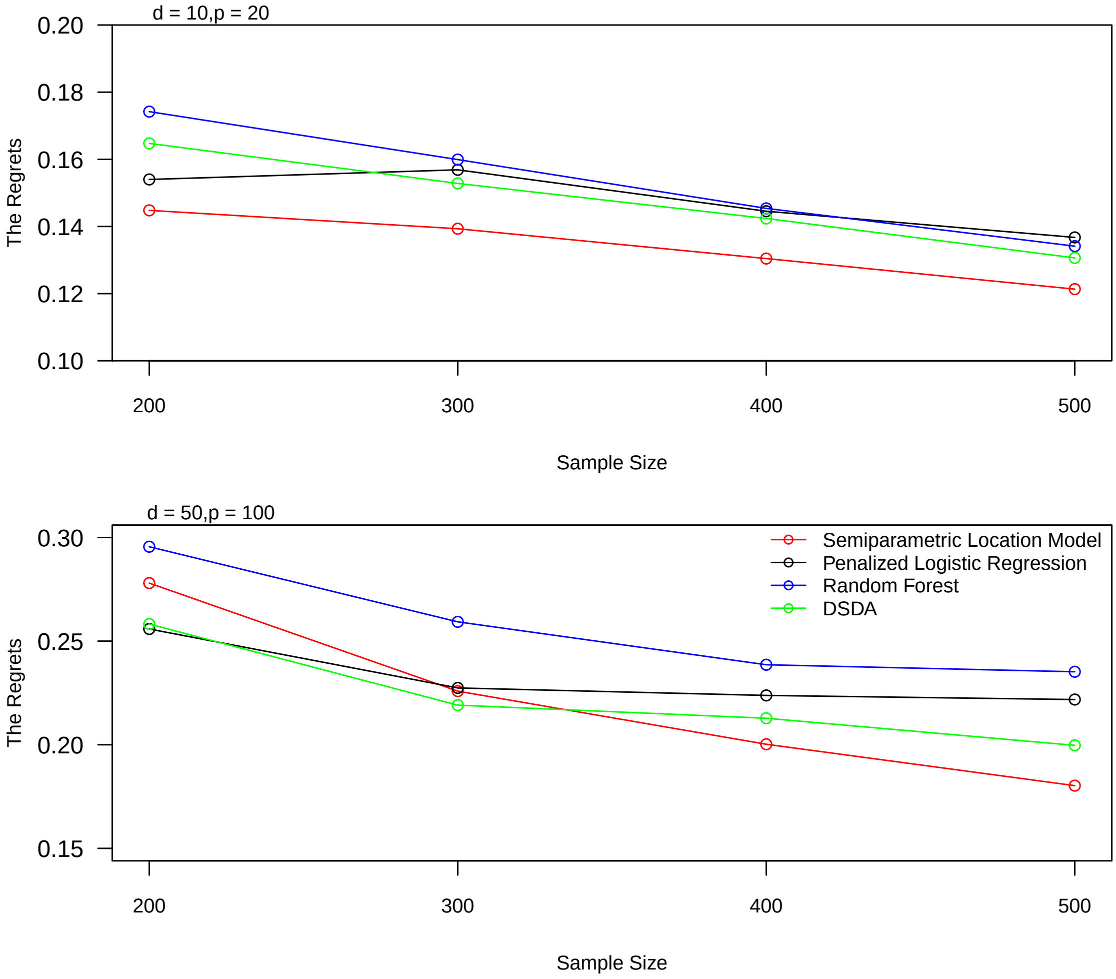

Figure 5.9: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 3

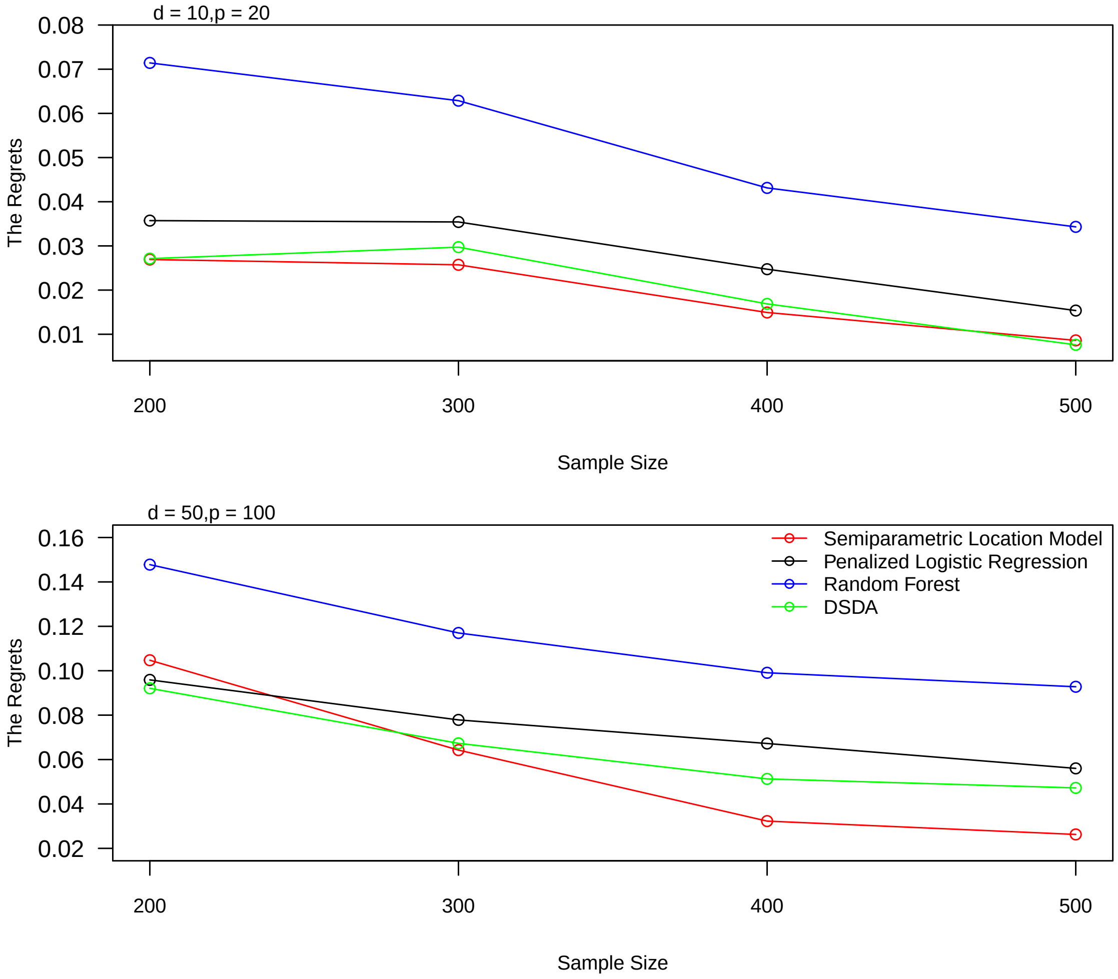

Figure 5.10: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 4

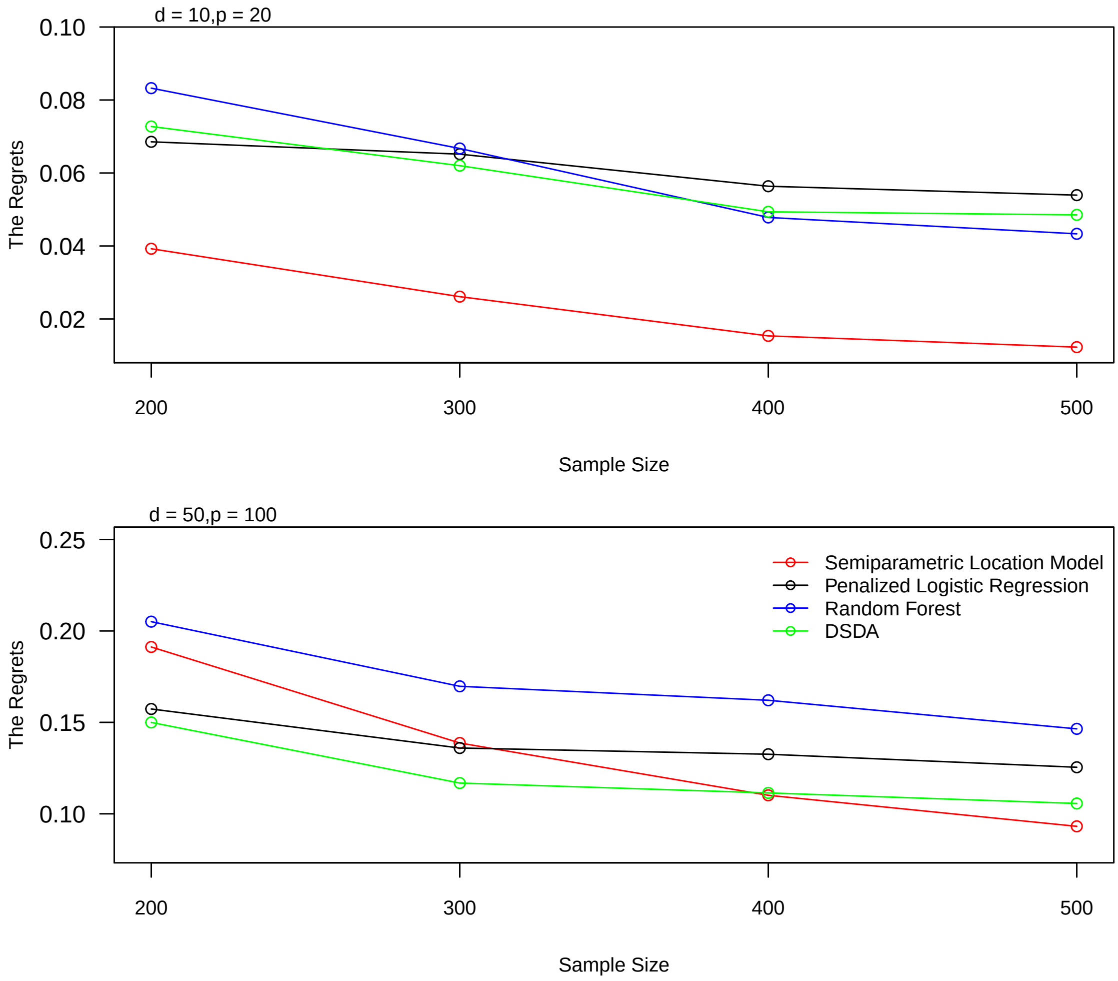

Figure 5.11: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 5

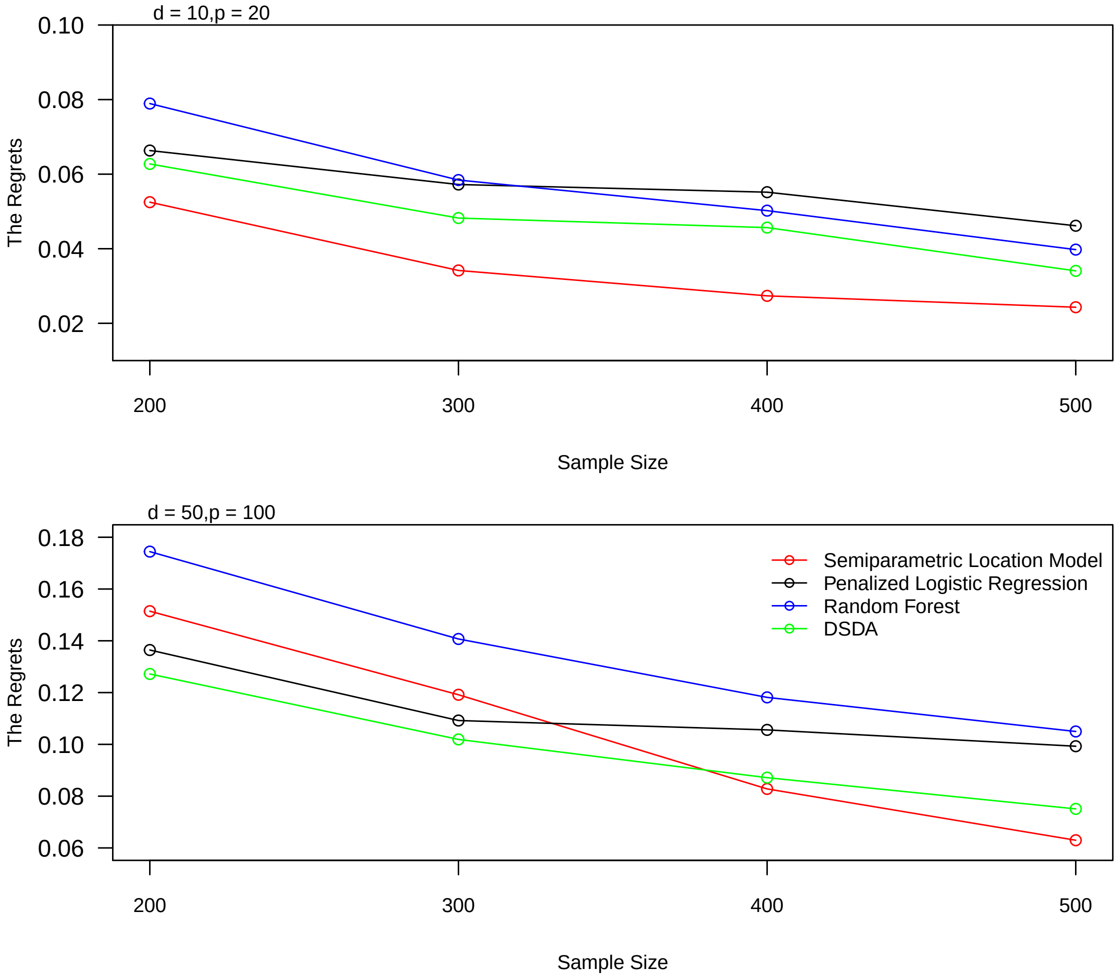

Figure 5.12: Simulation II : The regrets of SLM,PLG,RF,DSDA under Model 6

Chapter 6

Discussion

High dimensional data sets containing both discrete and continuous variables arise frequently in practice in the past decades. In order to classify data with high dimensional mixed variables simultaneously, we have proposed a semiparametric location model based on the optimal Bayes rule in this thesis. Recall that $\mathbf{U}$ is the discrete variable which is viewed as a random "location" in $\{0,1\}^d$. The traditional Bayes rule classifies observations based on the joint distribution of $(\mathbf{X},\mathbf{U})$ in each class. In our approach, the joint distribution is formulated via the distributions of $\mathbf{X}|\mathbf{U}$ and $\mathbf{U}$. Here, $\mathbf{X}|\mathbf{U}$ has Gaussian assumption, with the mean and covariance matrix defined as functions of the location. We have shown that under this location model, the direction $\beta(\mathbf{u})$ and the the intercept $\eta(\mathbf{u})$ of the classification rule can be estimated respectively. To address the curse of dimensionality, we consider the direction to be vary smoothly over the locations and intercept to be approximated by a linear expansion. The estimation methods we have adopted in this thesis have covered (low order) parametric approximation, and nonparametric smoothing. Different combinations of the parametric and nonparametric approaches could result in different theories and performance in different applications.

In this thesis, we expect to convey that discrete variables and continuous variables should be treated differently, and more dedicated modeling strategies should be considered in the presence of other complex structures. Apart from the semiparametric classifier we have developed, there are other possible variations that we can consider. First, we can impose different structures for $\beta(\mathbf{u})$ and $\eta(\mathbf{u})$. For example, similar to $\eta(\mathbf{u})$, we can also adopt a linear approximation for the classification direction $\beta(\mathbf{u})$ and the mean and covariance functions. The resulting classifier will reduce to a classifier with linear effects $\mathbf{X}$, $\mathbf{U}$ and their interactions. Second, we can also consider relaxing the Gaussian assumption on $\mathbf{X}$ to a copula model as in Jiang and Leng (2016) and Mai and Zou (2015). Add robust to the assumption of Gaussian assumption of continuous variables can be discussed. Third, it is interesting to apply the idea of the location model to develop ensemble classifiers using for example random subspace for high dimensional data with mixed variables (Tian and Feng, 2021). Furthermore, the proposed idea can be extended to handle the case where $\mathbf{X}$ admits a matrix or tensor structure (Pan et al., 2019). We leave these for future exploration. Last but not least, in practical cases, we can try to get the misclassification rates only based on continuous variables then compared it with the misclassification rates based on all discrete and continuous variables. From this we can see the practical effect of the proposed location model.

Bibliography

John Aitchison and Colin GG Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63(3):413–420, 1976.

TW Anderson. *An introduction to multivariate statistical analysis, 3rd ed. Wiley Series in Probability and Statistics. Wiley, New York*.

O Asparoukhov and Wojtek J Krzanowski. Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, 10(4):289–297, 2000.

Peter J Bickel and Elizaveta Levina. 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, 2004.

Denis Bosq. *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, volume 110. Springer Science & Business Media, 2012.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

William G Cochran and Carl E Hopkins. Some classification problems with multivariate qualitative data. *Biometrics*, 17(1):10–32, 1961.

JJ Daudin. Selection of variables in mixed-variable discriminant analysis. *Biometrics*, 42(3):473–481, 1986.

AR De Leon, A Soo, and T Williamson. Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 38(5):1021–1032, 2011.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Birgit Grund and Peter Hall. On the performance of kernel estimators for high-dimensional, sparse binary data. *Journal of Multivariate Analysis*, 44(2):321–344, 1993.

Peter Hall. On nonparametric multivariate binary discrimination. *Biometrika*, 68 (1):287–294, 1981.

Hashibah Hamid. A new approach for classifying large number of mixed variables. In *International Conference on Computer and Applied Mathematics*, pages 156–161, 2010.

Seok Young Hong and Oliver Linton. Asymptotic properties of a nadaraya-watson type estimator for regression functions of infinite order. *arXiv preprint arXiv:1604.06380*, 2016.

Binyan Jiang and Chenlei Leng. High dimensional discrimination analysis via a semiparametric model. *Statistics & Probability Letters*, 110:103–110, 2016.

Binyan Jiang, Xiangyu Wang, and Chenlei Leng. A direct approach for sparse quadratic discriminant analysis. *The Journal of Machine Learning Research*, 19 (1):1098–1134, 2018.

Binyan Jiang, Ziqi Chen, and Chenlei Leng. Dynamic linear discriminant analysis in high dimensional space. *Bernoulli*, 26(2):1234–1268, 2020.

James D Knoke. Discriminant analysis with discrete and continuous variables. *Biometrics*, 38(1):191–200, 1982.

Célestin C Kokonendji and Mona Ibrahim. Associated kernel discriminant analysis for multivariate mixed data. *Electronic Journal of Applied Statistical Analysis*, 9 (2):385–399, 2016.

WJ Krzanowski. The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10(1):25–49, 1993.

WJ Krzanowski. Quadratic location discriminant functions for mixed categorical and continuous data. *Statistics & Probability Letters*, 19(2):91–95, 1994.

WJ Krzanowski. Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Computational Statistics & Data Analysis*, 19(4):419–431, 1995.

Wojtek J Krzanowski. Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, 70(352):782–790, 1975.

Wojtek J Krzanowski. Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36(3):493–499, 1980.

Zhengyan Lin and Zhidong Bai. *Probability Inequalities*. Springer Science & Business Media, 2011.

Nor Idayu Mahat, Wojtek Janusz Krzanowski, and Adolfo Hernandez. Variable selection in discriminant analysis based on the location model for mixed variables. *Advances in Data Analysis and Classification*, 1(2):105–122, 2007.

Qing Mai and Hui Zou. A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics*, 55(2):243–246, 2013.

Qing Mai and Hui Zou. Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

Qing Mai, Yi Yang, and Hui Zou. Multiclass sparse discriminant analysis. *Statistica Sinica*, 29:97–111, 2019.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Ingram Olkin and Robert Fleming Tate. Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2): 448–465, 1961.

Yuqing Pan, Qing Mai, and Xin Zhang. Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association*, 114(527):1305–1319, 2019.

Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(1): 1–16, 2016.

Guilherme V Rocha, Xing Wang, and Bin Yu. Asymptotic distribution and sparsistency for ℓ_1-penalized parametric m-estimators with applications to linear svm and logistic regression. *arXiv preprint arXiv:0908.1940*, 2009.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39 (2):1241–1265, 2011.

Ye Tian and Yang Feng. Rase: A variable screening framework via random subspace ensembles. *Journal of the American Statistical Association*, (to appear):1–30, 2021.

Fengrong Wei, Jian Huang, and Hongzhe Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515–1540, 2011.

Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435, 1992.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

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