Fall 12-2016

Functional Principal Component Analysis for Discretely Observed Functional Data and Sparse Fisher?s Discriminant Analysis with Thresholded Linear Constraints

Jing Wang
Georgia State University

Functional Principal Component Analysis for Discretely Observed Functional Data and Sparse Fisher's Discriminant Analysis with Thresholded Linear Constraints

by

Jing Wang

Under the Direction of Xin Qi, PhD

ABSTRACT

We propose a new method to perform functional principal component analysis (FPCA) for discretely observed functional data by solving successive optimization problems. The new framework can be applied to both regularly and irregularly observed data, and to both dense and sparse data. Our method does not require estimates of the individual sample functions or the covariance functions. Hence, it can be used to analyze functional data

with multidimensional arguments (e.g. random surfaces). Furthermore, it can be applied to many processes and models with complicated or nonsmooth covariance functions. In our method, smoothness of eigenfunctions is controlled by directly imposing roughness penalties on eigenfunctions, which makes it more efficient and flexible to tune the smoothness. Efficient algorithms for solving the successive optimization problems are proposed. We provide the existence and characterization of the solutions to the successive optimization problems. The consistency of our method is also proved. Through simulations, we demonstrate that our method performs well in the cases with smooth samples curves, with discontinuous sample curves and nonsmooth covariance and with sample functions having two dimensional arguments (random surfaces), repectively. We apply our method to classification problems of retinal pigment epithelial cells in eyes of mice and to longitudinal CD4 counts data. In the second part of this dissertation, we propose a sparse Fisher's discriminant analysis method with thresholded linear constraints. Various regularized linear discriminant analysis (LDA) methods have been proposed to address the problems of the LDA in high-dimensional settings. Asymptotic optimality has been established for some of these methods when there are only two classes. A difficulty in the asymptotic study for the multiclass classification is that for the two-class classification, the classification boundary is a hyperplane and an explicit formula for the classification error exists, however, in the case of multiclass, the boundary is usually complicated and no explicit formula for the error generally exist. Another difficulty in proving the asymptotic consistency and optimality for sparse Fisher's discriminant analysis is that the covariance matrix is involved in the constraints of the optimization problems for high order components. It is not easy to estimate a general high-dimensional covariance matrix. Thus, we propose a sparse Fisher's discriminant analysis method which avoids the estimation of the

covariance matrix, provide asymptotic consistency results and the corresponding convergence rates for all components. To prove the asymptotic optimality, we provide an asymptotic upper bound for a general linear classification rule in the case of muticlass which is applied to our method to obtain the asymptotic optimality and the corresponding convergence rate. In the special case of two classes, our method achieves the same as or better convergence rates compared to the existing method. The proposed method is applied to multivariate functional data with wavelet transformations.

INDEX WORDS: Functional PCA, discretely observed functional data, successive optimization problems, roughness penalty, consistency, sparse Fisher's discriminant analysis, thresholded linear constraints, asymptotic consistency, asymptotic optimality, convergence rate

Functional Principal Component Analysis for Discretely Observed Functional Data and Sparse Fisher's Discriminant Analysis with Thresholded Linear Constraints

by

Jing Wang

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in the College of Arts and Sciences
Georgia State University
2016

Functional Principal Component Analysis for Discretely Observed Functional Data and Sparse Fisher's Discriminant Analysis with Thresholded Linear Constraints

by

Jing Wang

Committee Chair: Xin Qi

Committee: Gengsheng Qin
Yichuan Zhao
Ruiyan Luo

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
October 2016

DEDICATION

This dissertation is dedicated to my dear parents, my advisor,
and all my dear friends.

ACKNOWLEDGEMENTS

I would like to give my deep thanks and sincere gratitude to everyone for helping me to complete this dissertation.

First and foremost I want to express my sincere gratitude to my thesis advisor, Xin Qi, for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D study. He has been my advisor, mentor, collaborator and friend.

I appreciate Professors, Gengsheng Qin, Yichuan Zhao and Ruiyan Luo, who kindly agreed to serve on my dissertation committee and have given me continuous help throughout my graduate study. I am also thankful to them for reading my reports, commenting on my views and helping me understand and enrich my ideas. Without their precious support it would not be possible to conduct this research. Besides, I would like to thank other professors from our department for educating me in various courses.

At last, I want to thank everyone who has helped me throughout my graduate life, especially my immediate family to whom this dissertation is dedicated to. None of this would have been possible without the love and patience of my family. I am also grateful to friends. Their support and care helped me overcome setbacks and stay focused on my graduate study. I greatly value their friendship and I deeply appreciate their belief in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

- GSU - Georgia State University
- PCA - Principal Component Analysis
- FPCA - Functional Principal Component Analysis
- LDA - Linear Discriminant Analysis
- CV - Cross-validation

Chapter 1

INTRODUCTION

Principal component analysis (PCA) is one of the best known techniques in both multivariate analysis and functional data analysis. Different from classical PCA, functional principal component analysis (PCA) requires smoothing or regularizing of the estimated principal component curves (see Chapter 9 in Ramsay and Silverman [33]). Readers can find a general overview of many methods for computing the smoothed functional principal components when the sample curves are fully observed in Ramsay and Silverman [33]. Ferraty and Vieu [16] provides more discussions on nonparametric methods and developments for functional data analysis. However, in practice, the sample functions are usually observed at discrete points with measurement errors. The observation points might be irregular or sparse. Several FPCA methods for discretely sampled functional data or longitudinal data have been developed. Shi, Weiss and Taylor [38], Rice and Wu [35] and James, Hastie and Sugar [23] proposed mixed effects approaches in which individual sample curves or eigenfunctions of the covariance function are represented by basis function expansions. Staniswalis and Lee [44] and Yao, Müller and Wang [54] used nonparametric methods to estimate covariance functions and then obtained the eigenfunctions. Huang, Shen and Buja [22] proposed an FPCA method for regularly observed discrete functional data based on penalized rank one approximation to the data matrix. Peng and Paul [29] assumed a finite rank model for the covariance function, represented the eigenfunctions as basis function expansions, and proposed the restricted maximum likelihood method to estimate the parameters. Other nonparametric approaches to this problem tend to fall into two classes. The approaches in the first class smooth each individual curve by the smoothing spline method or other methods (see section 9.5 in Ramsay and Silverman [33]). Then the smoothed principal component curves can be obtained by the usual functional PCA or other methods. For example, motivated by the duality relation between row and column spaces of a data matrix, Benko et al. [6] proposed an FPCA method for regularly observed discrete functional data. The methods

in the second class assume that covariance functions are smooth. Smoothing methods such as kernel methods and free-knot spline smoothing are used to obtain smoothed estimates of mean functions and covariance functions. Then the principal components curves can be estimated by the eigenfunctions of the smoothed covariance function. However, some of these methods (Huang et al. [22] and Benko et al. [6]) cannot be applied to the discrete functional data in which the observation points are irregular and sparse. Some of them (Staniswalis and Lee [44] and Yao *et al.*[54]) need to estimate the covariance functions, hence it is hard to apply these methods to functional data with two or three dimensional arguments since we have to estimate four or six dimensional covariance functions.

In this dissertation, we first propose a new method to perform FPCA for discretely observed functional data by solving successive optimization problems. The new framework can be applied to both regularly and irregularly observed data, and to both dense and sparse data. First, our method does not need to estimate the individual sample functions or the covariance functions and we do not assume that they are smooth. Hence, it can be easily applied to discretely observed functional data with two or three dimensional arguements and to processes and models with complicated or nonsmooth covariance functions. Most of the current methods assume that either the sample functions or the covariance functions are smooth explicitly or implicitly. Some of them need to obtain the smoothed estimations of the sample functions or the covariance functions. However, there are many important processes and models with nonsmooth sample functions and nonsmooth covariance functions but with smooth eigenfunctions. Our methods can be applied to these processes and models. Some real functional data have complicated covariance functions in which we are not interested. In this case , our methods avoid estimating the complicated covariance functions. Second, our method controls the smoothness of eigenfunctions by directly imposing roughness penalties on eigenfunctions and can use different smoothing parameters for different eigenfunctions. Hence, it is efficient and flexible to tune the smoothness of eigenfunctions in this method. Our methods can also be easily extended to analyze the discretely observed functions defined on high-dimensional spaces, e.g. random surfaces. Section 5 in Müller (2005) listed some open problems concern the application of FDA methods including analysis of random surfaces and higher-dimensional functions. We applied our methods to a simulated discretely observed

random surface data. For high-dimensional data, the covariance functions are defined on higher dimension space with dimensions equal to two times of the dimensions of the sample functions. Hence, it is very hard to obtain good estimations of covariance functions in this case. Efficient algorithms for solving the successive optimization problems are proposed. We provide the existence and characterization of the solutions to the successive optimization problems. The consistency of our method is also proved. The following real example is used to motivate and illustrate the method developed in this dissertation.

The linear discriminant analysis (LDA) has been a favored tool for supervised classification in the settings of small p and large n. However, it faces major problems for high-dimensional data. In theory, Bickel and Levina [7] and Shao et al. [37] showed that the usual LDA can be as bad as the random guessing when $p > n$. In practice, the classic LDA methods have bad predictive performance in high-dimensional settings. To address these problems, various regularized discriminant analysis methods have been proposed, including Friedman [17], Krzanowski et al. [26], Dudoit et al. [13], Bickel and Levina [7], Guo et al. [19], Xu et al. [53], Tibshirani et al. [47], Witten and Tibshirani [52], Clemmensen et al. [11], Shao et al. [37], Cai and Liu [9], Fan et al. [15], Qi et al. [32] and many others. Asymptotic optimality has been established in some of these papers when there are two classes. Shao et al. [37] made sparsity assumptions on both the difference $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the population means of the two class, and the within-class covariance matrix $\boldsymbol{\Sigma}$. Then thresholding procedures were applied to both the difference between the two sample class means and the sample within-class covariance matrix $\widehat{\boldsymbol{\Sigma}}$. The asymptotic optimality and the corresponding convergence rate for their classification rule were obtained. Cai and Liu [9] observed that in the case of two classes, the optimal classification rule depends on $\boldsymbol{\Sigma}$ only through $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. Hence, they assumed l_1 sparsity for $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, proposed a sparse estimate of it through minimizing its l_1 norm with an l_∞ constraint, and provided asymptotic optimality of their classification rule. Fan et al. [15] imposed l_0 sparsity assumption on $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, estimated it through a minimization problem with an l_1 constraint and derived the asymptotic optimality. A major difficulty preventing the derivation of asymptotic optimality of the linear classification rules for multiple classes is that for the two-class classification, the classification boundary of LDA is a hyperplane and an explicit formula for the classification error exists, however,

for the multiclass classification, the classification boundary is usually complicated and no explicit formula for the classification error generally exist.

As a special case of LDA, the Fisher's discriminant analysis projects the original variables $\mathbf{X}$ to a low dimensional subspace to generate new predictor variables, $\mathbf{X}\boldsymbol{\alpha}_1, \mathbf{X}\boldsymbol{\alpha}_2, \ldots, \mathbf{X}\boldsymbol{\alpha}_{K-1}$, where the coefficient vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{K-1}$ are sequentially calculated and K is the number of classes. The coefficient vectors are found by maximizing the between class variation of the new predictor variables relative to their within class variation and the new predictors are orthogonal to each other, that is, the linear constraints $\boldsymbol{\alpha}_i^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = 0$ for any $1 \leq j < i < K$ are satisfied. Once the coefficients are determined and the classification rule is to assign a new observation to the class with the sample class mean closest to this observation in the projection subspace. Besides the complicated classification boundary for multiclass, the linear constraint $\boldsymbol{\alpha}_i^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = 0$ poses additional difficulty in studying the asymptotic consistency and optimality for the Fisher's discriminant analysis in high dimensional setting for $K > 2$ because the covariance matrix $\boldsymbol{\Sigma}$ is involved. It is not easy to find a consistent estimate for a general $\boldsymbol{\Sigma}$ in the high-dimensional settings. Qi et al. [32] introduced a sparse Fisher's discriminant analysis method, an advantage of which is that the proposed algorithm is applicable to any linear constraints imposed on the higher order components. In the second part of this dissertation, instead of aiming to find a consistent estimate of $\boldsymbol{\Sigma}$, we apply a soft-thresholding procedure to obtain a consistent estimate of the subspace $\{\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{i-1}\}$ which defines the linear constraints for $\boldsymbol{\alpha}_i$, for any $1 < i \leq K-1$. Then taking advantage of the algorithm in the paper above, we propose the estimates of $\boldsymbol{\alpha}_i$, for all $1 \leq i \leq K-1$, and an classification rule. We study the theoretical properties of this method in high dimensional settings, including the asymptotic consistency of the estimate of $\boldsymbol{\alpha}_i$ and the subspaces defining the orthogonal constraints, the asymptotic optimality, and the corresponding convergence rates, where the number K of classes can be any fixed positive integer. In the special case of $K = 2$, the asymptotic optimality of the our method is compared to the existing method and our method has the same or better convergence rate. We apply our method to the classification problems for multivariate functional data through the wavelet transformations.

The remainder of the dissertation is organized as follow. In Chapter 2, we present our new method to perform FPCA for discretely observed functional data by solving successive

optimization problems. We first give some background, basic notations and our main assumptions. The classic Silverman's method to perform smoothed FPCA is also introduced. We then present our method along with its theoretical properties, and an algorithm for solving the successive optimization problems in practice. Simulation results with comparison to other established method are reported to illustrate the effectiveness of our method. At last, we apply our method on 2 real data sets: the RPE data set and the Longitudinal CD4 counts data set. In Chapter 3, we propose a sparse Fisher's discriminant analysis method with thresholded linear constraints which avoids the estimation of the covariance matrix. We first introduce notations and briefly review the classic Fisher's discriminant analysis. Then our sparse Fisher's LDA method with thresholded linear constraints are introduced. We also present the main theoretical results along with simulation studies and applications. All proofs of our theorems can be found in the Chapter 4 of the dissertation.

Chapter 2

FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS FOR DISCRETELY OBSERVED FUNCTIONAL DATA

In this chapter, we present our new method to perform function principal analysis (FPCA) for discretely observed functional data by solving successive optimization problems. We first give some background, basic notations and our main assumptions. The classic Silverman's method to perform smoothed FPCA is introduced in Section 2.2. We then present our method along with its theoretical properties, and an algorithm for solving the successive optimization problems in practice. Simulation results are reported to illustrate the effectiveness of our method. At last, we apply our method on the RPE data set and the Longitudinal CD4 counts data set in Section 2.5. The proofs of all theorems are provided in Section 4.11 of Chapter 4.

2.1 Background and Notations

First, we introduce notations and definitions used in this chapter. Let $\mathbb{N}$ denote the collection of all the positive integers. In this chapter, we will mainly consider functions defined in a finite interval $[a, b]$ in the following two spaces, the space $L^2([a, b])$ of square integrable functions

$$L^2([a,b]) = \{f : f \text{ is a measurable function on } [a,b] \text{ and } \int_a^b |f(t)|^2 dt < \infty\}$$

and the Sobolev space $W_2^2([a, b])$ of functions with square integrable second derivatives,

$$W^2([a,b]) = \{f : f, f' \text{ are absolutely continuous on} [a,b] \text{ and} f'' \in L^2([a,b])\}$$

where f' and f'' denote the first and second derivatives of f, respectively. For any f, $g \in L^2([a,b])$, define the usual inner product

$$\langle f, g\rangle = \int_a^b f(t)g(t)dt$$

with corresponding squared norm $\|f\|^2 = \langle f, f\rangle$. Given a smoothing parameter $\alpha \geq 0$, for any $f, g \in W_2^2([a,b])$, define

$$[f, g] = \int_a^b f''(t)g''(t)dt$$

and the inner product

$$\langle f, g\rangle_\alpha = \langle f, g\rangle + \alpha[f, g]$$

with corresponding squared norm $\|f\|_\alpha^2 = \langle f, f\rangle_\alpha$. Here we use the same notations as those in Silverman [41].

Let $X(t)$, $a \leq t \leq b$ be a measurable stochastic process (random function) on $[a,b]$ and $X_1(t), X_2(t), \cdots, X_n(t)$ be i.i.d sample functions from the distribution of $X(t)$. Below we give three basic assumptions on $X(t)$ which are essentially the same as those in Silverman [41].

Assumption 1. $E\left[\|X\|^4\right] = E\left[\left(\int_a^b |X(t)|^2 dt\right)^2\right] < \infty$.

Under Assumption 1, $X(t) \in L^2([a,b])$ a.s.. Assume that mean function $EX(t) = \mu(t)$. Define the covariance function of $X(t)$

$$\Gamma(s,t) = E\left[(X(s) - \mu(s))\,(X(t) - \mu(t))\right], \forall s, t \in [a,b], \tag{2.1}$$

Under Assumption 1, Γ has a sequence of nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and the corresponding eigenfunctions $\gamma_1, \gamma_2, \cdots$. Every eigenfunction has been scaled to have L^2-norm 1. The set of all the eigenfunctions forms an orthonormal basis of $L^2([a,b])$. Furthermore, we have decomposition

$$\Gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j \gamma_j(s)\gamma_j(t)$$

. Suppose that we are interested in estimating the first K eigenvalues and eigenfunctions of the covariance function Γ.

Assumption 2. *Any eigenvalue* $\lambda_j, 1 \leq j \leq K$ *has multiplicity 1, so that* $\lambda_1 > \lambda_2 > \cdots > \lambda_K > \lambda_{K+1}$.

This assumption is just the third assumption in Section 5.2 of Silverman [41]. If an eigenvalue has multiplicity 1, then the corresponding eigenfunction is uniquely determined up to a sign. If the multiplicity is larger than 1, the eigenfunctions can not be uniquely determined up to a sign (Qi and Zhao [31]).

Assumption 3. *The eigenfunctions* $\gamma_j, 1 \leq j \leq K$ *belong to* $W_2^2([a,b])$.

If the covariance function Γ is smooth, then Assumption 3 holds. However, there are many important random processes whose covariance matrices are nonsmooth, but the eigenfunctions belong to $W_2^2([a,b])$ (Qi and Zhao [31]). For example, the continuous parameter AR(1) model in time series, Brownian motion, Poisson process and the stochastic differential equation models driven by them.

Example 1. *(Brownian motion and Poisson process). Consider the standard Brownian motion and the Poisson process with rate 1 in time interval* $[0,1]$. *Their covariance functions are the same and equal to* $\min(s,t), 0 \leq s,t \leq 1$ *(see Page 89 in the book Glasserman [18]) which is nonsmooth. The eigenvalues and eigenfunctions are*

$$\lambda_j = \left(\frac{2}{(2j-1)\pi}\right)^2, \gamma_j = \sqrt{2}\sin\left(\frac{(2j-1)\pi t}{2}\right), j = 1, 2, \cdots \tag{2.2}$$

Example 2. *(Stochastic differential equation models). SDE are widely used to model random processes in may areas. One example is the famous Black-Scholes Model in finance. Let* S_t *denote the price of a stock at time* t. *Then* S_t *satisfies the following SDE,*

$$dS_t = \nu S_t dt + \sigma S_t dW_t$$

where ν *is the instantaneous mean return,* σ *is the instantaneous return volatility and* W_t *is a Brownian motion.*

Another example is the counting processes model in survival analysis. Let N_t *be the number*

of the occurrences of the event in $[0, t]$. *Then* N_t *satisfies*

$$dN_t = \lambda(t)dt + dM_t$$

where $\lambda(t)$ *is a smooth intensity function and* M_t *is a martingale (Qi and Zhao [31]).*

Example 3. *(Continuous parameter models in time series). Consider the con-tinuous parameter AR(1) model in time series. Its covariance function is*

$$\Gamma(s,t) = \frac{e^{-\alpha|s-t|}}{\alpha}$$

where α *is a positive number (see Section 3.7 in Priestley [30]). This covariance function is nonsmooth.*

For these models, the covariance functions are nonsmooth but the eigenfunctions are smooth. In addition to these processes and models, some real functional data have covariance functions with complicated patterns.

2.2 Silverman's approach to smoothed functional PCA

In this section, the independent sample curves from the distribution of $X(t)$,

$$\{X_1(t), X_2(t), \cdots, X_n(t) : a \leq t \leq b\}$$

are assumed to be entirely observed and t could be a continuum in $[a, b]$, or in a two-dimensional region $[a, b] \times [c, d]$, or in higher-dimensional regions. . The covariance function is defined as 2.1 and covariance operator

$$(\Gamma\gamma)(t) = \int_a^b \Gamma(t,s)\gamma(s)ds.$$

For any $\beta, \gamma \in L^2([a, b])$,

$$Cov\left[\langle\beta, X\rangle, \langle\gamma, X\rangle\right] = \langle\beta, \Gamma\gamma\rangle$$

FPCA is one of the key techniques in functional data analysis for patterns discovery and dimension reduction in data sets. The first population functional principal component as a one-dimensional projection of X

$$\langle \gamma_1, X \rangle = \int_a^b \gamma_1(t) X(t) dt, \quad \|\gamma_1\| = 1$$

which maximizes the variance of principal component scores

$$Var(\langle \gamma_1, X \rangle) = \max_{\|\gamma\|=1} Var(\langle \gamma, X \rangle) = \max_{\|\gamma\|=1} \langle \gamma, \Gamma\gamma \rangle = \max_{\|\gamma\|=1} \frac{\langle \gamma, \Gamma\gamma \rangle}{\|\gamma\|^2} \tag{2.3}$$

for all nonzero linear functionals l in $L^2([a,b])$ with the norm $\|l\| = 1$. γ_1 is called the first principal component weight function or the first PC curve. Let λ_1 be the maximum value of (2.3). The pair (λ_1, γ_1) are the first eigenvalue and eigenfunction of Γ (see Section 2, Chapter 3 in Weinberger [51]),

$$\Gamma\gamma_1 = \lambda_1\gamma_1$$

The second functional principal component (γ_2, X): γ_2 is the solution to

$$\max_{\|\gamma\|=1, \langle \gamma, \gamma_1 \rangle = 0} \frac{\langle \gamma, \Gamma\gamma \rangle}{\|\gamma\|^2} \tag{2.4}$$

Let λ_2 be the maximum value of (2.4). The pair (λ_2, γ_2) are the second eigenvalue and eigenfunction of Γ,

$$\Gamma\gamma_2 = \lambda_2\gamma_2$$

Similary, the successive population functional principal components are defined.

However, we usually do not know the true covariance function Γ and the population principal component weight functions can not be obtained directly. We can use the sample covariance function $\hat{\Gamma}_n$ to estimate Γ and use the eigenvalues and eigenfunctions of $\hat{\Gamma}_n$ to estimate the eigenvalues and eigenfunctions of Γ, which are called non-smooth estimators. It was pointed out that the non-smooth principal component curves can show substantial variability (see Chapter 9 in Ramsay and Silverman [33]). Therefore, smoothing of the estimated principal component weight functions is necessary.

Silverman [41] (see also Chapter 9 in Ramsay and Silverman [33]) proposed an important method which incorporates smoothing by replacing the usual L^2 norm with a norm that takes the roughness of the functions into account. Qi and Zhao [31] summarizes the theoretical and practical advantages of Silverman's approach as follows:

> First, the weak assumptions underlying this method make it applicable to data from many fields. Silverman [41] did not make any assumptions on the mean curves and sample curves. Hence, in addition to data with smooth random curves, this method can be applied to analyze data where the sample curves can be unsmooth or even discontinuous, such as those encountered in financial engineering, survival analysis and other fields. For covariance functions, Silverman [41] only assumed that they have series expansions by their eigenfunctions without imposing smoothing constraint. This is attractive because the covariance functions are continuous but unsmooth in many important models such as stochastic differential equation models in financial engineering and counting process models in survival analysis. Second, Silverman's method controls the smoothness of eigenfunction curves by directly imposing roughness penalties on these functions instead of on sample curves or covariance functions. Furthermore, this approach changes the eigenvalue and eigenfunction problems in the usual L^2 space to problems in another Hilbert space, the Sobolev space (with a norm different from the usual norm in the Sobolev space). Therefore, many powerful tools from the theory of Hilbert space can be employed to study the properties of this method. Third, this approach incorporates the smoothing step into the step for computing eigenvalues and eigenfunctions. Therefore, this method is computationally efficient with the same computational load as the usual unsmoothed functional PCA. Fourth, the estimates produced by this method are invariant under scale transformations. As pointed out by Huang et al. [22], the invariance property under scale transformations should be a guiding principle in introducing roughness penalties to functional PCA.

Let α be a nonnegative smoothing parameter. Silverman defines the smoothed estimators

$\{(\hat{\lambda}_j^{[\alpha]}, \hat{\gamma}_j^{[\alpha]}) : j \in \mathbb{N}\}$of $\{(\lambda_j, \gamma_j) : j \in \mathbb{N}\}$ to be the solutions of the following successive optimization problems:
First,$\hat{\gamma}_1^{[\alpha]}$ is the solution of the optimization problem

$$\max_{\|\gamma\|=1} \frac{\langle \gamma, \hat{\Gamma}_n \gamma \rangle}{\langle \gamma, \gamma \rangle + \alpha[\gamma, \gamma]} = \max_{\|\gamma\|=1} \frac{\langle \gamma, \hat{\Gamma}_n \gamma \rangle}{\|\gamma\|_\alpha^2}. \tag{2.5}$$

Let $\hat{\gamma}_1^{[\alpha]}$ be the maximum value of (2.5). For any $k \in \mathbb{N}$, if we have obtained $\{\hat{\gamma}_j^{[\alpha]}, j = 1, 2, \cdots, k-1\}$ and $\{\hat{\lambda}_j^{[\alpha]}, j = 1, 2, \cdots, k-1\}$,$\hat{\gamma}_k^{[\alpha]}$ is the solution of the optimization problem

$$\max_{\substack{\|\gamma\|=1, \langle \gamma, \hat{\gamma}_j^{[\alpha]} \rangle_\alpha = 0, \\ 1 \le j \le k-1}} \frac{\langle \gamma, \hat{\Gamma}_n \gamma \rangle}{\|\gamma\|_\alpha^2} \tag{2.6}$$

and $\hat{\lambda}_k^{[\alpha]}$ is the maximum value of (2.6). Note that $\{(\hat{\lambda}_j^{[\alpha]}, \hat{\gamma}_j^{[\alpha]}) : j \in \mathbb{N}\}$ depends on both the sample size n and the smoothing parameter α (Qi and Zhao [31]).

2.3 Functional PCA for discretely observed functional data

We consider two sample scenarios for sample functions observed at discrete points, regular case and irregular case, respectively.

2.3.1 Regular case

In this case, we assume that the sample functions are observed at the same set $\{t_1, t_2 \cdots, t_m\}$ of discrete observation points across all the subjects with measurement errors, where m is the total number of observation points for each sample function. After sorting the observation points from the smallest to the largest, we get $a = t_{(1)} < t_{(2)} < \cdots t_{(m-1)} < t_{(m)} = b$. Let us consider the following model:

$$Y_{pq} = X_p(t_{(q)}) + \epsilon_{pq}, \quad p = 1, \cdots, n, \quad q = 1, \cdots, m, \tag{2.7}$$

where Y_{pq} is the observation of the sample function X_p at point $t_{(q)}$ with measurement error ϵ_{pq} and n is the total number of sample curves. Our estimates $\{\hat{\lambda}_k, \hat{\gamma}_k\}_{k \ge 1}$ of $\{\lambda_k, \gamma_k\}_{k \ge 1}$ are the solutions to the following successive optimization problems. The first pair of estimates

$\{\hat{\lambda}_1, \hat{\gamma}_1\}$ are the maximum value and the solution to the following optimization problem:

$$\max_{\gamma \in W_2^2([a,b]), \|\gamma\|=1} \frac{\sum_{q=1}^m \sum_{l=1}^m \hat{\Sigma}_{ql} \gamma(t_{(q)}) \gamma(t_{(l)}) w_q w_l}{\|\gamma\|^2 + \alpha_1 [\gamma, \gamma]}, \tag{2.8}$$

where $\alpha_1 > 0$ is a smoothing parameter, $\hat{\Sigma}_{ql} = \frac{1}{n} \sum_{p=1}^n \left(Y_{pq} - \bar{Y}_{\cdot q}\right) \left(Y_{pl} - \bar{Y}_{\cdot l}\right)$, $1 \le q, l \le m$, $\bar{Y}_{\cdot q} = \frac{1}{n}(Y_{1q} + \cdots + Y_{nq})$, and

$$w_q = \begin{cases} \left(t_{(2)} - t_{(1)}\right)/2 & q = 1 \\ \left(t_{(q+1)} - t_{(q-1)}\right)/2 & 1 < q < m \\ \left(t_{(m)} - t_{(m-1)}\right)/2 & q = m. \end{cases} \tag{2.9}$$

The higher order estimates $\{\hat{\lambda}_k, \hat{\gamma}_k\}, k \ge 2$ are the solutions to the following optimization problems:

$$\max_{\substack{\|\gamma\| = 1, \langle \gamma, \hat{\gamma}_j \rangle = 0, \\ j = 1, \cdots, k-1}} \frac{\sum_{q=1}^m \sum_{l=1}^m \hat{\Sigma}_{ql} \gamma(t_{(q)}) \gamma(t_{(l)}) w_q w_l}{\|\gamma\|^2 + \alpha_k [\gamma, \gamma]}, \tag{2.10}$$

where α_k is the smoothing parameter for the k-th estimates and the estimates of eigenfunctions are orthogonal to each other. We can choose different smoothing parameters for different principal components.

The idea behind our method is as follows. The true eigenvalues and eigenfunctions are the solutions to the following successive optimization problems:

$$\max_{\substack{\|\gamma\| = 1, \langle \gamma, \gamma_j \rangle = 0, \\ j = 1, \cdots, k-1}} \frac{\langle \gamma, \Gamma\gamma \rangle}{\|\gamma\|^2}.$$

where $\Gamma\gamma$ is the function defined by $(\Gamma\gamma)(t) = \int_a^b \Gamma(t,s)\gamma(s)ds$. These optimization problems depend on the covariance function Γ only through the inner product $\langle \gamma, \Gamma\gamma \rangle$. Hence, we use the numerators in (4.95) and (2.10) to approximate $\langle \gamma, \Gamma\gamma \rangle$. However, if there are no penalty terms in the denominators in (4.95) and (2.10), the maximum values of (4.95) and (2.10) are infinities. Since tuning α_k does not affect the first $k-1$ estimates, we can tune

the parameters one by one. We give a theorem on the existence and characterization of solutions of the successive optimization problems (4.95) and (2.10). Our methods solve the optimization problems in the Sobolev space, hence many powerful tools from the theory of Hilbert space can be used to study the asymptotic consistency of our method.

We give a theorem on the existence and characterization of solutions of the successive optimization problems (4.95) and (2.10).

Theorem 2.3.1. *The solutions* $\{\hat{\lambda}_k, \hat{\gamma}_k : k \geq 1\}$ *of the successive optimization problems* (4.95) *and* (2.10) *exist for any* $\{\alpha_k > 0, k \geq 1\}$. *Moreover, for each* k, $\hat{\gamma}_k$ *has continuous second derivatives on* $[a, b]$ *and, on any subinterval* $\{[t_{(q-1)}, t_{(q)}], 1 \leq q \leq m-1\}$, *it can be written as a linear combination of the following at most* $4k$ *functions,*

$$\exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \qquad \exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)$$
$$\exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \quad \exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),$$

where $1 \leq j \leq k$.

Hence, the first solution $\hat{\gamma}_1$ is similar to smoothing splines except that the solutions to the optimization problems in smoothing spline methods are cubic polynomials between any two adjacent observation points.

2.3.2 Irregular case

In this case, we assume the observation time points are

$$\{t_{pq} : p = 1, \cdots, n, \quad q = 1, \cdots, N_p\}$$

, where n is the number of sample curves and N_p is the number of the observation points of the p-th sample function X_p. The model is

$$Y_{pq} = X_p(t_{pq}) + \epsilon_{pq}, \quad p = 1, \cdots, n, \quad q = 1, \cdots, N_p, \tag{2.11}$$

where Y_{pq} is the observation of the random function X_p at time t_{pq} and ϵ_{pq} is the measurement error.

For irregular case, we assume that the mean function $\mu(t)$ is smooth and the observation points t_{pq} are random variables with a density function $h(t)$ which is bounded below away from zero on $[a, b]$. Our FPCA procedure for irregular case has three steps.

In the first step, we estimate the mean function $\mu(t)$ based on the pooled data from all individuals by local linear smoother. This step is the same as the first step of the procedure in Yao et al. [54]. We define the estimate $\hat{\mu}(t)$ of $\mu(t)$ by solving the following optimization problem

$$\min_{a,b} \sum_{p=1}^{n} \sum_{q=1}^{N_p} \kappa \left(\frac{t_{pq} - t}{\eta_\mu} \right) \{Y_{pq} - a - b(t - t_{pq})\}^2 , \tag{2.12}$$

where κ is the kernel, and η_μ is the bandwidth. Let $\hat{a}(t)$ and $\hat{b}(t)$ be the minimizers, then $\hat{\mu}(t) = \hat{a}(t)$.

In the second step, we estimate the density function $h(t)$ based on pooled observation time points by the maximum penalized likelihood estimation method (see Silverman [40], Silverman [42] and Chapter 6 in Ramsay and Silverman [33]). Let $\hat{g}(t)$ be the minimizer of the functional

$$-\frac{1}{N} \sum_{p=1}^{n} \sum_{q=1}^{N_p} g(t_{pq}) + \int_a^b e^{g(t)} dt + \eta_g [g, g], \tag{2.13}$$

where $N = \sum_{p=1}^{n} N_p$ and η_g is a smoothing parameter, then the estimate $\hat{h}(t) = e^{\hat{g}(t)}$. Here we use the maximum penalized likelihood estimation method instead of the kernel density estimation method because the density estimate in this step will appear in the denominators in the third step. Hence, the density estimate must be positive. In the maximum penalized likelihood estimation, the log density is first estimated, then its exponential is calculated as the density estimate. Hence, the maximum penalized likelihood density estimate is strictly positive.

The third step is to solve the following successive optimization problems. The first pair of estimates $\{\hat{\lambda}_1, \hat{\gamma}_1\}$ of $\{\lambda_1, \gamma_1\}$ are the maximum value and the solution to the optimization

problem:

$$\max_{\substack{\gamma \in W_2^2([a,b]),\\ \|\gamma\|=1}} \frac{\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}U_{ql}^{(p)}}{\|\gamma\|^2+\alpha_1[\gamma,\gamma]}, \tag{2.14}$$

where $\alpha_1 > 0$ is a smoothing parameter, $\chi_{[N_p>1]}$ is the indicator function of $N_p > 1$, $n' = \sum_{p=1}^{n}\chi_{[N_p>1]}$ is the total number of the sample functions with at least two observation points and

$$U_{ql}^{(p)} = \frac{\gamma(t_{pq})(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{\gamma(t_{pl})(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})}.$$

The higher order estimates $\{\hat{\lambda}_k, \hat{\gamma}_k\}, k \geq 2$ are the solutions to the following optimization problems:

$$\max_{\substack{\|\gamma\|=1, \langle\gamma,\hat{\gamma}_j\rangle=0,\\ j=1,\cdots,k-1}} \frac{\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}U_{ql}^{(p)}}{\|\gamma\|^2+\alpha_k[\gamma,\gamma]}, \tag{2.15}$$

where α_k is the positive smoothing parameter for the k-th estimates.

Now we intuitively explain (2.14) and (2.15). For each $1 \leq p \leq n$, if $N_p > 1$, then

$$\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}U_{ql}^{(p)}$$

is an approximation to the U-statistic

$$\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{\gamma(t_{pq})(Y_{pq}-\mu(t_{pq}))}{h(t_{pq})}\cdot\frac{\gamma(t_{pl})(Y_{pl}-\mu(t_{pl}))}{h(t_{pl})}. \tag{2.16}$$

For different p's with $N_p > 1$, (2.16) are independently and identically distributed random variables if we assume that N_p is a random variable independent of t_{pq} and the random

function X. Therefore, by the law of large numbers, the numerators in (2.14) and (2.15) are approximation to $\langle \gamma, \Gamma\gamma \rangle$. We give a similar theorem as Theorem 2.3.1 for the irregular case.

Theorem 2.3.2. *The solutions* $\{(\hat{\lambda}_k, \hat{\gamma}_k) : k \geq 1\}$ *of the successive optimization problems* (2.14) *and* (2.15) *exist for any* $\{\alpha_k > 0, k \geq 1\}$. *Moreover, for each* k, $\hat{\gamma}_k$ *has continuous second derivatives on* $[a, b]$ *and on the subinterval between any two adjacent pooled observation points, it can be written as a linear combination of the following at most* $4k$ *functions,*

$$\exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \qquad \exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)$$
$$\exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \quad \exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),$$

where $1 \leq j \leq k$.

2.3.3 Computational issues

Although Theorems 2.3.1 and 2.3.2 give the forms of the solutions to the successive optimization problems in our FPCA procedure, it is not convenient to compute the exact solutions in practice. Instead, we choose an appropriate basis and use the basis expansions to approximate the solutions to the successive optimization problems as did [33] in Section 9.4. We develop similar algorithms for computing the solutions to the successive optimization problems in our method as those in Section 9.4 of [33]. We first choose an appropriate basis $\{\phi_\nu\}_{\nu=1}^M$, where M is the number of basis functions. For example, we can choose the Fourier series as our basis for the periodic case and the B-spline basis for the nonperiodic case.

Let $\tilde{\gamma}_k = \sum_{\nu=1}^M c_{k\nu}\phi_\nu$, $k \geq 1$, be the solutions to (4.2) or (2.15) restricted to the linear space spanned by the basis functions. They are the approximations to $\{\hat{\gamma}_k\}_{k\geq 1}$. The coefficients $\mathbf{c}_k = (c_{k1}, \cdots, c_{kM})^T$ are solutions to the following successive optimization problems,

$$\max_{\substack{\mathbf{c} \in \mathbb{R}^M, \mathbf{c}^T\mathbf{J}\mathbf{c} = 1 \\ \mathbf{c}_j^T\mathbf{J}\mathbf{c} = 0, j = 1, \cdots, k-1}} \frac{\mathbf{c}^T\mathbf{V}\mathbf{c}}{\mathbf{c}^T\mathbf{J}\mathbf{c} + \alpha_k\mathbf{c}^T\mathbf{K}\mathbf{c}}. \tag{2.17}$$

$\mathbf{J}$ and $\mathbf{K}$ are $M \times M$ matrices with elements $\mathbf{J}_{\nu\nu'} = \int_a^b \phi_\nu(t)\phi_{\nu'}(t)dt$ and $\mathbf{K}_{\nu\nu'} = \int_a^b \phi''_\nu(t)\phi''_{\nu'}(t)dt$, $\nu, \nu' = 1, \cdots, M$, where ϕ''_ν is the second derivative of ϕ_ν. $\mathbf{V}$ is a $M \times M$ matrix with elements

$$\mathbf{V}_{\nu\nu'} = \sum_{q=1}^{m}\sum_{l=1}^{m} \hat{\Sigma}_{ql}\phi_\nu(t_{(q)})\phi_{\nu'}(t_{(l)})w_q w_l, \tag{2.18}$$

in regular case and

$$\mathbf{V}_{\nu\nu'} = \frac{1}{n'}\sum_{p=1}^{n} \frac{\chi_{[N_p>1]}}{N_p(N_p-1)} \sum_{l\neq q:1}^{N_p} \frac{\phi_\nu(t_{pq})(Y_{pq} - \hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})} \cdot \frac{\phi_{\nu'}(t_{pl})(Y_{pl} - \hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})}, \tag{2.19}$$

in irregular case.

The algorithm for solving (2.17) is as follows:

- Perform a Cholesky factorization $\mathbf{L}_1^T\mathbf{L}_1 = \mathbf{J} + \alpha_1\mathbf{K}$ and calculate the inverse matrix $\mathbf{L}_1^{-1}$ of $\mathbf{L}_1$.
- Let $\mathbf{B}_1 = (\mathbf{L}_1^{-1})^T\mathbf{V}\mathbf{L}_1^{-1}$ and compute the first eigenvector $\mathbf{d}_1$ of $\mathbf{B}_1$. Then $\mathbf{c}_1 = \frac{\mathbf{L}_1^{-1}\mathbf{d}_1}{r_1}$, where r_1 is a real number chosen such that $\mathbf{c}_1^T\mathbf{J}\mathbf{c}_1 = 1$.
- For $k > 1$, suppose that we have obtained $\mathbf{c}_1 \cdots, \mathbf{c}_{k-1}$. Perform the Cholesky factorization $\mathbf{L}_k^T\mathbf{L}_k = \mathbf{J} + \alpha_k\mathbf{K}$ and calculate the inverse matrix $\mathbf{L}_k^{-1}$ of $\mathbf{L}_k$.
- Let $\mathbf{C}_{k-1} = [\mathbf{c}_1 \cdots, \mathbf{c}_{k-1}]$, that is, $\mathbf{C}_{k-1}$ is an $M \times (k-1)$ matrix with the j-th column equal to $\mathbf{c}_j$.
- Perform a QR-decomposition $\mathbf{Q}_k\mathbf{R}_k = (\mathbf{L}_k^{-1})^T\mathbf{J}\mathbf{C}_{k-1}$, where $\mathbf{Q}_k$ is a $M \times (k-1)$ matrix with columns have norm 1 and orthogonal to each other and $\mathbf{R}_k$ is an upper triangular matrix.
- Calculate the projection matrix $\mathbf{P}_k = \mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^T$ onto the linear space orthogonal to the linear space spanned by the columns of $(\mathbf{L}_k^{-1})^T\mathbf{J}\mathbf{C}_{k-1}$, where $\mathbf{I}$ is the identity matrix of M dimension.
- Let $\mathbf{B}_k = \mathbf{P}_k(\mathbf{L}_k^{-1})^T\mathbf{V}\mathbf{L}_k^{-1}\mathbf{P}_k$ and compute the first eigenvector $\mathbf{d}_k$ of $\mathbf{B}_k$. Then $\mathbf{c}_k = \frac{\mathbf{L}_k^{-1}\mathbf{d}_k}{r_k}$, where r_k is a real number chosen such that $\mathbf{c}_k^T\mathbf{J}\mathbf{c}_k = 1$.

2.3.4 Consistency

We assume throughout this section that we want to estimates the first K principal component curves, where K is any fixed positive integer number.

First, we consider the regular model (2.7). For this model, we consider the following two cases for the distributions of t_q:

Case 1 (Nonrandom Case)**.** *$\{t_q, 1 \le q \le m\}$ are nonrandom. Define*

$$\delta_m = \max_{2 \le q \le m} \left(t_{(q)} - t_{(q-1)} \right). \tag{2.20}$$

Case 2 (Random Case)**.** *$\{t_q, 1 \le q \le m\}$ are i.i.d. random variables having a density functions $h(t)$ in $[a, b]$ with respect to Lebesgue measure and are independent of the random functions $X_p, 1 \le p \le n$. Furthermore, $h(t)$ has a positive lower bound c.*

In order to give the consistency result for the regular model (2.7), we need the following two more assumptions:

Assumption 4. *The measurement errors $\epsilon_{pq}, 1 \le p \le n, 1 \le q \le m$ are independent random variables and are independent of the random functions $X_p, 1 \le p \le n$ and the observation times $t_q, 1 \le q \le m$. For each q, $\{\epsilon_{1q}, \cdots, \epsilon_{nq}\}$ have the same distribution with mean 0 and variance σ_q^2. Furthermore,*

$$\sup_q \sigma_q^2 \le \sigma^2, \quad \sup_{q,l} E|\epsilon_{ql}|^3 \le \rho,$$

where σ and ρ are some positive numbers and do not depend on m.

Remark 2.3.3. *We do not assume that all the measurement errors have the same distributions. Instead we only assume that the errors arising at the same observation time have the same distribution, which is more general than the former.*

Assumption 5. *The covariance function $\Gamma(s, t)$ is a continuous function in $[a, b] \times [a, b]$.*

Define a function

$$\varpi(\delta) = \sup_{s,t \in [a,b], |s-t| \le \delta} \left[\Gamma(t, t) - 2\Gamma(s, t) + \Gamma(s, s) \right], \tag{2.21}$$

where $0 < \delta \leq b - a$. Note that

$$\Gamma(t,t) - 2\Gamma(s,t) + \Gamma(s,s) = E\left[((X(s) - \mu(s)) - (X(t) - \mu(t)))^2\right].$$

Under Assumption 5, we have $\lim_{\delta \to 0} \varpi(\delta) = 0$ and Γ is bounded. If G is smooth, then $\varpi(\delta) = O(\delta)$. Although the covariance functions of Brownian motion and Poisson process with rate 1 are not smooth, for both of them, we have

$$E\left[((X(s) - \mu(s)) - (X(t) - \mu(t)))^2\right] = |t - s|,$$

and therefore, $\varpi(\delta) = \delta$

Theorem 2.3.4. *Under Assumptions* $1 - 5$*, suppose that* $m, n \to \infty$*,* $\max_{1 \leq k \leq K} \alpha_k \to 0$ *and*

$$\frac{\max_{1 \leq k \leq K} \alpha_k}{\min_{1 \leq k \leq K} \alpha_k} = O_p(1). \tag{2.22}$$

If the following is satisfied that for Case 1*,*

$$\frac{1}{\min_{1 \leq k \leq K} \alpha_k} \left[\sqrt{\varpi(\delta_m)} + \delta_m + \sqrt{\frac{\delta_m}{n}}\right] \to 0$$

and for Case 2*,*

$$\frac{1}{\min_{1 \leq k \leq K} \alpha_k} \left[\sqrt{\varpi(\frac{3 \log m}{cm})} + \frac{\log m}{m} + \sqrt{\frac{\log m}{nm}}\right] \to 0,$$

then the estimators $\{(\hat{\lambda}_k, \hat{\gamma}_k) : 1 \leq k \leq K\}$ *are consistent.*

Second, we consider the irregular model (4.130). For this model, we make the following assumptions on the number of observation points, measurement errors, mean functions and density functions. They are actually parts of assumptions in Yao et al. [54] and Hall et al. [20].

Assumption 6. *The numbers of the observation points* N_p*,* $1 \leq p \leq n$*, are i.i.d random variables taking positive integer values with* $EN_p < \infty$ *and* $P(N_p > 1) > 0$*. The measurement*

errors $\epsilon_{pq}, 1 \leq p \leq n, 1 \leq q \leq m$ *are i.i.d random variables with mean zero and finite variance. The random functions, the observation points, the number of the observation points and the measurement errors are independent.*

Assumption 7. *Both the mean function and the density function have square integrable second derivatives, that is,* $\mu(t), h(t) \in W_2^2([a,b])$*. The kernel* κ *in* (4.91) *is compactly supported, symmetric and Hölder continuous. The smoothing parameter* η_μ *in* (4.91) *satisfies*

$$n^{\rho_1 - \frac{1}{2}} \leq \eta_\mu = o(1), \quad \eta_\mu = o(n^{-\frac{1}{4}}),$$

where $\rho_1 > 0$ *is some constant. There are two positive constants* $c < C$ *such that* $c \leq h(t) \leq C$ $\forall a \leq t \leq b$ *and the smoothing parameter* η_g *satisfies*

$$\eta_g \to 0, \quad n^{1-\rho_2}\eta_g \to \infty,$$

where $\rho_2 > 0$ *is some constant.*

Now we present the consistency result for the irregular model.

Theorem 2.3.5. *Under Assumptions* $1-3$ *and* $6-7$*, suppose that* $n \to \infty$*,* $\max_{1\leq k\leq K} \alpha_k \to 0$ *and*

$$\frac{\max_{1\leq k\leq K} \alpha_k}{\min_{1\leq k\leq K} \alpha_k} = O_p(1).$$

If the following is satisfied

$$\frac{1}{\min_{1\leq k\leq K} \alpha_k}\left[n^{-\frac{1}{2}}(\eta^{-1} + \eta_g^{-\frac{1}{2}-\epsilon}) + \eta_g^{\frac{3}{4}-\epsilon}\right] \to 0,$$

for some $\epsilon > 0$*, then the estimators* $\{(\hat{\lambda}_k, \hat{\gamma}_k) : 1 \leq k \leq K\}$ *are consistent.*

2.3.5 Extensions to FPCA for functional data with multidimensional arguments

Functional data with multidimensional arguments are collected in a growing number of fields. For example, in spatial data analysis, data are collected from different places

and at different times. Such data can be view as discretely observed functional data which are functions of both space and time. Analysis of such data is considered an important direction of functional data analysis (see Section 22.2 in Ramsay and Silverman [33] and Section 5 in Müller [28]). Our method can be easily extended in this context by defining similar successive optimization problems in multidimensional spaces. The numerators in the successive optimization problems (2.14) and (2.15) can be straightforwardly extended to the multidimensional case and the penalty terms in the denominators can be replaced with

$$J_m^d(\gamma) = \sum_{j_1+\cdots+j_d=m} \frac{m!}{j_1!\cdots j_d!} \int_\Omega \left(\frac{\partial^m \gamma}{\partial t_1^{j_1} \cdots \partial t_j^{j_d}} \right)^2 dt_1 \cdots dt_d$$

where d is the dimension of the space of the arguments, $\Omega \in \mathbb{R}^d$ is the region in which the function is defined and we assume that the eigenfunctions have square integrable m-th derivatives. Our method avoids the estimates of covariance functions which have $2d$ arguments and are not very easy to estimate when $d \geq 2$.

2.4 Simulation studies

To illustrate the performance of our method, we conduct three simulation studies. In the first study, the sample curves are smooth with both equally and unequally spaced observation time points, and we will compare our method with an alternative method (Method II) which first obtains the smooth estimate of mean curve and covariance functions, and then compute the eigenfunctions of the smoothed covariance function as the estimations of the PC curves. We use the software package PACE for the second method, which was developed by Yao *et al*, and downloaded the software from http://www.stat.ucdavis.edu/PACE/download. In the second study, the sample curves are simulated with 3 true principle curves and we will compare our mthod with Method II. In the third study, we simulate random surfaces and perform FPCA in a two-dimensional space with our method.

2.4.1 Smooth random curves with 2 PC curves

We first simulate 200 curves from the following random curve on $[0, 1]$,

$$X(t) = \sqrt{2}\sin(2\pi U)\sin(\frac{\pi t}{2}) + \cos(2\pi U)\sin(\frac{3\pi t}{2}),$$

where U is a random variable with uniform distribution on $[0, 1]$. The covariance function of $X(t)$ has two nonzero eigenvalues and the corresponding eigenfunctions are $\sqrt{2}\sin(\frac{\pi t}{2})$ and $\sqrt{2}\sin(\frac{3\pi t}{2})$. Figure 2.1 shows the plot of the first two principal component curves.

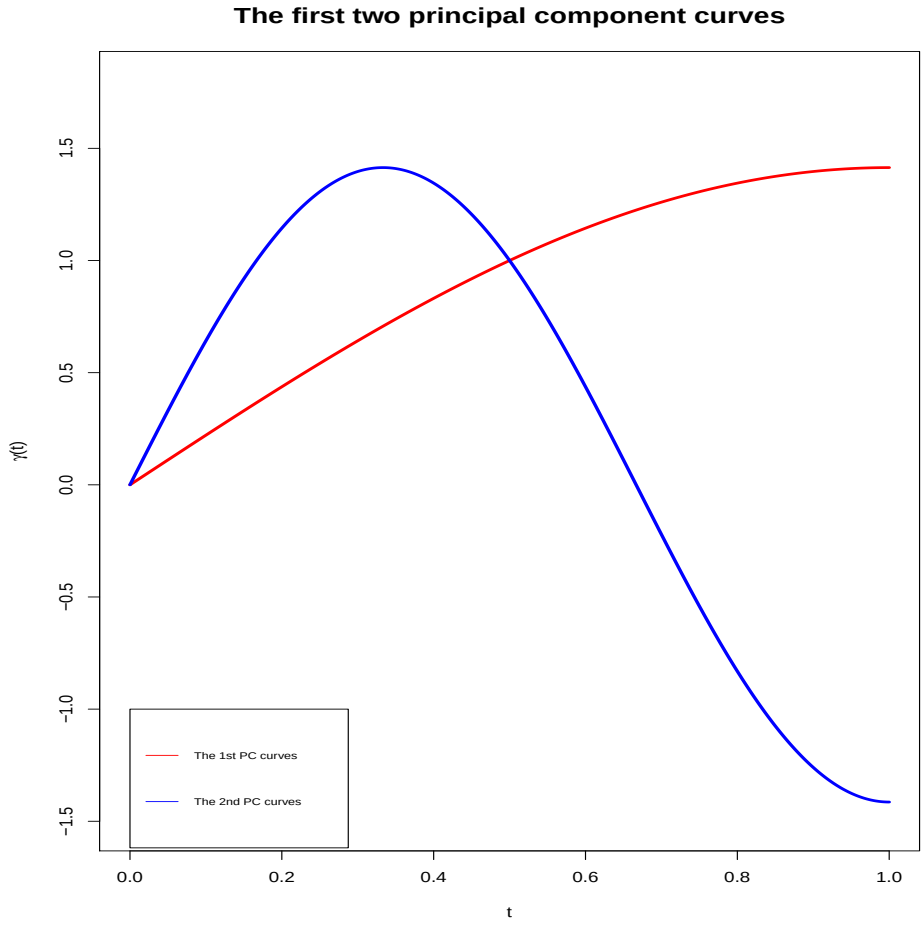

Figure 2.1 The First Two Principal Component Curves

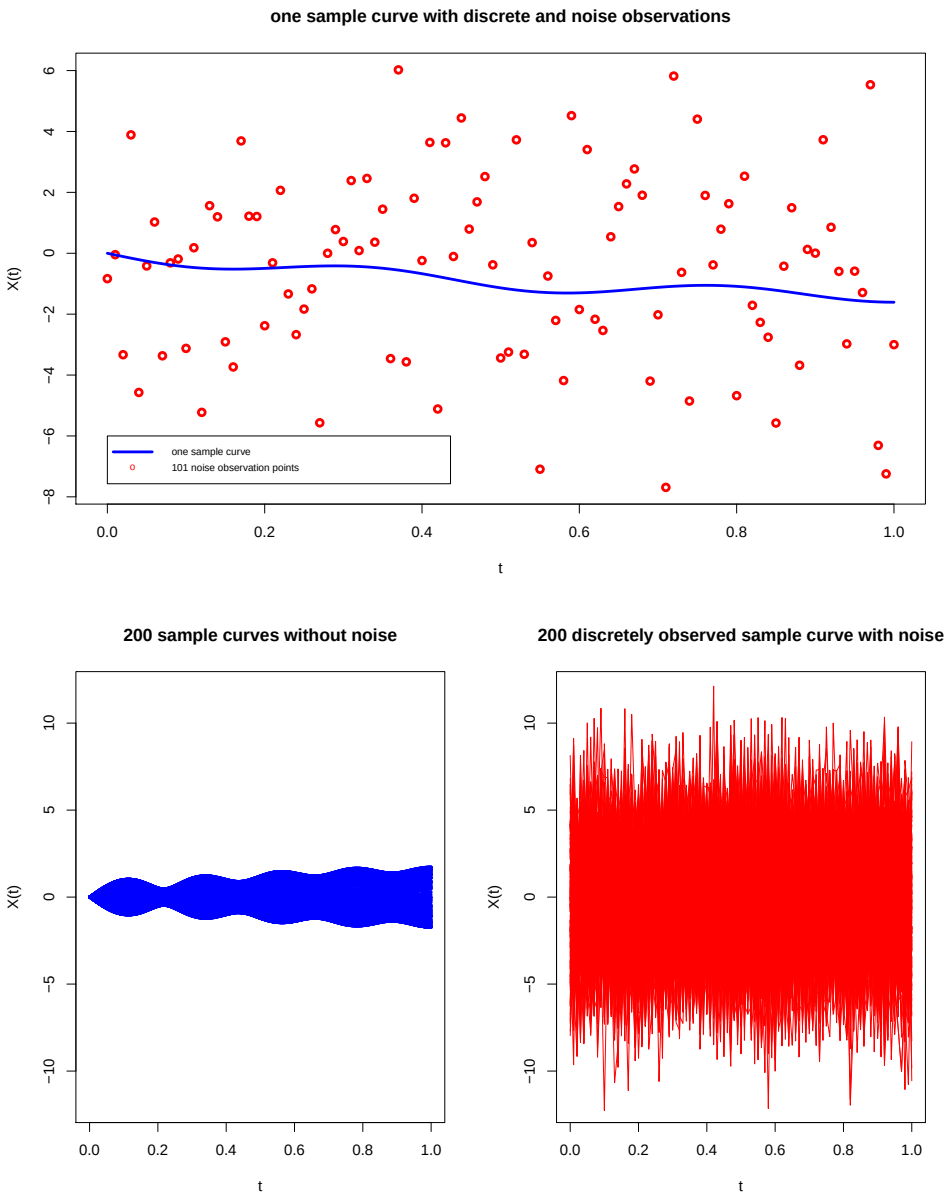

Figure 2.2 Simulated Sample Curves for the simulations in Section 2.4.1

For discrete observations, we will consider two cases.

Regular case The observed data are generated from the following model

$$Y_{pq} = X_p(t_q) + \epsilon_{pq}, \quad t_q = \frac{q-1}{100}, \quad q = 1, \cdots n, \quad p = 1, \cdots, 200.$$

where the measurement errors $\epsilon_{pq} \sim N(0, 3)$, the observation points $\{t_{pq}\}$ are equally spaced on $[0, 1]$. We consider $n = 101, 51$ and 21, that is, we sample different number of observation points: 101 equally spaced with measurement errors; 51 equally spaced with measurement errors; 21 equally spaced with measurement errors. Plots of the simulated sample curves are shown in 2.2.

We will estimate the first two eigenfunctions by our method and Method II respectively. We conducted 200 simulations and in each simulation, 200 observations are generated as training sample and 500 observations as test sample. For our method, use the usual cross-validation procedure to select the smoothing parameter α from $\{1e-10, 1e-09, 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02, 1e-01\}$, such that the total variance accounted for by all the principal components on the test data is maximized. We obtain the smoothing parameter $\alpha = 1e-04$. The parameters for Method II can be chosen by generalized cross-validation (GCV) method. Table 2.1 lists the cumulative variance of selected principal component scores. Under different settings, the first two estimated principle component curves obtained by our method explain larger total variation in the data.

Table 2.1 The averages and standard deviations of cumulative variance of selected principal component scores for the simulations in Section 2.4.1: Regular Case. For each sampling strategy shown in column 1, the first row is the average and standard deviation of the first estimated PC score variance; the second row is for the second esitmated PC score variance.

	Our Method		**Method II (PACE)**	
Selected PCs	Var.PC.Score (Mean)	Var.PC.Score (Variance)	Var.PC.Score (Mean)	Var.PC.Score (Variance)
101 Equally spaced				
1st PC	0.005264516	0.0002149985	0.0053	0.0035
2nd PC	0.008040198	0.0001687864	0.0056	0.0035
51 Equally spaced				
1st PC	0.01088641	0.0004329838	0.0108	0.0075
2nd PC	0.01695650	0.0004169770	0.0119	0.0075
21 Equally spaced				
1st PC	0.03002970	0.001443941	0.0322	0.0164
2nd PC	0.04807245	0.001607120	0.0391	0.0164

Irregular case In this case, the observed data are generated from the following model

$$Y_{pq} = X_p(t_q) + \epsilon_{pq}, \quad q = 1, \cdots n, \quad p = 1, \cdots, 200.$$

where the measurement errors $\epsilon_{pq} \sim N(0, 3)$, the observation points $\{t_{pq}\}$ are i.i.d random variables from Uniform$[0, 1]$. That is, we sample 200 curves and make different number of observation points: 101 unequally spaced with measurement errors; 51 unequally spaced with measurement errors; 21 unequally spaced with measurement errors.

We use the principal component curves estimated from both methods to approximate the true principal component curves. We conducted 200 simulations. Similarly to the regular case, we use the usual cross-validation procedure to obtain the parameters for our method such that the total variance accounted for by all the principal components on the test data is maximized. We obtain the smoothing parameter $\alpha = 1e - 04$ and the cumulative variance of selected PC scores are listed in Table 2.2. Under different settings, the estimated PC curves obtained by our method explain larger total variation in the data.

Table 2.2 The averages and standard deviations of cumulative variance of selected PC scores for the simulations in Section 2.4.1: Irregular Case. For each sampling strategy shown in column 1, the first row is the average and standard deviation of the first estimated PC score variance; the second row is for the second esitmated PC score variance.

	Our Method		**Method II (PACE)**	
Selected PCs	Var.PC.Score (Avg.)	Var.PC.Score (Std.)	Var.PC.Score (Avg.)	Var.PC.Score (Std.)
101 Unequally spaced				
1st PC	0.005266157	0.0001979994	0.0053	0.0037
2nd PC	0.007900791	0.0001649976	0.0056	0.0037
51 Unequally spaced				
1st PC	0.01037529	0.0004865600	0.0112	0.0061
2nd PC	0.01672025	0.0004334033	0.0123	0.0061
21 Unequally spaced				
1st PC	0.02997724	0.001536284	0.0288	0.0171
2nd PC	0.04821568	0.001759098	0.0359	0.0170

2.4.2 Smooth random curves with 3 PC curves

We consider the following random curve on $[0, 1]$,

$$X(t) = 3\beta_1 \sin(\frac{\pi t}{2}) + 2\beta_2 \sin(\frac{3\pi t}{2}) + \beta_3 \sin(\frac{5\pi t}{2}),$$

where $\beta_i, i = 1, 2, 3$ are random variables with normal distribution on $[0, 1]$. The covariance function of $X(t)$ has three nonzero eigenvalues and the corresponding eigenfunctions are $\sqrt{2}\sin(\frac{\pi t}{2})$, $\sqrt{2}\sin(\frac{3\pi t}{2})$ and $\sqrt{2}\sin(\frac{5\pi t}{2})$. Figure 2.3 shows the plot of the first three principal component curves.

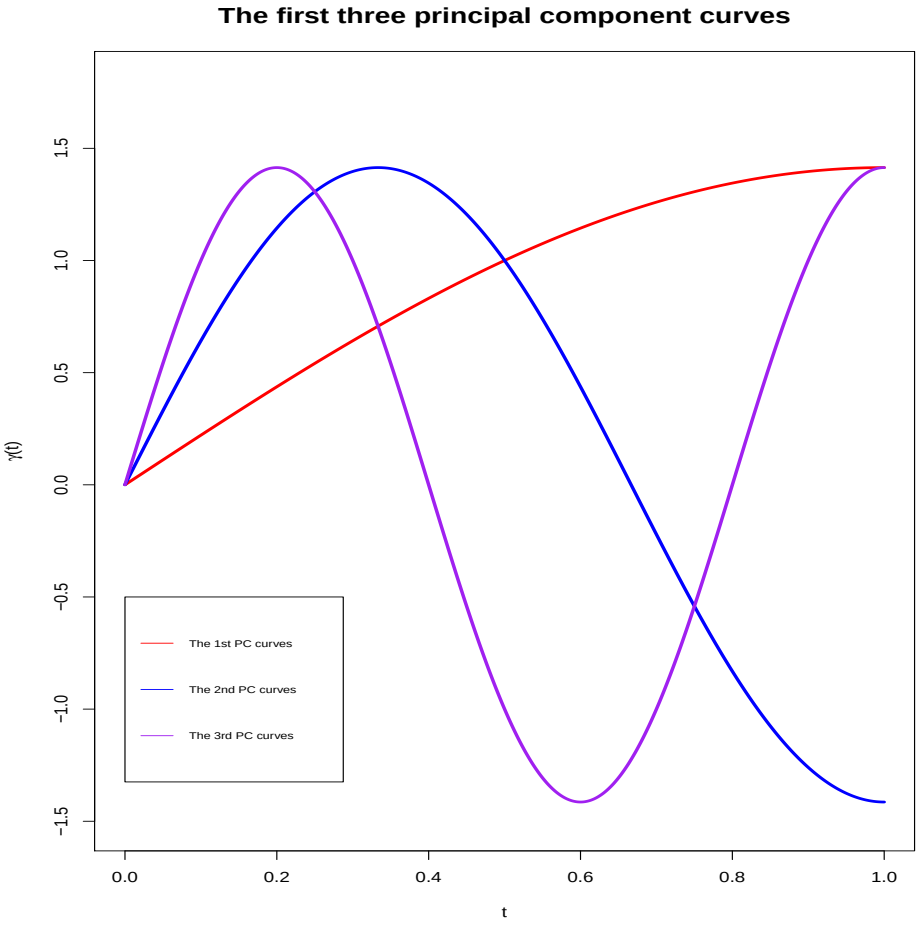

Figure 2.3 The First Three Principal Component Curves

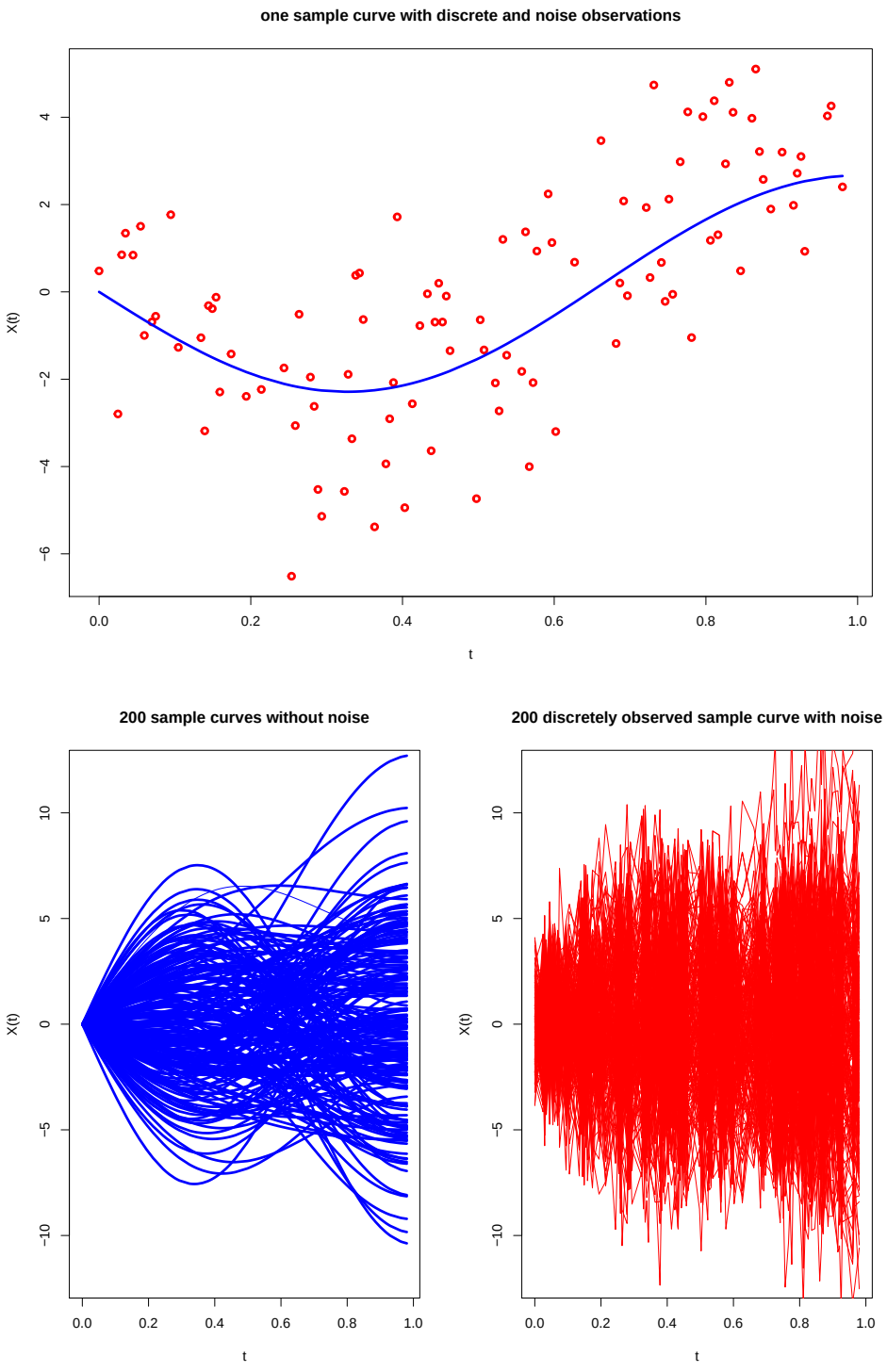

Figure 2.4 Simulated Sample Curves for the simulations in Section 2.4.2

Similar to the previous simulation, we will consider two cases.

Regular case The observed data are generated from the following model

$$Y_{pq} = X_p(t_q) + \epsilon_{pq}, \quad t_q = \frac{q-1}{100}, \quad q = 1, \cdots n, \quad p = 1, \cdots, 200.$$

where the measurement errors $\epsilon_{pq} \sim N(0, 3)$, the observation points $\{t_{pq}\}$ are equally spaced on $[0, 1]$. Plots of the simulated sample curves from one simulation are shown in 2.4.

We conducted 300 simulations with different number of observation points sampled: 101 equally spaced measurements; 51 equally spaced measurements; 21 equally spaced measurements. In each simulation,200 observations are generated as the training sample and 500 observations as the test sample. For our method, we use the usual cross-validation to obtain the smoothing parameter from $\{1e-10, 1e-09, 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02, 1e-01\}$, such that the total variance accounted for by all the principal components on the test data is maximized. The smoothing parameter chosen under different sampling strategies are listed in 2.3. We use the principal component curves estimated from both methods to approximate the true principal component curves. The cumulative variance of selected PC scores are listed in Table 2.4. We can see that the first three estimated principle component curves obtained by our method acccount for larger variation in the data.

Table 2.3 Selected smoothing parameter with the usual cross-validation procedure for the simulations in Section 2.4.2: Regular Case

Smoothing Parameter	101 Equally Spaced	51 Equally Spaced	21 Equally Spaced
α	1e-09	1e-06	1e-04

Irregular case In this case, the observated data are generated from the following model

$$Y_{pq} = X_p(t_q) + \epsilon_{pq}, \quad q = 1, \cdots n, \quad p = 1, \cdots, 200.$$

where the measurement errors $\epsilon_{pq} \sim N(0, 3)$, the observation points $\{t_{pq}\}$ are i.i.d random variables from Uniform$[0, 1]$. We use the principal component curves estimated from both

Table 2.4 The averages and standard deviations of cumulative variance of selected PC scores for the simulations in Section 2.4.2: Regular Case. For each sampling strategy shown in column 1, the first row is the average and standard deviation of the first estimated PC score variance; the second row is for the second esitmated PC score variance.

	Our Method		**Method II (PACE)**	
Selected PCs	Var.PC.Score (Avg.)	Var.PC.Score (Std.)	Var.PC.Score (Avg.)	Var.PC.Score (Std.)
101 Equally spaced				
1st PC	0.04248567	0.003085132	0.0452	0.0030
2nd PC	0.06662123	0.003577294	0.0456	0.0030
51 Equally spaced				
1st PC	0.08836079	0.005708103	0.0894	0.0058
2nd PC	0.13893059	0.006298565	0.0909	0.0058
21 Equally spaced				
1st PC	0.2178350	0.01486907	0.2221	0.0152
2nd PC	0.3435396	0.01646455	0.2313	0.0153
3rd PC	0.3506672	0.01646290	0.2344	0.0157

methods to approximate the true principal component curves with our method and Method II. Similarly to the regular case, we conducted 300 simulations with different number of observation points sampled: 101 unequally spaced measurements; 51 unequally spaced measurements; 21 unequally spaced measurements. We use the usual cross-validation procedure to obtain the parameters from for our method, such that the total variance accounted for by all the principal components on the test data is maximized. The smoothing parameter chosen under different sampling strategies are listed in 2.5 and the cumulative variance of selected PC scores are listed in Table 2.6.

Table 2.5 Selected smoothing parameter with the usual cross-validation procedure for the simulations in Section 2.4.2: Irregular Case

	101 Unqually Spaced	51 Unqually Spaced	21 Unqually Spaced
α	1e-06	1e-04	1e-03

Table 2.6 The averages and standard deviations of cumulative variance of selected PC scores for the simulations in Section 2.4.2: Irregular Case. For each sampling strategy shown in column 1, the first row is the average and standard deviation of the first estimated PC score variance; the second row is for the second esitmated PC score variance.

	Our Method		**Method II (PACE)**	
Selected PCs	Var.PC.Score (Avg.)	Var.PC.Score (Std.)	Var.PC.Score (Avg.)	Var.PC.Score (Std.)
101 Unequally spaced				
1st PC	0.04388102	0.002888812	0.0436	0.0029
2nd PC	0.06803005	0.003254693	0.0440	0.0029
3rd PC	0.06900475	0.003255002	0.0440	0.0029
51 Unequally spaced				
1st PC	0.08348801	0.005615962	0.0757	0.0051
2nd PC	0.12794842	0.006420639	0.0772	0.0051
3rd PC	0.12948339	0.006429075	0.0772	0.0051
21 Unequally spaced				
1st PC	0.2062827	0.01513302	0.2206	0.0140
2nd PC	0.3182653	0.01546205	0.2304	0.0140
3rd PC	0.3275431	0.01565964		

2.4.3 Random surface

We sample 200 surfaces from the distribution of

$$X(s,t) = 1 + e^s \cos(t) + (s-1)^2 t + \xi_1 \sin\left(\frac{\pi(s-t)}{2}\right) + \xi_2 \sin\left(\frac{\pi(s+t)}{2}\right),$$

where $0 \le s,t \le 1$, $\xi_1 \sim N(0,2)$, $\xi_2 \sim N(0,1)$. The summation of the first three terms is the mean function. The covariance function of $X(s,t)$ has two nonzero eigenvalues with eigenfunctions

$$\frac{1}{\sqrt{2}}(1-\frac{2}{\pi})\sin\left(\frac{\pi(s-t)}{2}\right), \quad \frac{1}{\sqrt{2}}(1+\frac{2}{\pi})\sin\left(\frac{\pi(s+t)}{2}\right), \quad 0 \le s,t \le 1.$$

For each sampled surface, we make 10 to 30 observations (irregular case) from a distribution with a truncated bivariate normal density with mean $(0.4, 0.6)$ and covariance matrix I restricted to the region $[0,1] \times [0,1]$. The number of the observations for each surface is a random variable with discrete uniform distribution on $\{10, 11, \cdots, 30\}$. The

measurement error $\epsilon \sim N(0, 0.2^2)$. The eigenfunctions and their estimates in one simulation are plotted in Figure 2.5. The smallest MISE of our method are 0.022 and 0.026 for the first two eigenfunctions respectively.

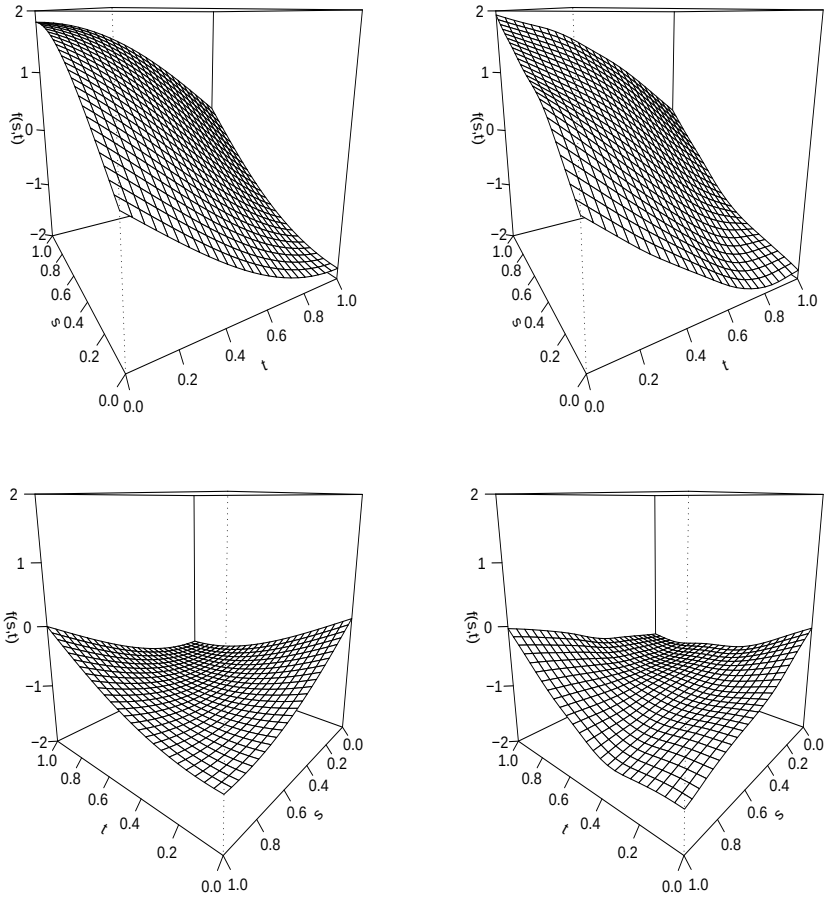

Figure 2.5 Eigenfunctions and their estimates in one simulation: The top left is the first true eigenfunction; the top right is the estimate of the first eigenfunction; the bottom left is the second eigenfunction; the bottom right is the estimate of the second eigenfunction.

2.5 Applications

2.5.1 Retinal pigment epithelium (RPE) data

The retinal pigment epithelium (RPE) is the pigmented cell layer between the choroid and the photoreceptor cell layer of eye. RPE is essential for visual function (see Strauss Strauss [45]). It provides multiple functions that support normal photoreceptor function, such as shielding the retina from excess incoming light, transporting water, nutrients and metabolic end products between the subretinal space and the blood, as well as secreting a variety of growth factors and signaling molecules (Zinn and Marmor Zinn and Marmor [55]). RPE is a key site of pathogenesis of age-related macular degeneration (AMD) which is a main source of vision loss even blindness in the elderly (Spaide and *et al.* Spaide and et al. [43]). The data is the collection of images of RPE cells of 88 mouse eyes provided in Emory Eye Center's L. F. Montgomery Lab at Emory University (Jiang and *et al.* Jiang and et al. [24]). The purpose of the study is to examine the relationship between the morphology of RPE layer and the age and disease status of the eye. Specifically, it is desirable to construct a classification rule based on the data so that the morphology of RPE of the eyes with different genotypes and in different age groups can be separated. There are two genotypes: wild and mutated, and two age groups: young (age$\leq$ 60 days) and elderly (age $>$ 60) groups in the data. Hence, we have four classes (that is, four combinations of genotypes and age groups). In each image, there are several thousands of cells. Several characteristics of each cell were measured including area, perimeter, aspect ratio, and so on. Local regions of two images with different genotypes, but having the same age equal to 60 days, are shown in Figure 2.6 (Chrenek and *et al.* Chrenek and et al. [10]). It can be seen that the distributions of the area and the shape of cells in the two images are quite different. Hence, we use the distributions of the area and the aspect ratio (a measure of shapes) of cells as classifiers respectively. The density curves of the area and the aspect ratio for each eye are estimated using the penalized likelihood method (see Section 5.4.3 in Ramsay, Hooker and Graves Ramsay et al. [34]), respectively, and the principal component scores are calculated and used to construct the classification rules. The eyes in different age groups can be separated using the distribution of the area of cells which cannot distinguish the eyes with different genotypes in the same

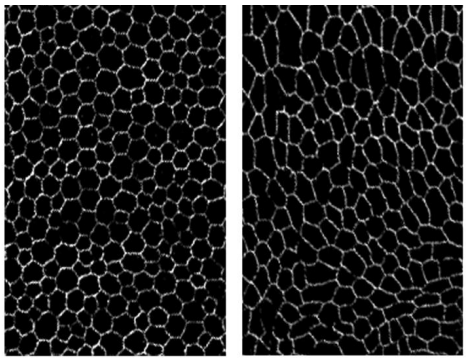

Figure 2.6 Local regions of two images with different genotypes, but same age equal to 60 days: (Left) RPE cells of the wild type and age 60 days; (Right) RPE cells of the mutated type and age 60 days.

age groups. Conversely, the distribution of the aspect ratio of cells can separate the eyes with different genotypes, but cannot distinguish those with the same genotypes in different age groups. Hence, we will combine the information of the area and the aspect ratio of cells together and apply our method to the joint density functions.

The data contains 88 images of mouse eyes. 27 are in the young age group with the wild genotype, 13 are in the elderly age group with the wild genotype, 27 are in the young age group with the mutated genotype and 21 are in the elderly age group with the mutated genotype. We first estimate the joint density function of the area and the aspect ratio of cells in each image using the kernel method (see Section 5.6 in Venables and Ripley [49]). The values of the density functions are calculated on a grid of 731×21 equally spaced points in the two-dimensional space (the area of cells are distributed between 0 and 730 μm^2 and the aspect ratio between 0 and 1). The mean joint densities of the four categories are plotted in Figure 2.7 which indicates the joint density curve is a good classifier of the genotype and the age group. We apply our method to the 88 joint density functions. Most of variations in the data are accounted for by the first four principal components which are plotted in Figure 2.8. Then we calculate four PC scores for each eye image, hence all the PC scores form a 88×4 matrix which is used to construct classification rules. We apply three classification methods, LDA (linear discriminant analysis), QDA (quadratic discriminant analysis) and SVM (support vector machine), to the matrix. Leave-one-out cross validation is used to

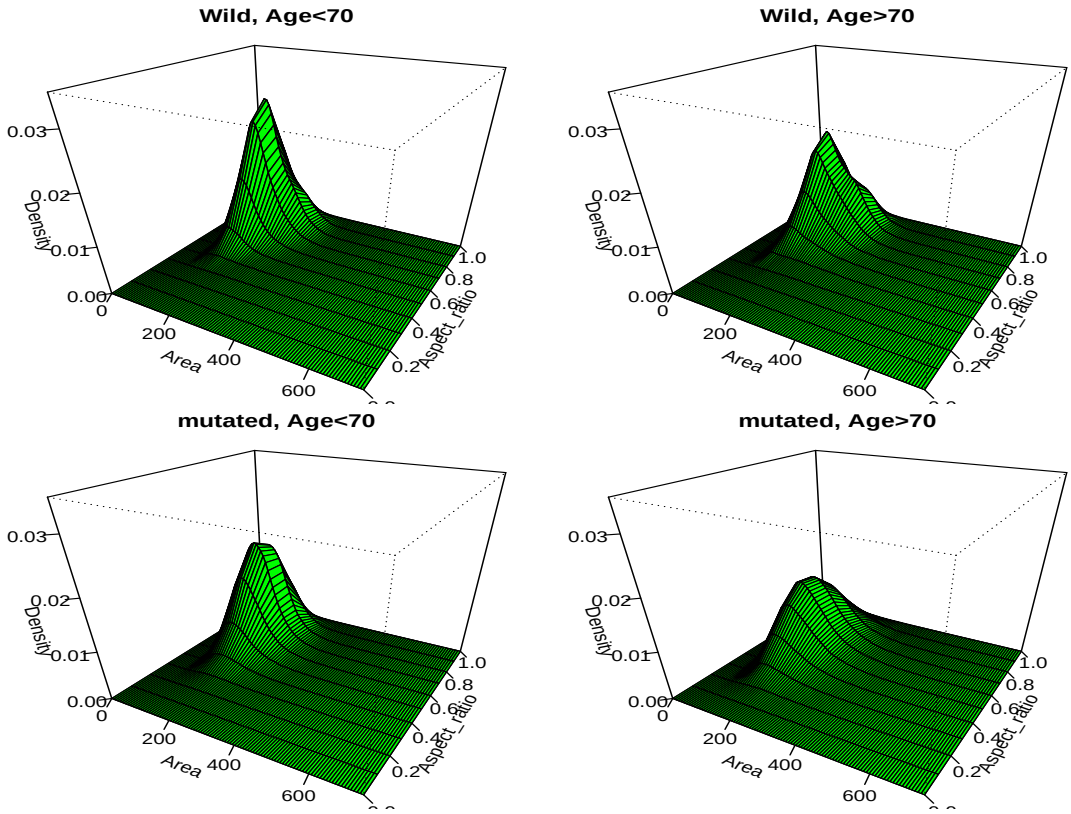

Figure 2.7 Mean joint densities of four categories.

assess the predictive accuracy. The PC scores of one eye image is selected as the test data and the PC scores of the remaining eye images are used as training data to construct the classification rule which is applied to the test data. This is repeated such that each eye image is used once as the test data. The predictive accuracy are 96.6% (85 are correctly classified among 88 eyes), 95.5%(84) and 95.5% (84) for LDA, QDA and SVM, respectively.

2.5.2 Longitudinal CD4 counts data

This dataset is from the Multicenter AIDS Cohort Study, which includes repeated measurements of physical exams, laboratory results, and CD4 percentages for 283 homosexual men who became HIV-positive between 1984 and 1991. The CD4 cell level is one of the important biomarkers to evaluate the disease progression of HIV infected subjects. All individuals were scheduled to have their measurements made at semiannual visits. However,

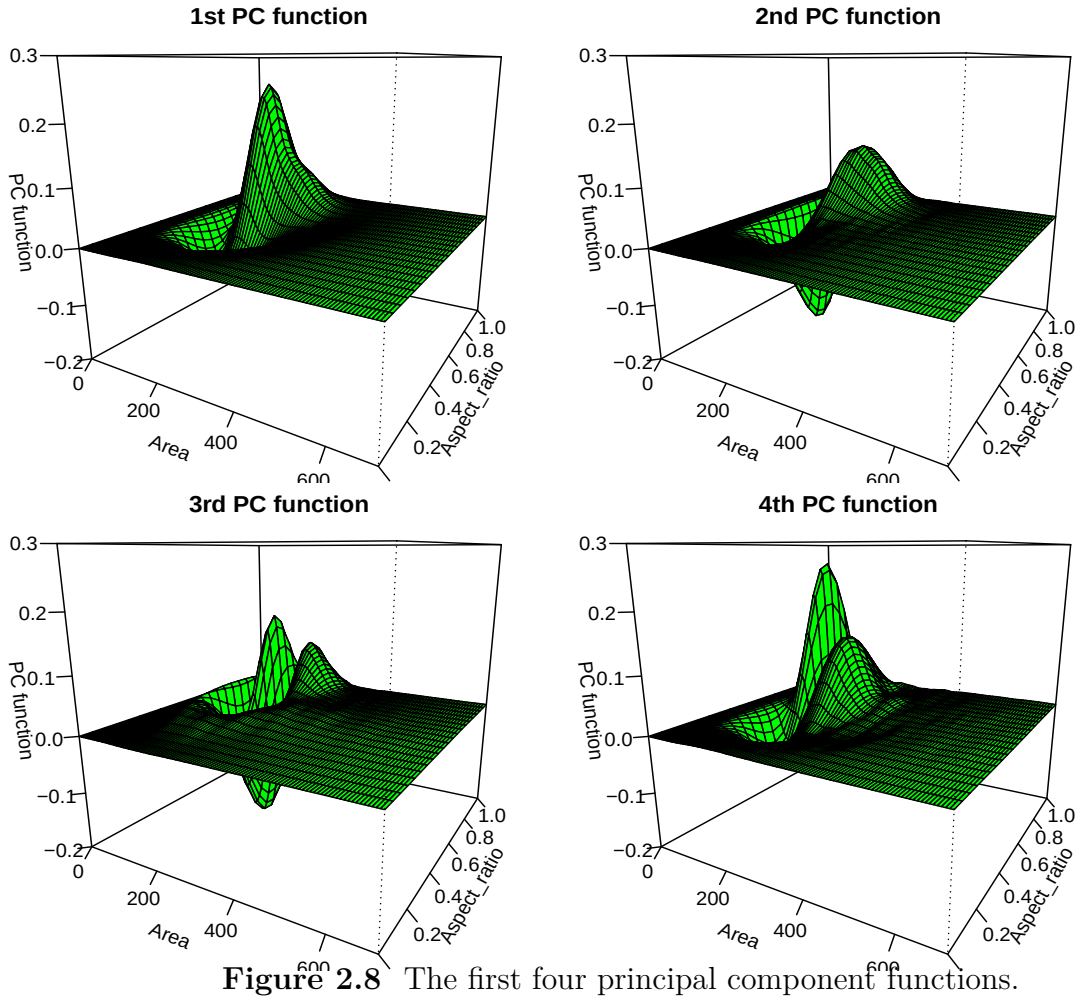

Figure 2.8 The first four principal component functions.

because many individuals missed scheduled visits and the HIV infections happened randomly during study, the data are sparse. The number of observations per subject ranged from 1 to 14, with a median of 6. Plots of sample curves from the data are shown in 2.9. As we can see, the CD4 count data are unbalanced, due to mistimed measurements and missing data that resulted from skipped visits and dropout.

This dataset has been studied by many authors (see Yao et al. [54]). The plots of all observed individual trajectories and the estimated mean curve can be found in Yao et al. [54]. We apply our method to this dataset and plot the estimates of the first three eigenfunctions in Figure 2.10. Our estimate of the first eigenfunction is similar to that in

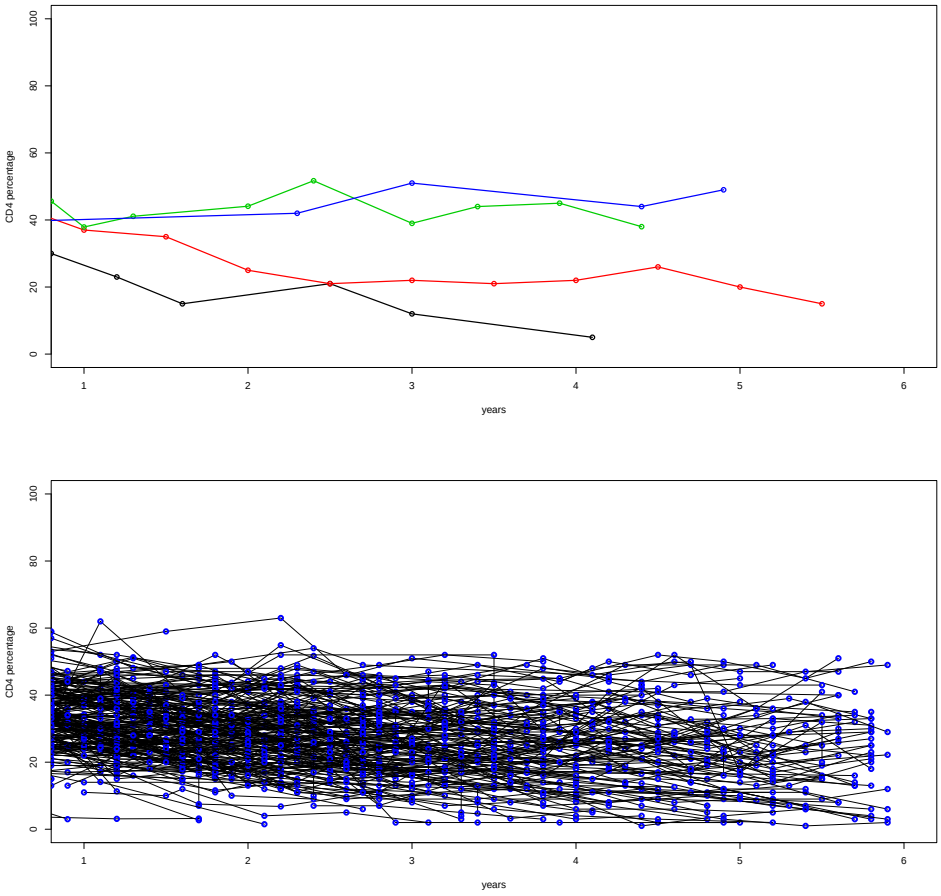

Figure 2.9 Sample curves from CD4 data: (Upper) 4 sample curves; (Lower)Sample curves of 283 patients between 1984 and 1991

Yao et al. [54]. However, there are some differences between our estimates of the second and third eigenfunctions and theirs. Our estimate of the second eigenfunction corresponds to the contrast between the cd4 counts before year 2.5 and those after that. The third estimate corresponds to the contrast between the cd4 counts during the middle course of the observations and the summation of those at the beginning and towards the end of the study period.

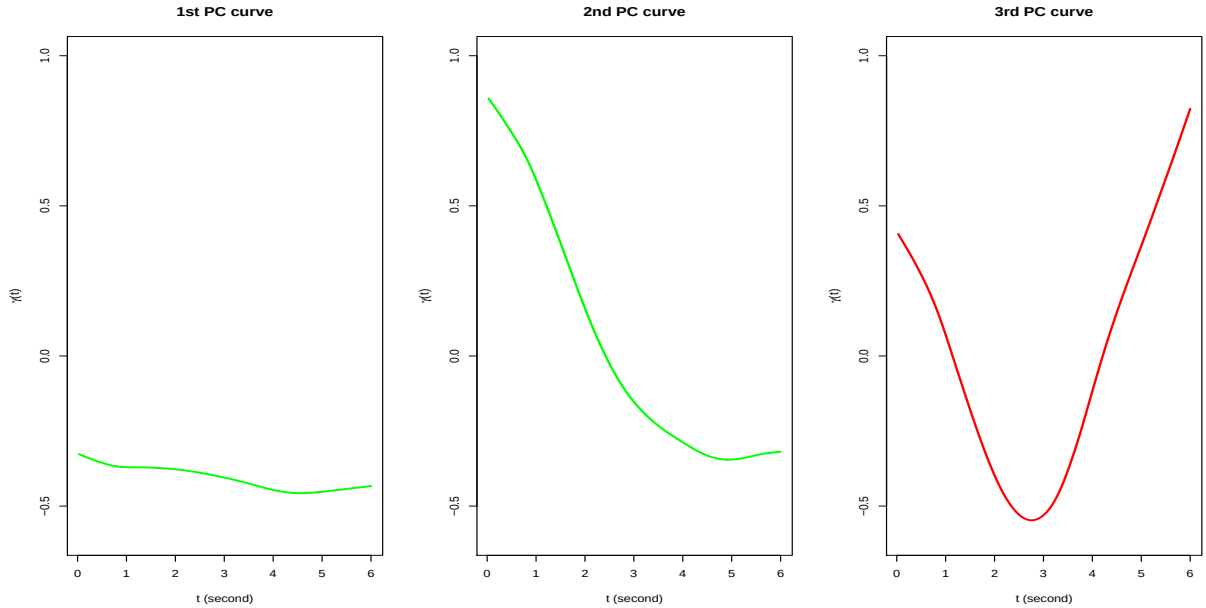

Figure 2.10 Estimates of the first three eigenfunctions for CD4 data

Chapter 3

SPARSE FISHER'S DISCRIMINANT ANALYSIS WITH THRESHOLDED LINEAR CONSTRAINTS

This chapter is organized as follows. In Section 3.1, we introduce notations and briefly review the classic Fisher's discriminant analysis. Our sparse Fisher's LDA method with thresholded linear constraints are introduced in Section 3.2. In Section 3.3, we present the main theoretical results. Sections 3.4 and 3.5 are simulation studies and applications, respectively. The proofs of all theorems are provided in Section 4.11 of Chapter 4.

3.1 Fisher's discriminant analysis

We first introduce the notations used throughout the chapter. For any vector $\mathbf{v} = (v_1, \cdots, v_p)^\mathsf{T}$, let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty = \max_{1\le i\le p}|v_i|$ denote the l_1, l_2, and l_∞ norms of $\mathbf{v}$, respectively. For any $p \times p$ symmetric matrix $\mathbf{M}$, we use $\lambda_{max}(\mathbf{M})$, $\lambda_{min}(\mathbf{M})$ and $\lambda^+_{min}(\mathbf{M})$ to denote the largest eigenvalue, the smallest eigenvalue and the smallest positive eigenvalue of $\mathbf{M}$, respectively. Now suppose that $\mathbf{M}$ is symmetric and nonnegative definite. We define two norms for $\mathbf{M}$,

$$\|\mathbf{M}\| = \sup_{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|_2=1} \|\mathbf{M}\mathbf{v}\|_2 = \lambda_{max}(\mathbf{M}), \text{ and } \|\mathbf{M}\|_\infty = \max_{1\le k,l\le p} |M_{kl}|\,, \tag{3.1}$$

where M_{kl} is the (k,l)th entry of $\mathbf{M}$. The first norm is the usual operator norm and is also called the spectral norm. The second is the max norm.

Throughout this chapter, we assume that the number K of classes is any fixed positive integer number. Suppose that the population in the i-th class has a multivariate normal distribution $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i$ is the true mean of the ith class, $1 \le i \le K$, and $\boldsymbol{\Sigma}$ is the true common within-class covariance matrix for all classes. We assume that the prior probabilities for all the classes are the same and equal to $1/K$. It will be seen that when we add a constant vector to all the observations (including all the training and the test data),

the classification results are not changed under the classification rules involved in this paper, therefore, without loss of generality, we assume that the overall mean of the whole population is zero, that is,

$$\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \cdots + \boldsymbol{\mu}_K = \mathbf{0}. \tag{3.2}$$

Define a $p \times K$ matrix $\mathbf{U} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K]$, which is the collection of the class means. Under the assumption (3.2), the between-class covariance matrix is defined as

$$\mathbf{B} = \sum_{i=1}^{K} \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathsf{T}} / K = \mathbf{U}\mathbf{U}^{\mathsf{T}} / K \tag{3.3}$$

Then Fisher's discriminant analysis method (when the true class means and the true covariance matrix are known) sequentially finds linear combinations $\mathbf{X}\boldsymbol{\alpha}_1, \cdots, \mathbf{X}\boldsymbol{\alpha}_{K-1}$ by solving the following generalized eigenvalue problem. Suppose that we have obtained $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{k-1}$, where $1 \leq i \leq K-2$, then $\boldsymbol{\alpha}_i$ is the solution to

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \boldsymbol{\alpha}^{\mathsf{T}} \mathbf{B} \boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\Sigma} \boldsymbol{\alpha} = 1, \quad \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{\Sigma} \boldsymbol{\alpha}_j = 0, \quad 1 \leq j \leq i-1. \tag{3.4}$$

The Fisher's classification rule is to assign a new observation $\mathbf{x}$ to the class i if

$$(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} \mathbf{D} (\mathbf{x} - \boldsymbol{\mu}_i) < (\mathbf{x} - \boldsymbol{\mu}_j)^{\mathsf{T}} \mathbf{D} (\mathbf{x} - \boldsymbol{\mu}_j) \tag{3.5}$$

for all $1 \leq j \neq i \leq K$, where $\mathbf{D} = \sum_{k=1}^{K-1} \boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^{\mathsf{T}}$.

It is well known that under our setting, that is, the population in each class has a normal distribution with the same covariance matrix and the prior probabilities for all classes are the same, the optimal classification rule is to assign a new observation $\mathbf{x}$ to class i if

$$(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) < (\mathbf{x} - \boldsymbol{\mu}_j)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \tag{3.6}$$

for all $1 \leq j \neq i \leq K$,(See Theorem 6.8.1 in Anderson [3] or Theorem 13.2 in Härdle and Simar [21]). Moreover, the optimal rule (3.6) is equivalent to the Fisher's discriminant rule

(3.5).

In practice, the true class means and the covariance matrix $\mathbf{\Sigma}$ are unknown. Consider a training data set, $\mathbf{X} = \{\mathbf{x}_{ij} : 1 \le i \le K, 1 \le j \le n_i\}$, where $\mathbf{x}_{ij}$ is the jth observation from the ith class and n_i is the number of the observations of the ith class. The numbers $(n_1, n_2, \cdots, n_K)$ can be either random or nonrandom. Let $n = \sum_{i=1}^{K} n_i$. Throughout this paper, we use

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, \quad \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\mathbf{x}_{ij}, \quad \widehat{\mathbf{\Sigma}} = \frac{1}{n-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)^{\mathsf{T}},$$
$$\widehat{\mathbf{B}} = \frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})^{\mathsf{T}}, \quad 1 \le i \le K, \tag{3.7}$$

to denote the sample class means, the sample overall mean, the sample within-class covariance matrix and the sample between-class covariance matrix, respectively. Then the classic Fisher's discriminant analysis is to sequentially obtain the estimates $\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1}$ of $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{K-1}$ by solving

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\mathbf{\Sigma}}\boldsymbol{\alpha} = 1, \quad \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\mathbf{\Sigma}}\widehat{\boldsymbol{\alpha}}_j = 0, \quad 1 \le j < i, \tag{3.8}$$

where $1 \le i \le K-1$. The classification rule is to assign a new observation $\mathbf{x}$ to the class i if

$$(\mathbf{x}-\bar{\mathbf{x}}_i)^{\mathsf{T}}\widetilde{\mathbf{D}}(\mathbf{x}-\bar{\mathbf{x}}_i) < (\mathbf{x}-\bar{\mathbf{x}}_j)^{\mathsf{T}}\widetilde{\mathbf{D}}(\mathbf{x}-\bar{\mathbf{x}}_j), \tag{3.9}$$

for all $1 \le j \ne i \le K$, where $\widetilde{\mathbf{D}} = \sum_{k=1}^{K-1}\widehat{\boldsymbol{\alpha}}_k\widehat{\boldsymbol{\alpha}}_k^{\mathsf{T}}$.

3.2 Sparse Fisher's discriminant analysis with thresholded linear constraints

In the high-dimensional setting, the classic Fisher's discriminant analysis has several drawbacks. First, the sample within-class covariance matrix $\widehat{\mathbf{\Sigma}}$ is not full rank, so the solution to (3.8) does not exist. Second, the sample between-class covariance matrix $\widehat{\mathbf{B}}$ and the sample within-class covariance matrix $\widehat{\mathbf{\Sigma}}$ as given in (3.7) are not consistent estimates in terms of the usual operator norm. Hence, $\widehat{\boldsymbol{\alpha}}_k$, $1 \le k \le K-1$, are not consistent. Third, suppose that we have obtained an estimate $\widetilde{\boldsymbol{\alpha}}_1$ of $\boldsymbol{\alpha}_1$, in order to estimate $\boldsymbol{\alpha}_2$, we have to estimate the

coefficient vector of the linear constraint in (3.4), $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$. However, even if $\widetilde{\boldsymbol{\alpha}}_1$ is a consistent estimate , $\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\alpha}}_1$ is not a consistent estimate of $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$ due to the inconsistency of $\widehat{\boldsymbol{\Sigma}}$. In this section, we describe our method and address these drawbacks. We will consider the cases that $K = 2$ and $K > 2$ separately because when $K = 2$, there is only a component and no linear constraints exist.

3.2.1 The case of $K = 2$

When there are two classes, there is only one component $\boldsymbol{\alpha}_1$ and $\mathbf{B} = (\boldsymbol{\mu}_1\boldsymbol{\mu}_1^\mathsf{T} + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\mathsf{T})/2 = \boldsymbol{\mu}_1\boldsymbol{\mu}_1^\mathsf{T}$ because $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$. It is easily seen that

$$\boldsymbol{\alpha}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}/\sqrt{\boldsymbol{\delta}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}$$

, where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Cai and Liu [9] and Fan et al. [15] imposed l_1 and l_0 sparsity assumptions on $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, respectively. Equivalently, we assume that $\boldsymbol{\alpha}_1$ is sparse in terms of l_1 norm as in Cai and Liu [9]. As in Qi et al. [32], we propose to get an estimate $\widehat{\boldsymbol{\alpha}}_1$ of $\boldsymbol{\alpha}_1$ by solving

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad \boldsymbol{\alpha}^\mathsf{T}\widehat{\mathbf{B}}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^\mathsf{T}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2 = 1, \tag{3.10}$$

where $\|\boldsymbol{\alpha}\|_\lambda^2 = (1-\lambda)\|\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1^2$ and both $\tau \geq 0$ and $0 \leq \lambda \leq 1$ are tuning parameters. The introduction of $\|\boldsymbol{\alpha}\|_2^2$ overcomes the issue that $\widehat{\boldsymbol{\Sigma}}$ is not full rank in high-dimensional setting, and the term $\|\boldsymbol{\alpha}\|_1^2$ encourages the sparsity of the solution. A difference between our penalty and the usual lasso or elastic-net penalty is that we use the squared l_1-norm, which leads to the following scale-invariant property. For any nonzero real number t, $t\widehat{\boldsymbol{\alpha}}_1$ is the solution to the penalized generalized eigenvalue problem,

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad \frac{\boldsymbol{\alpha}^\mathsf{T}\widehat{\mathbf{B}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\mathsf{T}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2}, \tag{3.11}$$

because the objective function is scale-invariant. Note that the problem (3.11) is equivalent to (3.10). This scale-invariant property is intensively used in our theoretical development. Once we obtain $\widehat{\boldsymbol{\alpha}}_1$, our classification rule is to assign a new observation $\mathbf{x}$ to class i if $(\mathbf{x} - \bar{\mathbf{x}}_i)^\mathsf{T}\widehat{\mathbf{D}}(\mathbf{x} - \bar{\mathbf{x}}_i) < (\mathbf{x} - \bar{\mathbf{x}}_j)^\mathsf{T}\widehat{\mathbf{D}}(\mathbf{x} - \bar{\mathbf{x}}_j)$ for $1 \leq j \neq i \leq 2$, where $\widehat{\mathbf{D}} = \widehat{\boldsymbol{\alpha}}_1\widehat{\boldsymbol{\alpha}}_1^\mathsf{T}$.

3.2.2 The case of $K > 2$

If $K > 2$, more than one components need to be estimated. The $\boldsymbol{\alpha}_1$ is estimated in the same way as that when $K = 2$. Since the higher order component $\boldsymbol{\alpha}_i$, $1 < i \leq K-1$, satisfies the constraints in (3.4), $\boldsymbol{\alpha}_i$ is actually orthogonal to the subspace spanned by $\{\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{i-1}\}$ in $\mathbb{R}^p$. Because $\boldsymbol{\alpha}_i$ is the eigenvector of the generalized eigenvalue problem (3.4), $\mathbf{B}\boldsymbol{\alpha}_j$ and $\boldsymbol{\Sigma}\boldsymbol{\alpha}_j$ $(1 \leq j < K-1)$ have the same direction and only differ by a scale factor, which is the j-th eigenvalue. Hence, the subspace spanned by $\{\mathbf{B}\boldsymbol{\alpha}_1, \cdots, \mathbf{B}\boldsymbol{\alpha}_{i-1}\}$ is the same as that of $\{\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{i-1}\}$.

Because neither $\widehat{\boldsymbol{\Sigma}}$ nor $\widehat{\mathbf{B}}$ are consistent estimates of $\boldsymbol{\Sigma}$ and $\mathbf{B}$ in terms of the operator norm, respectively, neither of the subspaces spanned by $\{\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_{i-1}\}$ and $\{\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_{i-1}\}$ is a consistent estimate of the subspace spanned by $\{\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{i-1}\}$ (or by $\{\mathbf{B}\boldsymbol{\alpha}_1, \cdots, \mathbf{B}\boldsymbol{\alpha}_{i-1}\}$), even if $\widehat{\boldsymbol{\alpha}}_j$, $1 \leq j \leq i-1$, are consistent estimates. Therefore, in order to estimate these subspaces, in addition to the sparsity assumption on $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{K-1}\}$, we also make sparsity assumptions on the vectors, $\mathbf{B}\boldsymbol{\alpha}_1, \cdots, \mathbf{B}\boldsymbol{\alpha}_{K-1}$, in terms of l_1 norm, which is equivalent to the sparsity on $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{K-1}$. Lemma 2 in Section 3.3 shows that making sparsity assumptions on $\boldsymbol{\Sigma}\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\Sigma}\boldsymbol{\alpha}_{K-1}$ is equivalent to assuming the sparsity of $\{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, 1 \leq i \neq j \leq K-1\}$ in terms of l_1 norm. This assumption has been made in Shao et al. [37]. Bickel and Levina [7] assumes that $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are sparse when $K = 2$, which implies that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse.

Under the above assumptions, suppose that we have obtained the estimate $\widehat{\boldsymbol{\alpha}}_j$ of $\boldsymbol{\alpha}_j$, $1 \leq j \leq i-1$, then we propose to estimate $\mathbf{B}\boldsymbol{\alpha}_j$ by applying the soft thresholding to $\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j$. This is equivalent to get the estimate $\widehat{\boldsymbol{\xi}}_j$ of $\mathbf{B}\boldsymbol{\alpha}_j$ by solving the following optimization problem:

$$\min_{\boldsymbol{\xi} \in \mathbb{R}^p} \left[\|\boldsymbol{\xi} - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j\|_2^2 + \kappa \|\boldsymbol{\xi}\|_1 \right], \tag{3.12}$$

where $\kappa \geq 0$ is a tuning parameter. It can be shown that the l-th coordinate of $\widehat{\boldsymbol{\xi}}_j$ is

$$(\widehat{\boldsymbol{\xi}}_j)_l = \mathbf{sign}((\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j)_l) \left[|(\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j)_l| - \kappa/2 \right] \mathbf{I}_{[|(\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j)_l| \geq \kappa/2]}, \quad 1 \leq l \leq p, \tag{3.13}$$

where $\mathbf{I}_{[|(\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j)_l| \geq \kappa/2]}$ is the indicator function of $[|(\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_j)_l| \geq \kappa/2]$. We will show that the

subspace spanned by $\{\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\boldsymbol{\xi}}_{i-1}\}$ is a consistent estimate of the subspace spanned by $\{\mathbf{B}\boldsymbol{\alpha}_1, \cdots, \mathbf{B}\boldsymbol{\alpha}_{i-1}\}$ and provide the convergence rate in Section 4. Now suppose that we have obtained the estimates $\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{i-1}$ and $\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\boldsymbol{\xi}}_{i-1}$, then $\widehat{\boldsymbol{\alpha}}_i$ is the solution to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2 = 1, \quad \boldsymbol{\alpha}^{\mathsf{T}}\widehat{\boldsymbol{\xi}}_j = 0, \quad j < i. \tag{3.14}$$

Once we obtain all the estimates $\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1}$, we build the classification rule which assigns a new observation $\mathbf{x}$ to class i if

$$(\mathbf{x} - \bar{\mathbf{x}}_i)^{\mathsf{T}}\widehat{\mathbf{D}}(\mathbf{x} - \bar{\mathbf{x}}_i) < (\mathbf{x} - \bar{\mathbf{x}}_j)^{\mathsf{T}}\widehat{\mathbf{D}}(\mathbf{x} - \bar{\mathbf{x}}_j), \tag{3.15}$$

for all $1 \le j \ne i \le K$, where

$$\widehat{\mathbf{D}} = (\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1})\,\widehat{\mathbf{K}}^{-1}\,(\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1})^{\mathsf{T}}, \tag{3.16}$$

and $\widehat{\mathbf{K}}$ is a symmetric $(K-1)\times(K-1)$ matrix with the (i,j)-th entry equal to $\widehat{\boldsymbol{\alpha}}_i^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_j$. The reason that we use the form (3.16) for $\widehat{\mathbf{D}}$ will be explained in Section 3.3.

3.2.3 Computation

The optimization problems (3.10) and (3.14) are special cases of the the following problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p} \quad \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\Pi}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{C}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_\lambda^2 \le 1, \quad \mathbf{L}\boldsymbol{\alpha} = 0, \tag{3.17}$$

where $\boldsymbol{\Pi}$ and $\mathbf{C}$ are any two $p\times p$ nonnegative definite symmetric matrices, and $\mathbf{L}$ is any matrix with p columns. For example, (3.14) is the special case of (3.17) with $\boldsymbol{\Pi} = \widehat{\mathbf{B}}$, $\mathbf{C} = \widehat{\boldsymbol{\Sigma}}$ and $\mathbf{L} = (\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\boldsymbol{\xi}}_{i-1})^{\mathsf{T}}$. In Qi et al. [32], (3.17) is solved by the following algorithm,

Algorithm 3.2.1. *1. Choose an initial vector $\boldsymbol{\alpha}^{(0)}$ with $\boldsymbol{\Pi}\boldsymbol{\alpha}^{(0)} \ne \mathbf{0}$.*

2. Iteratively compute a sequence $\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \cdots, \boldsymbol{\alpha}^{(i)}, \cdots$ until convergence as follows: for

any $i \geq 1$, compute $\boldsymbol{\alpha}^{(i)}$ by solving

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p}(\boldsymbol{\Pi}\boldsymbol{\alpha}^{(i-1)})^{\mathsf{T}}\boldsymbol{\alpha}, \quad \textit{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{C}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_{\lambda}^2 \leq 1, \quad \mathbf{L}\boldsymbol{\alpha} = \mathbf{0}. \tag{3.18}$$

The key step (3.18) of Algorithm 3.2.1 is a special case of the following problem with $\mathbf{c} = \boldsymbol{\Pi}\boldsymbol{\alpha}^{(i-1)}$:

$$\max_{\boldsymbol{\alpha}} \mathbf{c}^{\mathsf{T}}\boldsymbol{\alpha}, \quad \text{subject to} \quad \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{C}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_{\lambda}^2 \leq 1, \quad \mathbf{L}\boldsymbol{\alpha} = 0, \tag{3.19}$$

where $\mathbf{c}$ is any nonzero vector. The algorithm and the related theory to solve (3.19) have been developed and described in details in supplementary materials of Qi et al. [32].

3.3 Asymptotic consistency and asymptotic optimality

In this section, we study the asymptotic properties of the proposed method in Section 3.2. We first consider two mechanisms of class label generation. The first is a random mechanism in which sample observations are randomly drawn from any of K classes with equal probability $1/K$. Hence, $(n_1, n_2, \cdots, n_K)$ follows a multinomial distribution with parameters n and $(1/K, \cdots, 1/K)$. In this case, we have the following result.

Lemma 1. *Suppose that $(n_1, n_2, \cdots, n_K)$ follows a multinomial distribution with parameters n and $(1/K, \cdots, 1/K)$. Given any (K, n, p) satisfying that $p \geq 2$, $K \leq p+1$ and $\sqrt{K \log p/n}$ is bounded by some constant d_0, for any $M > 0$, we have*

$$P\left(\max_{1\leq i\leq K}\left|\frac{n_i}{n} - \frac{1}{K}\right| > C\sqrt{\frac{\log p}{Kn}}\right) \leq p^{-M} \tag{3.20}$$

for all $C \geq (M+3)(d_0+1)$.

In the following Condition 1, we will assume that the distributions of $\mathbf{x}_{ij}$ is independent of this random mechanism of class label generation. The second mechanism is nonrandom, that is, $(n_1, n_2, \cdots, n_K)$ are nonrandom numbers. In this case, we will impose the following Condition 1 (a) on these numbers.

Condition 1.

(a). If $(n_1, n_2, \cdots, n_K)$ are nonrandom, then there exists a constant C_0 (independent of n, p and K), such that we have $\max_{1\le i\le K} |n_i/n - 1/K| \le C_0\sqrt{\log p/(Kn)}$ for all large enough n. If $(n_1, n_2, \cdots, n_K)$ are random as in Lemma 1, we assume that they and $\mathbf{x}_{ij}$, $1 \le i \le K$ and $1 \le j \le n_i$, are independent.

(b). There exists a constant c_0 (independent of n, p and K) such that

$$c_0^{-1} \le \lambda_{min}(\mathbf{\Sigma}) \le \lambda_{max}(\mathbf{\Sigma}) \le c_0, \quad \text{and } \max_{1\le i\le K} \|\boldsymbol{\mu}_i\|_\infty \le c_0.$$

Lemma 1 and Condition 1 (a) ensure that the number of observations in different classes do not differ greatly in each of the two mechanisms. The regularity condition for $\mathbf{\Sigma}$ in Condition 1 (b) has been used by many authors. The condition about $\boldsymbol{\mu}_i$ can be achieved by scaling each of the p variables. Under Condition 1, we have the following two probability inequalities about $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty$ and $\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty$, which play basic roles in our theoretical development. Recall that $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{B}}$ are the sample within-class covariance matrix and the sample between-class covariance matrix, respectively, as defined in (3.7).

Theorem 3.3.1. *Suppose that Condition 1 holds, $p \ge 2$, $K \le p+1$ and $K \log p/n \to 0$ as $n \to \infty$. Then for any $M > 0$, we can find C large enough and independent of n, p and K such that*

$$P\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \le p^{-M}, \quad P\left(\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \le p^{-M}$$

for all large enough n.

Remark 3.3.2. *Theorem 3.3.1 holds even if $K \to \infty$ as $n \to \infty$ under the conditions in the theorem. However, since we need the condition that K is bounded in the following theorems, we fix K in this paper.*

We next impose some conditions on the maximum values of the generalized eigenvalue problem (3.4). Define

$$\mathbf{\Delta} = \mathbf{U}^\top \mathbf{\Sigma}^{-1} \mathbf{U}, \quad \mathbf{\Xi} = \mathbf{\Sigma}^{-1/2} \mathbf{B} \mathbf{\Sigma}^{-1/2}, \tag{3.21}$$

where $\mathbf{U} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K]$, $\boldsymbol{\Delta}$ is $K \times K$ and plays the same role as that of Δ_p in Shao et al. [37] and Cai and Liu [9] when $K = 2$, and $\boldsymbol{\Xi}$ is a $p \times p$ nonnegative definite matrix. Solving the generalized eigenvalue problem (3.4) is equivalent to computing the eigenvalues and eigenvectors of $\boldsymbol{\Xi}$. As $\boldsymbol{\alpha}_k$, $1 \le k \le K-1$, are the generalized eigenvectors of the problem (3.4), we have

$$\mathbf{B}\boldsymbol{\alpha}_k = \lambda_k \boldsymbol{\Sigma}\boldsymbol{\alpha}_k, \quad \text{and hence,} \quad \boldsymbol{\Xi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_k = \boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\alpha}_k = \lambda_k \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_k, \tag{3.22}$$

for any $1 \le k \le K-1$, where λ_k is the corresponding generalized eigenvalue. Therefore,

$$\boldsymbol{\gamma}_1 = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_1, \quad \boldsymbol{\gamma}_2 = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_2, \quad \cdots, \quad \boldsymbol{\gamma}_{K-1} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_{K-1}, \tag{3.23}$$

are the eigenvectors of $\boldsymbol{\Xi}$ with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$, respectively. So they are orthogonal to each other. In the following, we will use $\lambda_k(\boldsymbol{\Xi})$, $1 \le k \le K-1$, to denote the eigenvalues of $\boldsymbol{\Xi}$, which are also the generalized eigenvalues and the maximum values of (3.4). Since $\boldsymbol{\Xi}$ has the same rank as $\mathbf{B}$ which is equal to $K-1$ due to the constraint (3.2), $\boldsymbol{\Xi}$ has at most $K-1$ positive eigenvalues. By the conditions $\boldsymbol{\alpha}_k^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k = 1$, $1 \le k \le K-1$, we have $\|\boldsymbol{\gamma}_1\|_2 = \|\boldsymbol{\gamma}_2\|_2 = \cdots = \|\boldsymbol{\gamma}_{K-1}\|_2 = 1$. Let

$$\widehat{\boldsymbol{\gamma}}_1 = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\alpha}}_1, \quad \widehat{\boldsymbol{\gamma}}_2 = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\alpha}}_2, \cdots, \widehat{\boldsymbol{\gamma}}_{K-1} = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\alpha}}_{K-1}, \tag{3.24}$$

which are estimates of $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K-1}$, respectively. Since $-\widehat{\boldsymbol{\alpha}}_k$ is also the solution to the optimization problem in (3.10) or (3.14), without loss of generality, we choose the sign of $\widehat{\boldsymbol{\alpha}}_k$ such that $\widehat{\boldsymbol{\gamma}}_k^{\mathsf{T}}\boldsymbol{\gamma}_k \ge 0$, for $1 \le k \le K-1$. We impose the following regularity conditions on the eigenvalues of $\boldsymbol{\Xi}$.

Condition 2. *There exist positive constants c_1, c_2 and c_3 which are all independent of n, p and K such that*

(a). $\lambda_1(\boldsymbol{\Xi}) \ge \lambda_2(\boldsymbol{\Xi}) \ge \cdots \ge \lambda_{K-1}(\boldsymbol{\Xi}) \ge c_1$,

(b). $\min\left\{\frac{\lambda_1(\boldsymbol{\Xi})-\lambda_2(\boldsymbol{\Xi})}{\lambda_1(\boldsymbol{\Xi})}, \quad \frac{\lambda_2(\boldsymbol{\Xi})-\lambda_3(\boldsymbol{\Xi})}{\lambda_2(\boldsymbol{\Xi})}, \cdots, \frac{\lambda_{K-2}(\boldsymbol{\Xi})-\lambda_{K-1}(\boldsymbol{\Xi})}{\lambda_{K-2}(\boldsymbol{\Xi})}\right\} \ge c_2$,

(c). The ratio between the largest and the smallest eigenvalue: $\lambda_1(\boldsymbol{\Xi})/\lambda_{K-1}(\boldsymbol{\Xi}) \le c_3$.

Condition 2(b) prevents the cases that the spacing between adjacent eigenvalues is too small. Condition 2(c) excludes the situations where the effects of higher order components are dominated by those of lower order components and are negligible asymptotically.

Now we consider the choice of the tuning parameters, τ and λ, in the penalized optimization problem (3.10) and (3.14). We will show that the choice of λ is not essential for the asymptotic convergence rates as long as it is asymptotically bounded away from zero. In the following, we will choose tuning parameters (τ_n, λ_n) (which depend on the sample size n) satisfying

$$0 < \lambda_n < 1, \quad \liminf_{n\to\infty} \lambda_n > \lambda_0, \quad \tau_n = Cs_n, \quad \text{where} \quad s_n = \sqrt{\frac{K \log p}{n}}, \tag{3.25}$$

$\lambda_0 > 0$ and C are constants independent of n, p and K. The constant C is chosen based on Theorem 3.3.1 such that for all large enough n,

$$P\left(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty > \frac{C}{C_2} s_n\right) \le p^{-1}, \quad P\left(\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty > \frac{C}{C_2} s_n\right) \le p^{-1}, \tag{3.26}$$

where $C_2 = 2(1 + c_1^{-1})/\lambda_0$ and c_1 is the constant in Condition 2 (a). Define the event

$$\Omega_n = \left\{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \le \tau_n/C_2, \quad \|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty \le \tau_n/C_2\right\}, \text{ then } P(\Omega_n) \ge 1 - 2p^{-1} \tag{3.27}$$

by (3.26). Since the probability of the complement of Ω_n goes to zero as $n, p \to \infty$, we will only consider the elements in Ω_n.

We adopt the same definition of asymptotic optimality for a linear classification rule as in Shao et al. [37], Cai and Liu [9], Fan et al. [15] and other papers. Let T_{OPT} denote the optimal linear classification rule (3.5) or (3.6) and R_{OPT} represent its misclassification error rate.

Definition 3.3.3. *Let T be a linear classification rule with conditional misclassification rate $R_T(\mathbf{X})$, given the training sample $\mathbf{X}$. Then T is asymptotically optimal if*

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \le o_p(1). \tag{3.28}$$

Note that (3.28) implies that $R_T(\mathbf{X}) - R_{OPT} \leq o_p(1)$ because $0 \leq R_T(\mathbf{X}) \leq 1$. Hence we have $R_T(\mathbf{X}) \to R_{OPT}$ in probability and $E[R_T(\mathbf{X})] \to R_{OPT}$. If R_{OPT} is bounded away from 0, then $R_T(\mathbf{X}) - R_{OPT} \leq o_p(1)$ also implies (3.28). However, if $R_{OPT} \to 0$, (3.28) is stronger than the inequality $R_T(\mathbf{X}) - R_{OPT} \leq o_p(1)$. In the following, we will consider the asymptotic properties of our method for $K = 2$ and $K > 2$ separately because of the more complicated classification boundary and the additional linear constraints when $K > 2$. In both cases, we will assume that K is fixed, $p \to \infty$ and $s_n = \sqrt{K \log p/n} \to 0$ as $n \to \infty$.

3.3.1 The case of $K = 2$

In this case, there exists only one component $\boldsymbol{\alpha}_1$. The following theorem provides an upper bound for the sparsity (measured by the l_1 norm) and the consistency of the estimator $\widehat{\boldsymbol{\alpha}}_1$ obtained from (3.10).

Theorem 3.3.4. *Suppose that $K = 2$ and Conditions 1-2 hold. If $s_n \to 0$ and $\|\boldsymbol{\alpha}_1\|_1^2 s_n \to 0$ as $n, p \to \infty$, then for all large enough n, we have, in Ω_n,*

$$\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \leq 6\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0, \quad \|\widehat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\|_2^2 \leq C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n, \quad \|\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1\|_2^2 \leq c_0 C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n, \tag{3.29}$$

where C_5 is a constant independent of n and p, and c_0 is the constant in Condition 1 (b). Therefore, $\widehat{\boldsymbol{\alpha}}_1$ is a consistent estimate of $\boldsymbol{\alpha}_1$.

Next, we provide explicit formulas for the misclassification errors of the optimal rule and our rule in terms of $\mathbf{D} = \boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^\mathsf{T}$ and $\widehat{\mathbf{D}} = \widehat{\boldsymbol{\alpha}}_1\widehat{\boldsymbol{\alpha}}_1^\mathsf{T}$, respectively, when $K = 2$. Then based on Theorem 3.3.4, we can prove the asymptotic optimality of our method and provide the corresponding convergence rate.

Theorem 3.3.5. *Suppose that $K = 2$ and Conditions 1-2 hold. Then the misclassification rate of the optimal rule* (3.5) *and the conditional misclassification rate of our sparse LDA rule as given in Section 3.2.1 are*

$$\begin{aligned} R_{OPT} &= \Phi\left(-\frac{\boldsymbol{\delta}^\mathsf{T}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^\mathsf{T}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right), \\ R(\mathbf{X}) &= \frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^\mathsf{T}\widehat{\mathbf{D}}(2\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{2\|\widehat{\boldsymbol{\delta}}^\mathsf{T}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right) + \frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^\mathsf{T}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2 - 2\boldsymbol{\mu}_1)}{2\|\widehat{\boldsymbol{\delta}}^\mathsf{T}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right), \end{aligned} \tag{3.30}$$

respectively, where Φ *is the cumulative distribution function of the standard normal distribution,* $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ *and* $\widehat{\boldsymbol{\delta}} = \bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1$. *Moreover, if* $s_n \to 0$ *and* $\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n \to 0$ *as* $n, p \to \infty$, *our method is asymptotically optimal and we have*

$$\frac{R(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left(\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n\right). \tag{3.31}$$

Remark 3.3.6.

(1). The misclassification rate of the optimal rule is expressed in terms of $\boldsymbol{\Sigma}^{-1}$, *that is,* $R_{OPT} = \Phi\left(-\sqrt{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}/2\right)$ *in Equation (1) in Shao et al. [37] and Equation (5) in Cai and Liu [9]. Since it is established that* $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \mathbf{D}\boldsymbol{\delta}$ *in* (4.153) *(Supplementary Material) in the proof of Lemma 8, the* R_{OPT} *in* (4.54) *is the same as in those papers.*

(2). Cai and Liu [9] assumed sparsity on $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ *(where they used the notation* $\boldsymbol{\Omega}$ *rather than* $\boldsymbol{\Sigma}^{-1}$*), and obtained the convergence rate*

$$\frac{R(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left\{\left(\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\|_1\sqrt{\Delta_p} + \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\|_1^2\right)\sqrt{\frac{\log p}{n}}\right\}, \tag{3.32}$$

(see Theorem 3 in Cai and Liu [9]), where $\Delta_p = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$. *When* $K = 2$, $\boldsymbol{\alpha}_1 = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}/\sqrt{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}$. *By* (4.65) *(Supplementary Material) in the proof of Theorem 3.3.5, we have* $\boldsymbol{\delta}^\top\mathbf{D}\boldsymbol{\delta} = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = 4\lambda_1(\boldsymbol{\Xi})$. *Hence, our convergence rate on the right hand side of* (4.55) *is*

$$\begin{aligned} O_p\left(\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n\right) &= O_p\left\{(\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})\left\|\frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}{\sqrt{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\right\|_1^2\sqrt{\frac{K\log p}{n}}\right\} \\ &= O_p\left(\|\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\|_1^2\sqrt{\frac{\log p}{n}}\right). \end{aligned}$$

Compared to the convergence rate in (3.32), *our convergence rate does not have the first term in* (3.32). *Hence, the convergence rate of our method is the same as or better than that in Cai and Liu [9].*

3.3.2 The case of $K > 2$

In this subsection, we study the asymptotic properties of our method when $K > 2$. We first show that making sparsity assumptions on $\{\mathbf{\Sigma}\boldsymbol{\alpha}_1, \cdots, \mathbf{\Sigma}\boldsymbol{\alpha}_{K-1}\}$ is equivalent to or weaker than assuming the sparsity of $\{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j, 1 \leq i \neq j \leq K\}$.

Lemma 2. *Suppose that Conditions 1-2 hold. Then*

$$\frac{1}{(K-1)c_0\sqrt{2K\lambda_1(\mathbf{\Xi})}}\left(\max_{1\leq i\neq j\leq K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right) \leq \max_{1\leq i\leq K-1}\|\mathbf{\Sigma}\boldsymbol{\alpha}_i\|_1 \leq \frac{\sqrt{c_3}}{\sqrt{\lambda_1(\mathbf{\Xi})}}\left(\max_{1\leq i\neq j\leq K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right).$$

By Lemma 2, since $\lambda_1(\mathbf{\Xi}) \geq c_1$ by Condition 2, if $\lambda_1(\mathbf{\Xi})$ is bounded from the above, then $\max_{1\leq i\leq K-1}\|\mathbf{\Sigma}\boldsymbol{\alpha}_i\|_1$ has the same order as $\max_{1\leq i\neq j\leq K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1$. If $\lambda_1(\mathbf{\Xi}) \to \infty$, then we have $\max_{1\leq i\leq K-1}\|\mathbf{\Sigma}\boldsymbol{\alpha}_i\|_1 / \max_{1\leq i\neq j\leq K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1 \to 0$. We define the following measurement of sparsity on $\boldsymbol{\alpha}_i$ and $\mathbf{\Sigma}\boldsymbol{\alpha}_i$, $1 \leq i \leq K-1$:

$$\Lambda_p = \max_{1\leq i\leq K-1}\{\|\boldsymbol{\alpha}_i\|_1, \|\mathbf{\Sigma}\boldsymbol{\alpha}_i\|_1\}. \tag{3.33}$$

In the following theorem, we show that for each $1 \leq i \leq K-1$, the sparsity of the estimate $\widehat{\boldsymbol{\alpha}}_i$ (measured by the l_1 norm) is bounded by Λ_p multiplied by a constant which does not depend on n and p, and $\widehat{\boldsymbol{\alpha}}_i$ is a consistent estimate. Moreover, we show that the subspace spanned by $\{\widehat{\boldsymbol{\xi}}_1, \cdots, \widehat{\boldsymbol{\xi}}_i\}$ is a consistent estimate of the subspace spanned by $\{\mathbf{B}\boldsymbol{\alpha}_1, \cdots, \mathbf{B}\boldsymbol{\alpha}_i\}$ (or equivalently the subspace spanned by $\{\mathbf{\Sigma}\boldsymbol{\alpha}_1, \cdots, \mathbf{\Sigma}\boldsymbol{\alpha}_i\}$) and provide the convergence rates. In this paper, to measure whether the two subspaces with the same dimensions in $\mathbb{R}^p$ are close to each other, we use the operator norm of the difference between the projection matrices on the two subspaces.

Theorem 3.3.7. *Suppose that Conditions 1-2 hold. Let the tuning parameter in the optimization problem* (3.12)*, $\kappa_n = \widetilde{C}\lambda_1(\mathbf{\Xi})\Lambda_p s_n$, where $\widetilde{C}$ is a constant large enough and independent of n and p. For any $1 \leq i \leq K-1$, let $\mathbf{Q}_i$ and $\widehat{\mathbf{Q}}_i$ be the orthogonal projection matrices onto*

the following subspaces of $\mathbb{R}^p$*, respectively,*

$$\mathbf{W}_i = \text{span}\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_i\}, \quad \widehat{\mathbf{W}}_i = \text{span}\{\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \cdots, \widehat{\boldsymbol{\xi}}_i\}, \tag{3.34}$$

where $\boldsymbol{\xi}_i = \mathbf{B}\boldsymbol{\alpha}_i = \lambda_i(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_i$*. If* $s_n \to 0$ *and* $\Lambda_p^2 s_n \to 0$ *as* $n, p \to \infty$*, then for each* $1 \le i \le K-1$*, there exist constants* $D_{i,1}$*,* $D_{i,2}$ *and* $D_{i,3}$ *independent of* n *and* p *such that in* Ω_n*,*

$$\|\widehat{\boldsymbol{\alpha}}_i\|_1 \le D_{i,1}\Lambda_p, \quad \|\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i\|_2^2 \le D_{i,2}\Lambda_p^2 s_n, \quad \|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|^2 \le D_{i,3}\Lambda_p^2 s_n. \tag{3.35}$$

Hence, for each $1 \le i \le K-1$*,* $\widehat{\boldsymbol{\alpha}}_i$ *is a consistent estimate of* $\boldsymbol{\alpha}_i$*, and the projection matrix* $\widehat{\mathbf{Q}}_i$ *is a consistent estimate of* $\mathbf{Q}_i$*.*

Based on Theorem 3.3.7, we will prove the asymptotic optimality of our classification rule and provide the corresponding convergence rate. However, because the classification boundary is usually complicated and no explicit formula for the classification error generally exist when $K > 2$, we first prove a theorem which provides the asymptotic optimality and the corresponding convergence rate for a general linear classification rule. Then by applying the general result to our sparse Fisher's discriminant analysis method, we obtain the corresponding asymptotic optimality results. Define

$$\mathbf{a}_{ji} = \boldsymbol{\Sigma}^{1/2}\mathbf{D}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i), \quad \mathbf{b}_{ji} = \frac{1}{2}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i), \tag{3.36}$$

where $1 \le i, j \le K$. The Fisher's optimal rule T_{OPT} in (3.5) assigns a new observation $\mathbf{x}$ to the ith class if $\mathbf{a}_{ji}^\mathsf{T}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{b}_{ji}) < 0$ for all $j \ne i$. Consider a general linear classification rule T which assigns a new observation $\mathbf{x}$ to the ith class if

$$\widehat{\mathbf{a}}_{ji}^\mathsf{T}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \widehat{\mathbf{b}}_{ji}) < 0, \text{ for all } j \ne i, \tag{3.37}$$

where $\widehat{\mathbf{a}}_{ji}$ and $\widehat{\mathbf{b}}_{ji}$ are estimates of $\mathbf{a}_{ji}$ and $\mathbf{b}_{ji}$, respectively. Let $R_T(\mathbf{X})$ denote the conditional misclassification rate of the rule T given the training sample $\mathbf{X}$. The following theorem studies the asymptotic property of the general linear classification rule T. In fact, many

linear classification rules in practice can be written in this form. For example, the classic Fisher's rule (3.9) is of the form (3.37) with $\widehat{\mathbf{a}}_{ji} = \boldsymbol{\Sigma}^{1/2}\widetilde{\mathbf{D}}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i)$ and $\widehat{\mathbf{b}}_{ji} = \frac{1}{2}(\bar{\mathbf{x}}_j + \bar{\mathbf{x}}_i)$. The rule of our sparse Fisher's discriminant analysis method is also a special case of (3.37) with

$$\widehat{\mathbf{a}}_{ji} = \boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i), \quad \widehat{\mathbf{b}}_{ji} = \frac{1}{2}(\bar{\mathbf{x}}_j + \bar{\mathbf{x}}_i), \tag{3.38}$$

where $\widehat{\mathbf{D}}$ is defined in (3.16).

Theorem 3.3.8. *Suppose that Conditions 1 and 2 hold and the general classification rule T in* (3.37) *satisfies: $\widehat{\mathbf{a}}_{ji} = -\widehat{\mathbf{a}}_{ij}$ and $\widehat{\mathbf{b}}_{ji} = \widehat{\mathbf{b}}_{ij}$. Let $\{\delta_n : n \geq 1\}$ be a sequence of nonrandom positive numbers with $\delta_n \to 0$ and $\lambda_{max}(\boldsymbol{\Delta})\delta_n \to 0$ as $n \to \infty$. For any $1 \leq j \neq i \leq K$, let*

$$\mathbf{a}_{ji} = t_{ji}\widehat{\mathbf{a}}_{ji} + (\mathbf{a}_{ji})_{\perp}$$

be an orthogonal decomposition of $\widehat{\mathbf{a}}_{ji}$, where $t_{ji}\widehat{\mathbf{a}}_{ji}$ is the orthogonal projection of $\mathbf{a}_{ji}$ along the direction of $\widehat{\mathbf{a}}_{ji}$, t_{ji} is a real number, and $(\mathbf{a}_{ji})_{\perp}$ is orthogonal to $t_{ji}\widehat{\mathbf{a}}_{ji}$. Let

$$\widehat{d}_{ji} = \widehat{\mathbf{a}}_{ji}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i)\ , \quad d_{ji} = \mathbf{a}_{ji}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i) = \frac{1}{2}\|\mathbf{a}_{ji}\|_2^2. \tag{3.39}$$

If the following conditions are satisfied,

$$\|\mathbf{a}_{ji}\|_2^2 - \|\widehat{\mathbf{a}}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 O_p(\delta_n), \quad t_{ji} = 1 + O_p(\delta_n), \quad d_{ji} - \widehat{d}_{ji} = \|\widehat{\mathbf{a}}_{ji}\|_2^2 O_p(\delta_n), \tag{3.40}$$

where $O_p(\delta_n)$ are uniform for all $1 \leq j \neq i \leq K$, then we have

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left(K^2\sqrt{\lambda_{max}(\boldsymbol{\Delta})\delta_n \log\left[\{\lambda_{max}(\boldsymbol{\Delta})\delta_n)\}^{-1}\right]}\right). \tag{3.41}$$

The convergence rate in Theorem 3.3.8 is given in terms of $\lambda_{max}(\boldsymbol{\Delta})\delta_n$. By Lemma 9 (Supplementary Material), there is a simple relationship: $\lambda_{max}(\boldsymbol{\Delta}) = K\lambda_{max}(\boldsymbol{\Xi}) = K\lambda_1(\boldsymbol{\Xi})$. Moreover, in this paper, we assume that K is fixed. Then the inequality in (4.142) can be

given in terms of $\lambda_1(\mathbf{\Xi})\delta_n$ as follows,

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left(\sqrt{\lambda_1(\mathbf{\Xi})\delta_n \log\left[\{\lambda_1(\mathbf{\Xi})\delta_n\}^{-1}\right]}\right).$$

Applying Theorem 3.3.8 to our classification rule (3.15), which is a special case of the general classification rule (3.37) with $\widehat{\mathbf{a}}_{ji}$ and $\widehat{\mathbf{b}}_{ji}$ as given in (3.38), we get the asymptotic optimality of our classification rule as stated in the following theorem.

Theorem 3.3.9. *Suppose that Conditions 1-2 hold, $s_n \to 0$ and $\lambda_1(\mathbf{\Xi})\Lambda_p^2 s_n \to 0$ as $n, p \to \infty$. Then the classification rule* (3.15) *of our sparse Fisher's discriminant analysis method is asymptotically optimal. Moreover, we have*

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left(\sqrt{\lambda_1(\mathbf{\Xi})\Lambda_p^2 s_n \log\left[\{\lambda_1(\mathbf{\Xi})\Lambda_p^2 s_n\}^{-1}\right]}\right). \tag{3.42}$$

Remark 3.3.10. *Now we explain why we choose the particular form* (3.16) *of $\widehat{\mathbf{D}}$ in our classification rule* (3.15)*. By* (4.106) *in the proof of Theorem 3.3.9, $\mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{\Sigma}^{-1/2}$ is equal to the projection matrix onto the subspace spanned by $\{\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K-1}\}$. By* (4.103)*,* (4.104) *and* (4.109) *in the proof of the theorem, one can see that $\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{D}}\mathbf{\Sigma}^{-1/2}$ is a consistent estimate of the projection matrix onto the subspace spanned by $\{\widehat{\boldsymbol{\gamma}}_1, \cdots, \widehat{\boldsymbol{\gamma}}_{K-1}\}$ and hence it is a consistent estimate of the projection matrix $\mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{\Sigma}^{-1/2}$.*

3.4 Simulation studies

In this section, we compare the proposed sparse Fisher's discriminant analysis with thresholded linear constraints (SFDA-threshold) with the sparse Fisher's discriminant analysis without thresholding (SFDA) (Qi et al. [32]), regularized discriminant analysis (RDA) (Guo et al. [19], R package "rda") and penalized discriminant analysis (PDA) (Witten and Tibshirani [52], R package "penalizedLDA"). Three simulation models are considered. In each simulation, 50 independent data sets are simulated each of which has 1500 observations and three classes. In each dataset, for each observation, we randomly select a class label and then generate the value of $\mathbf{x}$ based on the distribution of that class. Then the 1500 observations in each dataset are randomly split into the training set with 150 observations and the test set

with 1350 observations. There are 500 features (that is, p=500) in these datasets. For our methods, SFDA-threshold and SFDA, we use the usual cross-validation procedure to select tuning parameters τ from $\{0.5, 1, 5, 10\}$, and μ from $\{0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$. For SFDA-threshold, we choose κ in (3.12) from the three values which are equal to $\|\hat{\alpha}_j\|_1$ multiplied by 0, 0.001 and 0.01, respectively. For RDA and PDA, the default cross-validation procedure in the corresponding packages are used. The details of the three simulation studies are provided below.

(a). *Simulation 1:* There is no overlap between the features for different classes, but there are correlations among some feature variables. Specifically, let x_{ij} be the i^{th} observation on the j^{th} variable, $1 \leq j \leq 500$ and $1 \leq i \leq 1500$. If the i^{th} observation is in class $k(= 1, 2, 3)$, then $x_{ij} = \mu_{kj} + Z_i + \epsilon_{ij}$ if $1 \leq j \leq 30$, and $x_{ij} = \mu_{kj} + \epsilon_{ij}$ if $j \geq 31$, where $Z_i \sim \text{Normal}(0, 1)$ and $\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$ are independent. Here $\mu_{1j} \sim \text{Normal}(1, 0.8^2)$ if $1 \leq j \leq 20$, $\mu_{2j} \sim \text{Normal}(4, 0.8^2)$ if $21 \leq j \leq 30$, $\mu_{3j} \sim \text{Normal}(1, 0.8^2)$ if $31 \leq j \leq 50$ and $\mu_{kj} = 0$ otherwise. We consider the cases that $\sigma^2 = 1$, 1.5^2 and 4, respectively.

(b). *Simulation 2:* There are overlaps between the features for different classes and the variables are correlated. The i^{th} observation, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{i,500}) \sim \text{Normal}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_k = (\mu_{k,1}, \mu_{k,2}, \cdots, \mu_{k,500})$, if it is in class k, $1 \leq k \leq 3$. The covariance matrix $\boldsymbol{\Sigma}$ is block diagonal, with five blocks each of dimension 100×100. The five blocks are the same and have (j, j') element $0.6^{|j-j'|} \times \sigma^2$. Also, $\mu_{1j} \sim \text{Normal}(1, 1)$, $\mu_{2j} \sim \text{Normal}(2, 1)$ and $\mu_{3j} \sim \text{Normal}(3, 1)$ if $1 \leq j \leq 10$ or $101 \leq j \leq 110$ and $\mu_{kj} = 0$ otherwise. We consider $\sigma^2 = 1$, 2 and 3.

(c). *Simulation 3:* Observations from different classes have different distributions about the class means. If the i^{th} observation is in class k, $\mathbf{x}_i \sim \text{Normal}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. We take $\mu_{1j} = 3$ if $1 \leq j \leq 10$, $\mu_{2j} = 2$ if $1 \leq j \leq 20$, $\mu_{3j} = 1$ if $1 \leq j \leq 30$, and $\mu_{kj} = 0$ otherwise. The covariance matrix $\boldsymbol{\Sigma}_1$ is diagonal with the diagonal elements generated from the uniform distribution in $(0.5, 2) \times \sigma^2$. $\boldsymbol{\Sigma}_2$ is block diagonal, with five blocks each of dimension 100×100. The blocks have (j, j') element $0.9^{|j-j'|} \times \sigma^2$. And $\boldsymbol{\Sigma}_3$ is block diagonal, with five blocks each of dimension 100×100. The blocks have (j, j') element $0.6 \times \sigma^2$ if $j \neq j'$ and σ^2 otherwise. We consider $\sigma^2 = 1$, 2 and 3.

The mean misclassification rates (percentages) of 50 data sets for each simulation are shown in Table 3.1, with standard deviations in parentheses. SFDA-threshold performs similarly with SFDA and both methods have good prediction accuracies in all the simulations. Therefore, in addition to the theoretical advantages, the newly proposed method has good empirical performance.

Table 3.1 The averages and standard deviations of misclassification rates (%) for the simulations in Section 3.4.

	σ^2	**SFDA-threshold**	**SFDA**	**RDA**	**PDA**
Simulation 1	1	0.21(0.26)	0.24(0.26)	0.32(0.39)	2.37(1.46)
	1.5^2	1.52(0.77)	1.54(0.71)	1.75(0.96)	5.40(2.07)
	4	8.78(4.06)	8.60(3.71)	10.20(4.41)	12.73(4.32)
Simulation 2	1	0.48(0.43)	0.48(0.47)	0.79(0.73)	0.86(0.57)
	2	3.15(2.40)	3.29(2.38)	3.61(2.15)	4.84(2.45)
	3	5.05(2.57)	5.10(2.43)	6.05(2.99)	8.55(3.52)
Simulation 3	1	4.86(1.12)	4.85(1.12)	7.71(2.03)	9.51(4.20)
	2	13.02(2.73)	12.84(2.79)	18.74(2.84)	20.42(5.72)
	3	21.49(3.45)	21.48(3.35)	26.56(3.58)	29.74(7.61)

3.5 Application to multivariate functional data

With the advance of techniques, multiple curves can be extracted and recorded simultaneously for one subject in a single experiment. In this section, we consider two real datasets where observations are classified into multiple categories and for each subject, multiple curves were measured. We first apply the wavelet transformation to those curves, and then apply our method to the obtained wavelet coefficients. The setting for the tuning parameters is the same as that in the simulation studies.

3.5.1 Daily and sports activities data

The daily and sports activities data set, available in UCI Machine Learning Repository Bache and Lichman [4], recorded motion sensor data of 19 daily and sports activities each performed by 8 subjects (4 female, 4 male, between the ages 20 and 30) in their own style for

5 minutes. Five Xsens MTx units are used on the torso, arms, and legs Altun and Barshan [1], Altun et al. [2], Barshan and Yüksek [5]. Nine sensors (x, y, z accelerometers, x, y, z gyroscopes, x, y, z magnetometers) were placed on each of five body parts (torso, right arm, left arm, right leg, left leg) and calibrated to acquire data at 25 Hz sampling frequency. The data from each sensor were recorded as 60 segments with each segment contained 125 discrete time points. Each segment was considered as a sample observation from a sensor and a subject for an activity. There are totally $8 \times 60 = 480$ observations for each activity and a sensor. The purpose of the study is to classify the activities based on these observations each of which has 45 curves recorded by 45 sensors at 125 time points. We first apply the Fast Fourier Transformation to each of 45 curves to convert it from time domain to the frequency domain and get its spectrum curve. After filtering the higher frequency, we use the first 64 frequency points for each of 45 frequency curves. Then we apply wavelet transformation with 64 wavelet basis functions to the 64 frequency points for each curve and obtain 64 wavelet coefficients. In this way, for each observation, a vector with $64 \times 45 = 2880$ wavelet coefficients is obtained as the features to make classifications.

We focus on nine activities which can be divided into three groups. Group 1 includes three activities: walking in a parking lot, ascending and descending stairs; Group 2 has three activities: running on a treadmill with a speed of 8 km/h, exercising on a stepper and exercising on a cross trainer; Group 3 includes rowing, jumping and playing basketball. We will consider seven classification problems. In each of the first three problems, we consider the classification of the three activities in each of the three groups. In each of the next three problems, we combine any two of the thee groups and consider the classification of the six activities in the combined groups. The last problem is the classification of all nine activities. In each problem, for each class, we randomly select 30 observations as the training sample and all other 450 observations as the test sample. The procedure is repeated 50 times for each of the seven problems and the averages and standard deviations of misclassification rates are reported in Table 3.2.

Table 3.2 The averages and standard deviations of the misclassification rates (%) for the daily and sports activities data. For each classification problem shown in column 1, the first row is the average of misclassification rates, and the second row is the standard deviations.

Classes included	**SFDA-threshold**	**SFDA**	**RDA**	**PDA**
Group 1	0.23(0.23)	0.23(0.23)	1.94(1.91)	1.96(2.10)
Group 2	0.14(0.43)	0.14(0.44)	0.58(0.66)	0.21(0.58)
Group 3	0.12(0.07)	0.12(0.08)	0.58(1.08)	0.23(0.36)
Group 1+2	0.45(0.44)	0.46(0.43)	1.13(0.79)	2.39(1.52)
Group 1+3	1.50(0.84)	1.54(0.96)	1.92(0.99)	4.79(2.33)
Group 2+3	0.53(0.26)	0.54(0.24)	1.06(0.72)	0.80(0.37)
Group 1+2+3	1.63(0.60)	1.53(0.63)	1.78(0.65)	4.20(2.01)

3.5.2 Australian sign language data

The data is available in UCI Machine Learning Repository and the details of the experiments can be founded in Kadous [25]. This data consists of sample of Auslan (Australian Sign Language, it is the language used by the Deaf in Australia) signs. Twenty seven examples of each sign were captured from a native signer using high-quality position trackers and instrumented gloves. This was a two-hand system. For each hand, 11 time series curves were recorded simultaneously, including the measurements of x (left/right),y (up/down), z (backward/forward) positions, the direction of palm(is the palm pointing up or down?) and five finger bends. The frequency curve of each of the 22 curves were extracted by the Fast Fourier Transformation and then were transformed by 16 wavelet basis functions. Hence, for each sign, we obtained 352 features. We choose nine signs and divide them into three groups: Group 1 contains the three signs with meanings “innocent”, “responsible” and “not-my-problem”, respectively; Group 2 contains “read”, “write” and “draw”; Group 3 contains “hear”, “answer” and “think”. As in the previous example, we consider seven classification problems. For each class, we randomly choose 20 observations as the training sample and the other 7 as the test sample. The procedure is repeated 50 times and the averages and standard deviations of misclassification rates are reported in Table 3.3.

Table 3.3 The averages and standard deviations of the misclassification rates (%) for the Australian sign language data. For each classification problem shown in column 1, the first row is the average of misclassification rates, and the second row is the standard deviations.

Classes included	**SFDA-threshold**	**SFDA**	**RDA**	**PDA**
Group 1	0(0)	0(0)	1.24(2.32)	0.19(0.94)
Group 2	0(0)	0(0)	1.43(2.76)	4.57(5.61)
Group 3	1.24(2.11)	1.14(2.05)	3.05(3.94)	3.9(6.5)
Group 1+2	0.19(0.65)	0.62(1.26)	0.76(1.31)	3.81(2.93)
Group 1+3	0.81(1.71)	0.62(1.16)	1.29(1.94)	2.24(2.38)
Group 2+3	0.93(1.45)	1.06(1.68)	1.72(2.13)	5.16(4.34)
Group 1+2+3	0.73(1.02)	0.57(0.95)	1.14(1.11)	6.0(2.78)

Chapter 4

PROOFS OF THEOREMS

Here we provide the proofs of all our theoretical results. The proofs of some technical lemmas can be founded in Section 4.11.

4.1 Proof of Theorem 2.3.1

Theorem (Theorem 2.3.1)**.** *The solutions* $\{\hat{\lambda}_k, \hat{\gamma}_k : k \geq 1\}$ *of the successive optimization problems* (4.95) *and* (2.10) *exist for any* $\{\alpha_k > 0, k \geq 1\}$*. Moreover, for each* k*,* $\hat{\gamma}_k$ *has continuous second derivatives on* $[a, b]$ *and, on any subinterval* $\{[t_{(q-1)}, t_{(q)}], 1 \leq q \leq m-1\}$*, it can be written as a linear combination of the following at most* $4k$ *functions,*

$$\exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \qquad \exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)$$
$$\exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right), \quad \exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),$$

where $1 \leq j \leq k$*.*

Proof of Theorem 2.3.1. Consider the Sobolev space $W_2^2([a, b])$. Given a smoothing parameter $\alpha > 0$, for any $f, g \in W_2^2([a, b])$, we define an inner product,

$$\langle f, g\rangle_\alpha = \langle f, g\rangle + \alpha\,[f, g] \tag{4.1}$$

and the corresponding norm

$$\|f\|_\alpha = \sqrt{\langle f, f\rangle + \alpha\,[f, f]}$$

. Under this inner product, $W_2^2([a, b])$ becomes a reproducing kernel Hilbert space. For an introduction on the reproducing kernel Hilbert space, we refer the reader to Wahba [50]. For the reproducing kernel Hilbert space $(W_2^2([a, b], \langle\cdot\rangle_\alpha)$, the reproducing kernel has the following

form (see Thomas-Agnan [46]):

$$K^{(\alpha)}(s,t) = \begin{cases} \sum_{j=1}^{4}\sum_{k=1}^{4} l_{jk} b_j^{(\alpha)}(s) b_k^{(\alpha)}(t) & \text{for } t \le s \\ \\ \sum_{j=1}^{4}\sum_{k=1}^{4} l_{kj} b_j^{(\alpha)}(s) b_k^{(\alpha)}(t) & \text{for } t \ge s, \end{cases}$$

where

$$b_1^{(\alpha)}(t) = \exp\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right), \quad b_2^{(\alpha)}(t) = \exp\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right),$$
$$b_3^{(\alpha)}(t) = \exp\left(-\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right), \quad b_4^{(\alpha)}(t) = \exp\left(-\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_k^{\frac{1}{4}}}\right).$$

The coefficients matrix $(l_{jk})_{j,k=1}^4$ is nonsymmetric. Hence, $K^{(\alpha)}(s,t)$ has different forms for $s \le t$ and $s \ge t$. Given any fixed s, $K^{(\alpha)}(s,t)$ is a function of t with continuous second derivatives and is the linear combinations of $b_1^{(\alpha)}$, $b_2^{(\alpha)}$, $b_3^{(\alpha)}$ and $b_4^{(\alpha)}$ in the intervals $a \le t \le s$ and $s \le t \le b$ respectively, but the coefficients of the linear combinations may be different in these two intervals. For any given s, define $K_s^{(\alpha)}(t) = K^{(\alpha)}(s,t)$.

Because the numerators in

$$\max_{\substack{\|\gamma\| = 1, \langle \gamma, \hat{\gamma}_j \rangle = 0, \\ j = 1, \cdots, k-1}} \frac{\sum_{q=1}^{m}\sum_{l=1}^{m} \hat{\Sigma}_{ql}\gamma(t_{(q)})\gamma(t_{(l)})w_q w_l}{\|\gamma\|^2 + \alpha_k [\gamma,\gamma]}, \tag{4.2}$$

depend on γ only through their values at $\{t_q\}_{q=1}^m$, we denote the numerators by $f(\gamma(t_{(1)}), \cdots, \gamma(t_{(m)}))$ where f is a function on $\mathbb{R}^m$. By the properties of reproducing kernel, we have

$$\gamma(t_{(q)}) = \left\langle K_{t_{(q)}}^{(\alpha_1)}, \gamma \right\rangle_{\alpha}, \quad q = 1, \cdots, m.$$

We first consider $\hat{\gamma}_1$ which is the solution to (4.2) for $k = 1$. Let $\mathbf{V}_1$ be the finite-dimensional space spanned by $\{K_{t_{(q)}}^{(\alpha_1)} : 1 \le q \le m\}$ in $W_2^2([a,b])$ and let $\mathbf{V}_1^{\perp}$ be the orthogonal complement of $\mathbf{V}_1$ in $(W_2^2([a,b]), \langle\cdot\rangle_{\alpha_1})$. For any $\gamma \in W_2^2([a,b])$, it has unique

decomposition $\gamma = \gamma_1 + \gamma_2$, where $\gamma_1 \in \mathbf{V}_1$ and $\gamma_2 \in \mathbf{V}_1^{\perp}$. We have

$$\gamma(t_{(q)}) = \left\langle K_{t_{(q)}}^{(\alpha_1)}, \gamma \right\rangle_{\alpha_1} = \left\langle K_{t_{(q)}}^{(\alpha_1)}, \gamma_1 \right\rangle_{\alpha_1} + \left\langle K_{t_{(q)}}^{(\alpha_1)}, \gamma_2 \right\rangle_{\alpha_1} = \left\langle K_{t_{(q)}}^{(\alpha_1)}, \gamma_1 \right\rangle_{\alpha_1} = \gamma_1(t_{(q)}).$$

Hence,

$$\frac{f(\gamma(t_{(1)}), \cdots, \gamma(t_{(m)}))}{\|\gamma\|_{\alpha_1}^2} = \frac{f(\gamma_1(t_{(1)}), \cdots, \gamma_1(t_{(m)}))}{\|\gamma_1\|_{\alpha_1}^2 + \|\gamma_2\|_{\alpha_1}^2} \leq \frac{f(\gamma_1(t_{(1)}), \cdots, \gamma_1(t_{(m)})}{\|\gamma_1\|_{\alpha_1}^2},$$

where the second term and the third term are equal if and only if $\gamma_2 = 0$. Therefore,

$$\begin{aligned} &\max_{\gamma \in W_2^2([a,b]), \|\gamma\|=1} \frac{f(\gamma(t_{(1)}), \cdots, \gamma(t_{(m)})}{\|\gamma\|_{\alpha_1}^2} = \max_{\gamma \in W_2^2([a,b]), \|\gamma\| \neq 0} \frac{f(\gamma(t_{(1)}), \cdots, \gamma(t_{(m)})}{\|\gamma\|_{\alpha_1}^2} \\ &\leq \max_{\gamma_1 \in \mathbf{V}_1, \|\gamma_1\| \neq 0} \frac{f(\gamma_1(t_{(1)}), \cdots, \gamma_1(t_{(m)})}{\|\gamma_1\|_{\alpha_1}^2} = \max_{\gamma_1 \in \mathbf{V}_1, \|\gamma_1\|=1} \frac{f(\gamma_1(t_{(1)}), \cdots, \gamma_1(t_{(m)})}{\|\gamma_1\|_{\alpha_1}^2}. \end{aligned}$$

We have that the solution to the optimization problem

$$\max_{\gamma_1 \in \mathbf{V}_1, \|\gamma_1\|=1} \frac{f(\gamma_1(t_{(1)}), \cdots, \gamma_1(t_{(m)}))}{\|\gamma_1\|_{\alpha_1}^2} \tag{4.3}$$

is also the solution to

$$\max_{\gamma \in W_2^2([a,b]), \|\gamma\|=1} \frac{f(\gamma(t_{(1)}), \cdots, \gamma(t_{(m)}))}{\|\gamma\|_{\alpha_1}^2}$$

and does not have other solutions. Since $\mathbf{V}_1$ is a finite-dimensional space, the solution to (4.3) exists. Hence, the solution to (4.2) with $k = 1$, $\hat{\gamma}_1$, exists and belongs to $\mathbf{V}_1$. Since $\mathbf{V}_1$ is spanned by $\{K_{t_{(q)}}^{(\alpha_1)} : 1 \leq q \leq m\}$ in $W_2^2([a,b])$, it follows from the properties of $\{K_{t_{(q)}}^{(\alpha_1)} : 1 \leq q \leq m\}$ that $\hat{\gamma}_1$ has continuous second derivatives and is a linear combination of $b_1^{(\alpha_1)}$, $b_2^{(\alpha_1)}$, $b_3^{(\alpha_1)}$ and $b_4^{(\alpha_1)}$ on each of the intervals $[t_{(q)}, t_{(q+1)}]$, $q = 1, \cdots, m-1$.

Now suppose that $\hat{\gamma}_1, \cdots, \hat{\gamma}_{k-1}$ satisfy Theorem 2.3.1. We prove that $\hat{\gamma}_k$ also satisfies Theorem 2.3.1. Let $\mathbf{V}_k$ be the finite-dimensional space spanned by $\{K_{t_{(q)}}^{(\alpha_j)} : 1 \leq q \leq m, 1 \leq j \leq k\}$ in $W_2^2([a,b])$ and let $\mathbf{V}_k^{\perp}$ be the orthogonal complement of $\mathbf{V}_k$ in $(W_2^2([a,b], \langle \cdot \rangle_{\alpha_k})$. We have $\hat{\gamma}_1, \cdots, \hat{\gamma}_{k-1} \in \mathbf{V}_k$. For any $\gamma \in W_2^2([a,b])$, it has unique decomposition $\gamma = \gamma_1 + \gamma_2$,

where $\gamma_1 \in \mathbf{V}_k$ and $\gamma_2 \in \mathbf{V}_k^{\perp}$. Then the optimization problem (4.2),

$$\begin{aligned}
&\max_{\|\gamma\|=1, \langle\gamma,\hat{\gamma}_j\rangle=0,\ j=1,\cdots,k-1} \frac{f(\gamma(t_{(1)}),\cdots,\gamma(t_{(m)}))}{\|\gamma\|^2+\alpha_k[\gamma,\gamma]} \\
= &\max_{\gamma=\gamma_1+\gamma_2,\gamma_1\in\mathbf{V}_k,\gamma_2\in\mathbf{V}_k^{\perp},\ \langle\gamma_1,\hat{\gamma}_j\rangle=0, j=1,\cdots,k-1} \frac{f(\gamma_1(t_{(1)}),\cdots,\gamma_1(t_{(m)}))}{\|\gamma_1\|^2_{\alpha_k}+\|\gamma_2\|^2_{\alpha_k}} \\
\leq &\max_{\gamma_1\in\mathbf{V}_k,\langle\gamma_1,\hat{\gamma}_j\rangle=0,\ j=1,\cdots,k-1} \frac{f(\gamma_1(t_{(1)}),\cdots,\gamma_1(t_{(m)}))}{\|\gamma_1\|^2_{\alpha_k}}
\end{aligned}$$

Now by the same arguments as those for $\hat{\gamma}_1$, $\hat{\gamma}_k$ satisfies Theorem 2.3.1. □

4.2 Proof of Theorem 2.3.2

Theorem (Theorem 2.3.2)**.** *The solutions* $\{(\hat{\lambda}_k,\hat{\gamma}_k): k\geq 1\}$ *of the successive optimization problems* (2.14) *and* (2.15) *exist for any* $\{\alpha_k>0, k\geq 1\}$*. Moreover, for each* k*,* $\hat{\gamma}_k$ *has continuous second derivatives on* $[a,b]$ *and on the subinterval between any two adjacent pooled observation points, it can be written as a linear combination of the following at most* $4k$ *functions,*

$$\exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),\qquad \exp\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)$$
$$\exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\sin\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),\quad \exp\left(-\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right)\cos\left(\frac{t-a}{\sqrt{2}\alpha_j^{\frac{1}{4}}}\right),$$

where $1\leq j\leq k$*.*

Proof of Theorem 2.3.2. The proof is the same as the proof of Theorem 2.3.1. □

4.3 Proof of Theorem 2.3.4

Theorem (Theorem 2.3.4)**.** *Under Assumptions* $1-5$*, suppose that* $m, n \to \infty$*,* $\max_{1\le k\le K}\alpha_k \to 0$ *and*

$$\frac{\max_{1\le k\le K}\alpha_k}{\min_{1\le k\le K}\alpha_k} = O_p(1). \tag{4.4}$$

If the following is satisfied that for Case 1*,*

$$\frac{1}{\min_{1\le k\le K}\alpha_k}\left[\sqrt{\varpi(\delta_m)} + \delta_m + \sqrt{\frac{\delta_m}{n}}\right] \to 0$$

and for Case 2*,*

$$\frac{1}{\min_{1\le k\le K}\alpha_k}\left[\sqrt{\varpi(\frac{3\log m}{cm})} + \frac{\log m}{m} + \sqrt{\frac{\log m}{nm}}\right] \to 0,$$

then the estimators $\{(\hat{\lambda}_k, \hat{\gamma}_k) : 1 \le k \le K\}$ *are consistent.*

Proof of Theorem 2.3.4. Without loss of generality, we assume that the mean function $\mu(t)$ of $X(t)$ is zero. We still consider the Hilbert space $W_2^2([a,b])$ equipped with the inner product $\langle\cdot,\cdot\rangle_\alpha$ and the corresponding norm $\|\cdot\|_\alpha$ (see the definition (4.1)). For any $f \in W_2^2([a,b])$ we have the following Taylor expansion

$$f(t) = f(a) + (t-a)f'(a) + \int_a^b G(t,s)f''(s)ds, \tag{4.5}$$

where

$$G(t,s) = (t-s)_+ = \max\{(t-s), 0\}.$$

We first give upper bounds for $|f(a)|$ and $|f'(a)|$ in terms of $\|f\|_\alpha$.

Lemma 3.

$$|f(a)| \le \frac{1+\frac{\|G\|}{\sqrt{\alpha}}}{(\sqrt{2}-1)\sqrt{b-a}}\|f\|_\alpha, \quad |f'(a)| \le \frac{2(\sqrt{2}+2)(1+\frac{\|G\|}{\sqrt{\alpha}})}{(b-a)\sqrt{b-a}}\|f\|_\alpha,$$

where $\|G\| = \left[\int_a^b \int_a^b |G(s,t)|^2 ds dt\right]^{\frac{1}{2}}$.

The proof of Lemma 3 is in Section 4.11. Given any $\alpha > 0$, for any $\beta_1, \beta_2 \in W_2^2([a,b])$, by (4.97),

$$\begin{aligned}
&\sum_{q=1}^m \sum_{l=1}^m \hat{\Sigma}_{ql} \beta_1(t_{(q)}) \beta_2(t_{(l)}) w_q w_l \qquad (4.6)\\
=&\tilde{A}_{00}\beta_1(a)\beta_2(a) + \tilde{A}_{10}\beta_1'(a)\beta_2(a) + \tilde{A}_{01}\beta_1(a)\beta_2'(a) + \tilde{A}_{11}\beta_1'(a)\beta_2'(a)\\
&+ \beta_1(a)\int_a^b \tilde{\xi}_0(t)\beta_2''(t)dt + \beta_2(a)\int_a^b \tilde{\xi}_0(t)\beta_1''(t)dt + \beta_1'(a)\int_a^b \tilde{\xi}_1(t)\beta_2''(t)dt\\
&+ \beta_2'(a)\int_a^b \tilde{\xi}_1(t)\beta_1''(t)dt + \int_a^b\int_a^b \tilde{\Xi}(s,t)\beta_1''(s)\beta_2''(t)dsdt,
\end{aligned}$$

defines a bounded symmetric bilinear form in
$(W_2^2([a,b]), \langle\cdot,\cdot\rangle_\alpha)$, where

$$\begin{aligned}
&\tilde{A}_{00} = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql} w_q w_l, \quad \tilde{\Xi}(s,t) = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql} G(t_{(q)},s)G(t_{(l)},t) w_q w_l, \qquad (4.7)\\
&\tilde{A}_{01} = \tilde{A}_{10} = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql}(t_{(l)}-a) w_q w_l, \quad \tilde{\xi}_0(t) = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql} G(t_{(l)},t) w_q w_l\\
&\tilde{A}_{11} = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql}(t_{(q)}-a)(t_{(l)}-a) w_q w_l, \quad \tilde{\xi}_1(t) = \sum_{q=1}^m\sum_{l=1}^m \hat{\Sigma}_{ql}(t_{(q)}-a) G(t_{(l)},t) w_q w_l.
\end{aligned}$$

Hence, there is a unique bounded symmetric operator $\hat{R}_\alpha^{(m)}$ in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_\alpha)$, such that for any $\beta_1, \beta_2 \in W_2^2([a,b])$, (4.6) is equal to $\left\langle \beta_1, \hat{R}_\alpha^{(m)}\beta_2\right\rangle_\alpha$ (see Section 84 in Riesz and Sz.-Nagy [36]). Similarly, let

$$\hat{\Gamma}_n(s,t) = \frac{1}{n}\sum_{p=1}^n (X_p(s) - \bar{X}(s))(X_p(t) - \bar{X}(t)), a \le s,t \le b$$

be the sample covariance function, then there exists unique bounded symmetric operators $\hat{R}_\alpha$ and R_α in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_\alpha)$ for $\hat{\Gamma}_n$ and the true covariance function Γ respectively, such that for any $\beta_1, \beta_2 \in W_2^2([a,b])$,

$$\left\langle \beta_1, \hat{R}_\alpha\beta_2\right\rangle_\alpha = \left\langle \beta_1, \hat{\Gamma}_n\beta_2\right\rangle, \quad \langle \beta_1, R_\alpha\beta_2\rangle_\alpha = \langle \beta_1, \Gamma\beta_2\rangle, \qquad (4.8)$$

where $\hat{\Gamma}_n\beta_2$ denotes the function $\int_a^b \hat{\Gamma}_n(t,s)\beta_2(s)ds$. It is easy to see that

$$\|\hat{R}_\alpha - R_\alpha\|_\alpha \leq \|\hat{\Gamma}_n - \Gamma\|,$$

where $\|\cdot\|_\alpha$ and $\|\cdot\|$ are operator norms in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_\alpha)$ and L^2 space, respectively. By the central limit theorem in Hilbert space (see Chapter 10 in Ledoux and Talagrand [27]) and Assumption 1, we have

$$\|\hat{R}_\alpha - R_\alpha\|_\alpha \leq \|\hat{\Gamma}_n - \Gamma\| = O_p(\frac{1}{\sqrt{n}}). \tag{4.9}$$

We derive an upper bound for $\|\hat{R}_\alpha^{(m)} - \hat{R}_\alpha\|_\alpha$.

Lemma 4.

$$\|\hat{R}_\alpha^{(m)} - \hat{R}_\alpha\|_\alpha \leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\delta_m)}) + O_p(\delta_m) + O_p(\sqrt{\frac{\delta_m}{n}})\right], \tag{4.10}$$

in Case 1 (nonrandom case), and

$$\|\hat{R}_\alpha^{(m)} - \hat{R}_\alpha\|_\alpha \leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\frac{3\log m}{cm})}) + O_p(\frac{\log m}{m}) + O_p(\sqrt{\frac{\log m}{nm}})\right]. \tag{4.11}$$

in Case 2 (random case).

The proof of Lemma 4 is in Section 4.11. Define $\{\hat{\hat{\lambda}}_k, \hat{\hat{\gamma}}_k\}, k \geq 1$ to be the solutions to the following successive optimization problems:

$$\max_{\substack{\|\gamma\|=1, \left\langle\gamma, \hat{\hat{\gamma}}_j\right\rangle = 0, \\ j=1,\cdots,k-1}} \frac{\left\langle\gamma, \hat{R}_{\alpha_k}\gamma\right\rangle_{\alpha_k}}{\|\gamma\|_{\alpha_k}^2} = \max_{\substack{\|\gamma\|=1, \left\langle\gamma, \hat{\hat{\gamma}}_j\right\rangle = 0, \\ j=1,\cdots,k-1}} \frac{\left\langle\gamma, \hat{\Gamma}_n\gamma\right\rangle}{\|\gamma\|_{\alpha_k}^2},$$

and $\{\lambda_k^{[\alpha_k]}, \gamma_k^{[\alpha_k]}\}, k \geq 1$ to be the solutions to the following successive optimization problems:

$$\max_{\substack{\|\gamma\|=1, \left\langle \gamma, \gamma_j^{[\alpha_k]} \right\rangle = 0, \\ j=1,\cdots,k-1}} \frac{\langle \gamma, R_{\alpha_k}\gamma \rangle_{\alpha_k}}{\|\gamma\|_{\alpha_k}^2} = \max_{\substack{\|\gamma\|=1, \left\langle \gamma, \gamma_j^{[\alpha_k]} \right\rangle = 0, \\ j=1,\cdots,k-1}} \frac{\langle \gamma, \Gamma\gamma \rangle}{\|\gamma\|_{\alpha_k}^2}.$$

Note that $\{\lambda_k^{[\alpha_k]}, \gamma_k^{[\alpha_k]}\}, k \geq 1$ are nonrandom. Now we consider the first pair of estimates $\{\hat{\lambda}_1, \hat{\gamma}_1\}$ which are the first eigenvalue and eigenfunction of $\hat{R}_{\alpha_1}^{(m)}$. Since $\{\hat{\hat{\lambda}}_1, \hat{\hat{\gamma}}_1\}$ are the first eigenvalue and eigenfunction of $\hat{R}_{\alpha_1}$, by Corollary 4 in Section XI.9 of Dunford and Schwartz [14],

$$\begin{aligned} |\hat{\lambda}_1 - \hat{\hat{\lambda}}_1| &\leq \|\hat{R}_{\alpha_1}^{(m)} - \hat{R}_{\alpha_1}\|_{\alpha_1}, \\ |\hat{\hat{\lambda}}_1 - \lambda_1^{[\alpha_1]}| &\leq \|\hat{R}_{\alpha_1} - R_{\alpha_1}\|_{\alpha_1}. \end{aligned} \tag{4.12}$$

Hence, under the conditions in Theorem 2.3.4, it follows from (4.9), Lemma 4, (4.12) and the following Lemma 5 that $\hat{\lambda}_1$ is consistent.

Lemma 5. *Under Assumptions 1-3, for any* $1 \leq k \leq K$ *and* $0 \leq \alpha \leq \alpha_0$, $\left[\gamma_1^{[\alpha_1]}, \gamma_1^{[\alpha_1]}\right] \leq 2kL_k^2$ *and*

$$\begin{aligned} 0 \leq \lambda_k - \lambda_k^{[\alpha]} &\leq \sqrt{2}\sqrt{k}L_k^2\lambda_k\alpha + o(\alpha), \\ \|\gamma_k^{[\alpha]} - \gamma_k\| &\leq \sqrt{\alpha}\sqrt{\frac{4\sqrt{2}\sqrt{k}L_k^2\lambda_k}{\lambda_k - \lambda_{k+1}}} + o(\sqrt{\alpha}), \end{aligned}$$

where $L_k = \max_{1\leq j\leq k}\sqrt{[\gamma_j, \gamma_j]}$ *and*

$$\alpha_0 = \min_{1\leq k\leq K}\left\{\min\left\{\frac{\sqrt{1+\frac{2k(\lambda_{k-1}-\lambda_k)^2}{(k-1)\lambda_k\|\Gamma\|}}-1}{2kL_k^2}, \frac{\lambda_k - \lambda_{k+1}}{(8\sqrt{k}+16k)L_k^2\lambda_k}, \right.\right.$$
$$\left.\left.\frac{(\lambda_{k-1}-\lambda_k)\left\{1+\frac{2\|\Gamma\|}{\lambda_k-\lambda_{k+1}}\right\}^{-\frac{1}{2}}}{4\sqrt{2k(k-1)}L_k^2\lambda_k}\right\}\right\}. \tag{4.13}$$

Lemma 5 is just Theorem 4.1 in Qi and Zhao [31]. In order to compute $\|\hat{\gamma}_1 - \hat{\hat{\gamma}}_1\|_{\alpha_1}$,

we define $\hat{E}_1$ and $\hat{\hat{E}}_1$ to be the orthogonal projections onto the space spanned by $\hat{\gamma}_1$ and $\hat{\hat{\gamma}}_1$, respectively. When $\|\hat{R}_{\alpha_1}^{(m)} - \hat{R}_{\alpha_1}\|_{\alpha_1}$ is small enough, we can carry a similar calculation as that in Section 2.1.1 of Dauxois et al. [12] and obtain the inequality

$$\|\hat{E}_1 - \hat{\hat{E}}_1\|_\alpha \leq O_p(\|\hat{R}_{\alpha_1}^{(m)} - \hat{R}_{\alpha_1}\|_{\alpha_1}).$$

Define

$$\hat{e}_1 = \frac{\hat{\gamma}_1}{\|\hat{\gamma}_1\|_{\alpha_1}}, \quad \hat{\hat{e}}_1 = \frac{\hat{\hat{\gamma}}_1}{\|\hat{\hat{\gamma}}_1\|_{\alpha_1}}.$$

Then we have

$$\|\hat{e}_1\| = \frac{1}{\|\hat{\gamma}_1\|_{\alpha_1}}, \quad \|\hat{\hat{e}}_1\| = \frac{1}{\|\hat{\hat{\gamma}}_1\|_{\alpha_1}},$$

since $\|\hat{\gamma}_1\| = 1$ and $\|\hat{\hat{\gamma}}_1\| = 1$. Now

$$\begin{aligned}
&\|\hat{e}_1 - \hat{\hat{e}}_1\|_{\alpha_1}^2 = \|\hat{e}_1\|_{\alpha_1}^2 - 2\left\langle \hat{e}_1, \hat{\hat{e}}_1 \right\rangle_{\alpha_1} + \|\hat{\hat{e}}_1\|_{\alpha_1}^2 = 2\left(1 - \left\langle \hat{e}_1, \hat{\hat{e}}_1 \right\rangle_{\alpha_1}\right) \\
\leq & 2\left(1 - \left\langle \hat{e}_1, \hat{\hat{e}}_1 \right\rangle_{\alpha_1}\right)\left(1 + \left\langle \hat{e}_1, \hat{\hat{e}}_1 \right\rangle_{\alpha_1}\right) = 2|\left\langle \hat{e}_1, \hat{\hat{e}}_1 \right\rangle_{\alpha_1}^2 - 1| \\
= & 2\left|\left\langle \hat{\hat{e}}_1, (\hat{E}_1 - \hat{\hat{E}}_1)\hat{\hat{e}}_1 \right\rangle_{\alpha_1}\right| \leq 2\|\hat{E}_1 - \hat{\hat{E}}_1\|_{\alpha_1} \\
\leq & O_p(\|\hat{R}_{\alpha_1}^{(m)} - \hat{R}_{\alpha_1}\|_{\alpha_1}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\hat{\gamma}_1 - \hat{\hat{\gamma}}_1\|_{\alpha_1} = \|\frac{\hat{e}_1}{\|\hat{e}_1\|} - \frac{\hat{\hat{e}}_1}{\|\hat{\hat{e}}_1\|}\|_{\alpha_1} = \|\frac{\hat{e}_1}{\|\hat{e}_1\|} - \frac{\hat{e}_1}{\|\hat{\hat{e}}_1\|} + \frac{\hat{e}_1}{\|\hat{\hat{e}}_1\|} - \frac{\hat{\hat{e}}_1}{\|\hat{\hat{e}}_1\|}\|_{\alpha_1} \\
\leq & \frac{2\|\hat{e}_1 - \hat{\hat{e}}_1\|_{\alpha_1}}{\|\hat{\hat{e}}_1\|} = \|\hat{\hat{\gamma}}_1\|_{\alpha_1}\|\hat{e}_1 - \hat{\hat{e}}_1\|_{\alpha_1} \leq \|\hat{\hat{\gamma}}_1\|_{\alpha_1} O_p(\|\hat{R}_{\alpha_1}^{(m)} - \hat{R}_{\alpha_1}\|_{\alpha_1}^{\frac{1}{2}}).
\end{aligned} \tag{4.14}$$

A similar calculation leads to the following inequality

$$\|\hat{\hat{\gamma}}_1 - \gamma_1^{[\alpha_1]}\|_{\alpha_1} \leq \|\gamma_1^{[\alpha_1]}\|_{\alpha_1} O_p(\|\hat{R}_{\alpha_1} - R_{\alpha_1}\|_{\alpha_1}^{\frac{1}{2}}). \tag{4.15}$$

By Lemma 5, $\left[\gamma_1^{[\alpha_1]}, \gamma_1^{[\alpha_1]}\right] = O(1)$ as $\alpha_1 \to 0$, hence, $\|\gamma_1^{[\alpha_1]}\|_{\alpha_1} \to 1$. Then by (4.15), $\|\hat{\hat{\gamma}}_1 - \gamma_1^{[\alpha_1]}\|_{\alpha_1} \to 0$ and hence $\|\hat{\hat{\gamma}}\|_{\alpha_1} \to 1$ in probability. Now by (4.14), $\|\hat{\gamma}_1 - \hat{\hat{\gamma}}_1\|_{\alpha_1} \to 0$ in probability. Moreover, by Lemma 5,

$$\|\gamma_1^{[\alpha_1]} - \gamma_1\| \leq O(\sqrt{\alpha_1}).$$

Hence, $\hat{\gamma}_1$ is consistent. Then for any $1 \leq k \leq K$, the consistent results follow from the following lemma by induction.

Lemma 6. *Under the conditions in Theorem 2.3.4, if for all* $1 \leq j \leq k-1$,

$$\|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\|_{\alpha_j} \to 0, \|\hat{\hat{\gamma}}_j - \gamma_j^{[\alpha_j]}\|_{\alpha_j} \to 0, \|\gamma_j^{[\alpha_j]} - \gamma_j\| \to 0, \|\gamma_j^{[\alpha_j]}\|_{\alpha_j} \to 1, \tag{4.16}$$

then (4.16) *is also true for* k, *moreover,* $\hat{\lambda}_k$ *is consistent.*

□

4.4 Proof of Theorem 2.3.5

Theorem (Theorem 2.3.5)**.** *Under Assumptions* $1-3$ *and* $6-7$, *suppose that* $n \to \infty$, $\max_{1\leq k\leq K} \alpha_k \to 0$ *and*

$$\frac{\max_{1\leq k\leq K} \alpha_k}{\min_{1\leq k\leq K} \alpha_k} = O_p(1).$$

If the following is satisfied

$$\frac{1}{\min_{1\leq k\leq K} \alpha_k}\left[n^{-\frac{1}{2}}(\eta^{-1} + \eta_g^{-\frac{1}{2}-\epsilon}) + \eta_g^{\frac{3}{4}-\epsilon}\right] \to 0,$$

for some $\epsilon > 0$, *then the estimators* $\{(\hat{\lambda}_k, \hat{\gamma}_k) : 1 \leq k \leq K\}$ *are consistent.*

Proof of Theorem 2.3.5. Under the assumptions in this theorem, we have

$$\begin{aligned}
&\sup_{a\le t\le b}|\hat{\mu}(t)-\mu(t)| = O_p(\frac{1}{\sqrt{n}\eta_\mu}),\\
&\sup_{a\le t\le b}\left|\hat{h}(t)-h(t)\right| = O_p(n^{-\frac{1}{2}}\eta_g^{-\frac{1}{2}-\epsilon}+\eta_g^{\frac{3}{4}-\epsilon}),
\end{aligned} \tag{4.17}$$

where ϵ is any positive constant. For the proof of the first equality in (4.17), we refer the reader to Theorem 1 in Yao et al. [54] or the proofs of the main results in Hall et al. [20]. The second equality in (4.17) is Theorem 8.1 in Silverman [40].

Similar to the proof of Theorem 2.3.4, for any $\alpha>0$, we define a bounded symmetric operator $\hat{S}_\alpha$ in $(W_2^2([a,b]),\langle\cdot,\cdot\rangle_\alpha)$, such that for any $\beta_1(t),\beta(t)_2\in W_2^2([a,b])$,

$$\begin{aligned}
&\left\langle\beta_1,\hat{S}_\alpha\beta_2\right\rangle_\alpha\\
=&\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{\beta_1(t_{pq})(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{\beta_2(t_{pl})(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})}\\
=&\tilde{B}_{00}\beta_1(a)\beta_2(a)+\tilde{B}_{10}\beta_1'(a)\beta_2(a)+\tilde{B}_{01}\beta_1(a)\beta_2'(a)+\tilde{B}_{11}\beta_1'(a)\beta_2'(a)\\
&+\beta_1(a)\int_a^b\tilde{\psi}_0(t)\beta_2''(t)dt+\beta_2(a)\int_a^b\tilde{\psi}_0(t)\beta_1''(t)dt+\beta_1'(a)\int_a^b\tilde{\psi}_1(t)\beta_2''(t)dt\\
&+\beta_2'(a)\int_a^b\tilde{\psi}_1(t)\beta_1''(t)dt+\int_a^b\int_a^b\tilde{\Psi}(s,t)\beta_1''(s)\beta_2''(t)dsdt,
\end{aligned}$$

defines a bounded symmetric bilinear form in $(W_2^2([a,b]),\langle\cdot,\cdot\rangle_\alpha)$, where

$$\begin{aligned}
\tilde{B}_{00}&=\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})},\\
\tilde{B}_{01}=\tilde{B}_{10}&=\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{(t_{pq}-a)(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})}\\
\tilde{B}_{11}&=\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{(t_{pq}-a)(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{(t_{pl}-a)(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})},\\
\psi_0(t)&=\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{G(t_{pq},t)(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})},\\
\psi_1(t)&=\frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{G(t_{pq},t)(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{(t_{pl}-a)(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})},
\end{aligned} \tag{4.18}$$

$$\Psi(t) = \frac{1}{n'}\sum_{p=1}^{n}\chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\frac{G(t_{pq},t)(Y_{pq}-\hat{\mu}(t_{pq}))}{\hat{h}(t_{pq})}\cdot\frac{G(t_{pl},s)(Y_{pl}-\hat{\mu}(t_{pl}))}{\hat{h}(t_{pl})}.$$

Similarly, we define a bounded symmetric operator S_α in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_\alpha)$, such that for any $\beta_1(t), \beta(t)_2 \in W_2^2([a,b])$,

$$\langle\beta_1, S_\alpha\beta_2\rangle_\alpha = \langle\beta_1, \Gamma\beta_2\rangle.$$

Lemma 7. *Under the assumptions in the theorem,*

$$\|\hat{S}_\alpha - S_\alpha\|_\alpha \leq \frac{1}{\alpha}\left[O_p(\frac{1}{\sqrt{n}\eta_\mu}) + O_p(n^{-\frac{1}{2}}\eta_g^{-\frac{1}{2}-\epsilon} + \eta_g^{\frac{3}{4}-\epsilon}) + O_p(\frac{1}{\sqrt{n}})\right].$$

as $\alpha \to 0$ and $n \to \infty$.

Then by the same arguments as in the proof of Theorem 2.3.4, the theorem follows Lemma 7. □

4.5 Proof of Theorem 3.3.1

We first provide several lemmas whose proofs can be found in Section 4.11 Section.

Lemma 8. *For any $1 \leq i \neq j \leq K-1$, we have*

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_{ij} = \mathbf{D}\boldsymbol{\delta}_{ij}, where \boldsymbol{\delta}_{ij} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i.$$

Lemma 9. *Define a $K \times K$ matrix,*

$$\boldsymbol{\Delta} = \mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U}.$$

Under Condition 2, $\boldsymbol{\Delta}$ has $K-1$ positive eigenvalues denoted by

$$\lambda_{min}^{+}(\boldsymbol{\Delta}) = \lambda_{K-1}(\boldsymbol{\Delta}) \leq \cdots \leq \lambda_2(\boldsymbol{\Delta}) \leq \lambda_1(\boldsymbol{\Delta}) = \lambda_{max}(\boldsymbol{\Delta}).$$

Then we have

$$\lambda_1(\boldsymbol{\Delta}) = K\lambda_1(\boldsymbol{\Xi}), \cdots, \lambda_{K-1}(\boldsymbol{\Delta}) = K\lambda_{K-1}(\boldsymbol{\Xi}),$$

and

$$2Kc_1 \leq 2\lambda^{+}_{min}(\boldsymbol{\Delta}) \leq \min_{i\neq j}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$
$$\leq \max_{i\neq j}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \leq 2\lambda_{max}(\boldsymbol{\Delta}).$$

Lemma 10. *Suppose that Condition 1 holds, $p \geq 2$, $K \leq p+1$ and $K\log p/n \to 0$ as $n \to \infty$. Then for any $M > 0$, we can find a constant C large enough and independent of n, p and K, such that*

$$P\left(\max_{1\leq j\leq K} \|\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \leq p^{-M}, \tag{4.19}$$

for all n large enough.

Now we prove the main theorems.

Theorem (Theorem 3.3.1)**.** *Suppose that Condition 1 holds, $p \geq 2$, $K \leq p+1$ and $K\log p/n \to 0$ as $n \to \infty$. Then for any $M > 0$, we can find C large enough and independent of n, p and K such that*

$$P\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \leq p^{-M},$$
$$P\left(\|\widehat{\mathbf{B}} - \mathbf{B}\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \leq p^{-M}$$

for all large enough n.

Proof of Theorem 3.3.1. We only consider the case that $(n_1, n_2, \cdots, n_K)$ follows a multinomial distribution. For the nonrandom case, a similar argument can prove the theorem. Let $\widehat{\sigma}_{kl}$ and σ_{kl} be the (k,l) element of $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$, respectively, $1 \leq k, l \leq p$. By the definition of

$\widehat{\mathbf{\Sigma}}$ in (3.7),

$$\left(\frac{n-K}{n}\right)|\widehat{\sigma}_{kl}-\sigma_{kl}| = \left|\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}^k-\bar{\mathbf{x}}_i^k)(\mathbf{x}_{ij}^l-\bar{\mathbf{x}}_i^k)-(1-\frac{K}{n})\sigma_{kl}\right| \tag{4.20}$$

$$\begin{aligned}
=&\left|\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}^k-\boldsymbol{\mu}_i^k)(\mathbf{x}_{ij}^l-\boldsymbol{\mu}_i^l)-\frac{1}{n}\sum_{i=1}^{K}n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)-(1-\frac{K}{n})\sigma_{kl}\right| \\
\leq&\frac{1}{n}\left|\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\mathbf{x}_{ij}^k-\boldsymbol{\mu}_i^k)(\mathbf{x}_{ij}^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}\right]\right|+\frac{1}{n}\left|\sum_{i=1}^{K}\left[n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}\right]\right|,
\end{aligned}$$

where $\mathbf{x}_{ij}^k$ and $\boldsymbol{\mu}_i^k$ denotes the k-th coordinate of $\mathbf{x}_{ij}$ and $\boldsymbol{\mu}_i$, respectively. Note that both $\mathbf{x}_{ij}-\boldsymbol{\mu}_i$ and $\sqrt{n_i}(\bar{\mathbf{x}}_i-\boldsymbol{\mu}_i)$ have the distributions $N(\mathbf{0},\mathbf{\Sigma})$. By Lemma A.3. in Bickel and Levina [8], we have

$$P\left(\left|\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\mathbf{x}_{ij}^k-\boldsymbol{\mu}_i^k)(\mathbf{x}_{ij}^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}\right]\right|>n\nu_1\right)\leq C_1\exp(-C_2n\nu_1^2),$$

and hence

$$P\left(\max_{k,l}\left|\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\mathbf{x}_{ij}^k-\boldsymbol{\mu}_i^k)(\mathbf{x}_{ij}^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}\right]\right|>n\nu_1\right)\leq C_1p^2\exp(-C_2n\nu_1^2), \tag{4.21}$$

for any ν_1 less than a constant δ, where C_1, C_2 and δ are constants only depending on the upper bound c_0 of the eigenvalues of $\mathbf{\Sigma}$. For any $C>0$, taking $\nu_1=C\sqrt{\frac{\log p}{n}}$, we can obtain

$$P\left(\max_{k,l}\frac{1}{n}\left|\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[(\mathbf{x}_{ij}^k-\boldsymbol{\mu}_i^k)(\mathbf{x}_{ij}^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}\right]\right|>C\sqrt{\frac{\log p}{n}}\right)\leq C_1p^2p^{-C^2C_2}. \tag{4.22}$$

For any $M>0$, we can find C large enough such that the right hand side of (4.22) is less than p^{-M} for all $p>1$. For the second term in in the last line of (4.20), define $Z_{ik}=\sqrt{n_i}(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)$,

for any $1 \leq i \leq K$ and $1 \leq k \leq p$. Then

$$\begin{aligned}\sum_{i=1}^{K}\left[n_i(\bar{\mathbf{x}}_i^k - \boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l - \boldsymbol{\mu}_i^l) - \sigma_{kl}\right] &= \sum_{i=1}^{K}[Z_{ik}Z_{il} - \sigma_{kl}] \\ &= \frac{1}{4}\sum_{i=1}^{K}[(Z_{ik}+Z_{il})^2 - (\sigma_{kk}+\sigma_{ll}+2\sigma_{kl})] \\ &-\frac{1}{4}\sum_{i=1}^{K}[(Z_{ik}-Z_{il})^2 - (\sigma_{kk}+\sigma_{ll}-2\sigma_{kl})].\end{aligned} \tag{4.23}$$

We will derive the upper bound for the first sum in the last line of (4.23). Let

$$Y_i = (Z_{ik}+Z_{il})^2/(\sigma_{kk}+\sigma_{ll}+2\sigma_{kl}) - 1.$$

Then $Y_1, \cdots, Y_K$, are i.i.d. random variables with the distribution $\chi_1^2 - 1$. We will apply the Bernstein's inequality (see Lemma 2.2.11 in Van Der Vaart and Wellner [48] or page 855 of Shorack and Wellner [39]) for unbounded random variables to $Y_1 + \cdots + Y_K$. We first verify the moment condition required by the Bernstein's inequality. For any positive integer $m \geq 3$, noting that $Var(Y_i) = 2$, we have

$$\begin{aligned}&E[|Y_i|^m] = E[|\chi_1^2 - 1|^m] \leq 2^{m-1}\left(E[|\chi_1^2|^m] + 1\right) \leq 2^m E[|\chi_1^2|^m] \\ =&2^m[1\cdot 3\cdot 5\cdots(2m-1)] \leq 2^m[2\cdot 4\cdot 6\cdots(2m)] = 2^m[2^m m!] \\ =&4^m m! Var(Y_i)/2 \leq D^{m-2}m!Var(Y_i)/2,\end{aligned}$$

where $D = 64$. When $m = 2$,

$$E[|Y_i|^m] = Var(Y_i) = D^{m-2}m!Var(Y_i)/2.$$

Hence, the moment condition for the Bernstein's inequality holds for Y_i. Now by the

Bernstein's inequality, for any $\nu_2 > 0$, let $x = n\nu_2/(\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl})$, then we have

$$\begin{aligned}
&P\left(\left|\sum_{i=1}^{K}[(Z_{ik} + Z_{il})^2 - (\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl})]\right| > n\nu_2\right) \\
=&P\left(|Y_1 + \cdots + Y_K| > \frac{n\nu_2}{\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl}}\right) \\
=&P\left(|Y_1 + \cdots + Y_K| > x\right) \\
\leq&2\exp\left(-\frac{1}{2}\frac{x^2}{KVar(Y_1) + Dx}\right) \\
=&2\exp\left(-\frac{1}{2}\frac{x^2}{2K + Dx}\right).
\end{aligned}$$

Note that $-x^2/(2K + Dx)$ is a decreasing function for $x > 0$, and that

$$\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl} = \mathbf{v}_{kl}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{v}_{kl} \leq \lambda_{\max}(\mathbf{\Sigma})\|\mathbf{v}_{kl}\|_2^2 = 2\lambda_{\max}(\mathbf{\Sigma}) \leq 2c_0, \tag{4.24}$$

where $\mathbf{v}_{kl}$ is the p-dimensional vector with all coordinates equal to 0 except the k-th and l-th coordinates which are equal to 1 and the last inequality is due to Condition 1 (b). Then we have $x \geq n\nu_2/(2c_0)$ and

$$\exp\left(-\frac{1}{2}\frac{x^2}{2K + Dx}\right) \leq \exp\left(-\frac{1}{2}\frac{(n\nu_2)^2}{8c_0^2K + 2c_0Dn\nu_2}\right).$$

Hence,

$$\begin{aligned}
&P\left(\max_{k,l}\left|\sum_{i=1}^{K}[(Z_{ik} + Z_{il})^2 - (\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl})]\right| > n\nu_2\right) \\
\leq&2p^2\exp\left(-\frac{1}{2}\frac{(n\nu_2)^2}{8c_0^2K + 2c_0Dn\nu_2}\right).
\end{aligned}$$

For any $C > 0$, let $\nu_2 = C\sqrt{\frac{\log p}{n}}$. Since $\log p/n \to 0$, when n is large enough, we have

$\log p \leq n$, and hence

$$\begin{aligned}
&2p^2 \exp\left(-\frac{1}{2}\frac{(n\nu_2)^2}{8c_0^2K + 2c_0Dn\nu_2}\right) \\
=&2p^2 \exp\left(-\frac{1}{2}\frac{C^2 n\log p}{8c_0^2K + 2c_0DC\sqrt{n\log(p)}}\right) \\
\leq&2p^2 \exp\left(-\frac{1}{2}\frac{C^2 n\log p}{8c_0^2n + 2c_0DC\sqrt{nn}}\right) \\
=&2p^2 p^{-\frac{1}{2}\frac{C^2}{8c_0^2+2c_0DC}}, \qquad (4.25)
\end{aligned}$$

where we use $K \leq n$ as n is large enough due to the condition $K\log p/n \to 0$. Now for any $M > 0$, we can find C large enough such that the right hand side of (4.25) is less than p^{-M} for all $p \geq 2$. We can obtain the similar result for the second sum in the last line of (4.23). Hence, for any $M > 0$, we can find C large enough such that

$$P\left(\max_{k,l}\frac{1}{n}\left|\sum_{i=1}^{K}\left[n_i(\bar{\mathbf{x}}_i^k - \boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l - \boldsymbol{\mu}_i^l) - \sigma_{kl}\right]\right| > C\sqrt{\frac{\log p}{n}}\right) \leq p^{-M}. \qquad (4.26)$$

It follows from (4.20), (4.22) and (4.26), for any $M > 0$, we can find C large enough and independent of n, p and K, such that

$$\begin{aligned}
&P\left(\max_{k,l}|\widehat{\sigma}_{kl} - \sigma_{kl}| > C\sqrt{\frac{K\log p}{n}}\right) \\
\leq&P\left(\max_{k,l}|\widehat{\sigma}_{kl} - \sigma_{kl}| > C\sqrt{\frac{\log p}{n}}\right) \leq p^{-M},
\end{aligned}$$

for all n large enough, where we use $K/n \to 0$ due to the condition $K\log p/n \to 0$.

In order to estimate $\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty$, we first calculate the (k,l) element of $\widehat{\mathbf{B}}-\mathbf{B}$.

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i^k-\bar{\mathbf{x}}^k)(\bar{\mathbf{x}}_i^l-\bar{\mathbf{x}}^l)-\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^{K} n_i\bar{\mathbf{x}}_i^k\bar{\mathbf{x}}_i^l-\bar{\mathbf{x}}^k\bar{\mathbf{x}}^l-\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)+\frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)\boldsymbol{\mu}_i^l+\frac{1}{n}\sum_{i=1}^{K} n_i\boldsymbol{\mu}_i^k(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)\right. \\
&\left.+\frac{1}{n}\sum_{i=1}^{K} n_i\boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l-\bar{\mathbf{x}}^k\bar{\mathbf{x}}^l-\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l\right| \\
\le&\frac{K}{n}|\sigma_{kl}|+\left|\frac{1}{n}\sum_{i=1}^{K}[n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)-\sigma_{kl}]\right|+\left|\frac{1}{n}\sum_{i=1}^{K} n_i(\bar{\mathbf{x}}_i^k-\boldsymbol{\mu}_i^k)\boldsymbol{\mu}_i^l\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{K} n_i\boldsymbol{\mu}_i^k(\bar{\mathbf{x}}_i^l-\boldsymbol{\mu}_i^l)\right|+\left|\sum_{i=1}^{K}\left(\frac{n_i}{n}-\frac{1}{K}\right)\boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l\right|+|\bar{\mathbf{x}}^k\bar{\mathbf{x}}^l| \\
=&\frac{K}{n}|\sigma_{kl}|+I+II+III+IV+V \\
\le&\frac{K}{n}c_0+I+II+III+IV+V. \qquad (4.27)
\end{aligned}
$$

Note that the term I is just that in (4.26). By Condition 1 (b),

$$\max_{k,l} II \le \frac{1}{n}\sum_{i=1}^{K} n_i \max_{1\le j\le K}\|\bar{\mathbf{x}}_j-\boldsymbol{\mu}_j\|_\infty \max_{1\le j\le K}\|\boldsymbol{\mu}_j\|_\infty \le c_0 \max_{1\le j\le K}\|\bar{\mathbf{x}}_j-\boldsymbol{\mu}_j\|_\infty,$$

which combined with Lemma 10 gives that for any $M>0$, we can find a constant C large enough and independent of n, p and K, such that

$$P\left(\max_{k,l} II > C\sqrt{\frac{K\log p}{n}}\right)\le p^{-M}, \qquad (4.28)$$

for all n large enough. The same bound can be obtained for the term III. As to the term

IV, by Lemma 1 and Condition 1 (b), for any $M > 0$, we can find a constant C such that

$$\begin{aligned} P\left(\max_{1\le k,l\le p} \left| \sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i^k \boldsymbol{\mu}_i^l \right| > C\sqrt{\frac{K\log p}{n}} \right) \\ \le P\left(Kc_0^2 \max_{1\le i\le K} \left| \frac{n_i}{n} - \frac{1}{K} \right| > C\sqrt{\frac{K\log p}{n}} \right) \\ = P\left(\max_{1\le i\le K} \left| \frac{n_i}{n} - \frac{1}{K} \right| > \frac{C}{c_0^2}\sqrt{\frac{\log p}{Kn}} \right) \le p^{-M} \end{aligned} \tag{4.29}$$

for all n large enough. For the term V, because

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} \mathbf{x}_{ij} = \bar{\mathbf{y}} + \frac{1}{n}\sum_{i=1}^{K} n_i \boldsymbol{\mu}_i = \bar{\mathbf{y}} + \sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i, \tag{4.30}$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \boldsymbol{\mu}_i)$ and the last equality is due to (3.2). Then we have

$$\begin{aligned} \bar{\mathbf{x}}^k\bar{\mathbf{x}}^l = \bar{\mathbf{y}}^k\bar{\mathbf{y}}^l + \sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i^l \bar{\mathbf{y}}^k + \sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i^k \bar{\mathbf{y}}^l \\ + \left[\sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i^l \right] \left[\sum_{i=1}^{K} \left(\frac{n_i}{n} - \frac{1}{K} \right) \boldsymbol{\mu}_i^k \right]. \end{aligned} \tag{4.31}$$

We will consider the four terms on the right hand side of (4.31), respectively. Note that $\bar{\mathbf{y}}$ has the normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}/n$. $\bar{\mathbf{y}}^k\bar{\mathbf{y}}^l = (\bar{\mathbf{y}}^k + \bar{\mathbf{y}}^l)^2/4 - (\bar{\mathbf{y}}^k - \bar{\mathbf{y}}^l)^2/4$. Note that $\bar{\mathbf{y}}^k + \bar{\mathbf{y}}^l$ has a normal distribution with mean zero and variance $(\sigma_{kk} + \sigma_{ll} + 2\sigma_{kl})/n \le 2c_0/n$ by (4.24). By the same arguments as in the proof of Lemma 10, we can show that for any $M > 0$, we can find C large enough such that

$$P\left(\max_{k,l} \frac{1}{\sqrt{Kc_0^2}} \left| \frac{\bar{\mathbf{y}}^k + \bar{\mathbf{y}}^l}{2} \right| > C\sqrt{\frac{\log p}{n}} \right) \le p^{-M},$$

for all n large enough. Since when n is large enough, we have $Cc_0\sqrt{K\log p/n} \leq 1$, then,

$$\begin{aligned}
&P\left(\max_{k,l}\left|\frac{(\bar{\mathbf{y}}^l+\bar{\mathbf{y}}^l)^2}{4}\right| > Cc_0\sqrt{\frac{K\log p}{n}}\right)\\
\leq &P\left(\max_{k,l}\left|\frac{(\bar{\mathbf{y}}^l+\bar{\mathbf{y}}^l)^2}{4}\right| > \left[Cc_0\sqrt{\frac{K\log p}{n}}\right]^2\right)\\
= &P\left(\max_{k,l}\frac{1}{\sqrt{Kc_0^2}}\left|\frac{\bar{\mathbf{y}}^k+\bar{\mathbf{y}}^l}{2}\right| > C\sqrt{\frac{\log p}{n}}\right) \leq p^{-M}. \qquad (4.32)
\end{aligned}$$

and the same inequality for $(\bar{\mathbf{y}}^k-\bar{\mathbf{y}}^l)^2/4$. Therefore, we have that for any $M>0$, we can find C large enough and independent of n, p and K, such that

$$P\left(\max_{k,l}\left|\bar{\mathbf{y}}^k\bar{\mathbf{y}}^l\right| > C\sqrt{\frac{K\log p}{n}}\right) \leq p^{-M}. \qquad (4.33)$$

Using the same arguments as in (4.29), we can obtain the same probability bounds for the last three terms on the right hand side of (4.31). Then by combining (4.27)-(4.33) and using Lemmas 1 and 10, for any $M>0$, we can find C large enough such that

$$\begin{aligned}
&P\left(\|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty > C\sqrt{\frac{K\log p}{n}}\right) \qquad (4.34)\\
= &P\left(\max_{k,l}\left|\frac{1}{n}\sum_{i=1}^K n_i(\bar{\mathbf{x}}_i^k-\bar{\mathbf{x}}^k)(\bar{\mathbf{x}}_i^l-\bar{\mathbf{x}}^l) - \frac{1}{K}\sum_{i=1}^K \boldsymbol{\mu}_i^k\boldsymbol{\mu}_i^l\right| > C\sqrt{\frac{K\log p}{n}}\right) \leq p^{-M},
\end{aligned}$$

for all n large enough.

□

4.6 Proof of Theorem 3.3.4

Theorem (Theorem 3.3.4)**.** *Suppose that $K=2$ and Conditions 1-2 hold. If $s_n\to 0$ and $\|\boldsymbol{\alpha}_1\|_1^2 s_n \to 0$ as $n,p\to\infty$, then for all large enough n, we have, in Ω_n,*

$$\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \leq 6\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0, \quad \|\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1\|_2^2 \leq C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n, \quad \|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1\|_2^2 \leq c_0C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n, \qquad (4.35)$$

where C_5 is a constant independent of n and p, and c_0 is the constant in Condition 1 (b). Therefore, $\widehat{\boldsymbol{\alpha}}_1$ is a consistent estimate of $\boldsymbol{\alpha}_1$.

Proof of Theorem 3.3.4. In this proof, we only consider elements in the event Ω_n. First, note that $\boldsymbol{\alpha}_1$ and $\widehat{\boldsymbol{\alpha}}_1$ are the solutions to

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^p,\boldsymbol{\alpha}\neq\mathbf{0}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}}, \quad \text{and} \quad \max_{\boldsymbol{\alpha}\in\mathbb{R}^p,\boldsymbol{\alpha}\neq\mathbf{0}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}+\tau_n\|\boldsymbol{\alpha}\|_{\lambda_n}^2}, \tag{4.36}$$

respectively, with

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 1, \qquad \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1 + \tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_{\lambda_n} = 1, \tag{4.37}$$

Hence, we have

$$\begin{aligned}
&\frac{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1}{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1} \le \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1} = \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1, \\
&\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1 = \frac{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1}{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1+\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_{\lambda_n}^2} \ge \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_1+\tau_n\|\boldsymbol{\alpha}_1\|_{\lambda_n}^2}.
\end{aligned} \tag{4.38}$$

The first inequality in (4.38) leads to

$$\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1 \le (\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1). \tag{4.39}$$

By the definition of Ω_n in (3.27),

$$|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1| \le \|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 = \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2,$$

and similarly,

$$\begin{aligned}
|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1| &\le \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2, \\
|\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1| &\le \frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2, \\
|\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1| &\le \frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2.
\end{aligned} \tag{4.40}$$

By Condition 2 (a),

$$\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1 = \lambda_1(\boldsymbol{\Xi}) \geq c_1. \tag{4.41}$$

Moreover, we have

$$\lambda_n\|\widehat{\boldsymbol{\alpha}}\|_1^2 \leq \|\widehat{\boldsymbol{\alpha}}_1\|_{\lambda_n}^2 = (1-\lambda_n)\|\widehat{\boldsymbol{\alpha}}\|_2^2 + \lambda_n\|\widehat{\boldsymbol{\alpha}}\|_1^2 \leq \|\widehat{\boldsymbol{\alpha}}\|_1^2. \tag{4.42}$$

Then by (4.37) , (4.39), (4.40), (4.41) and (4.42),

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1 \leq{}& \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1 + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \leq (\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1) + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \\
\leq{}&(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)\left(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1 + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2\right) + \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1}{c_1}\frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \\
={}&(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)\left(1-\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_{\lambda_n}^2 + \frac{1+c_1^{-1}}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2\right) \\
\leq{}&(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)\left(1-\tau_n\lambda_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 + \frac{1+c_1^{-1}}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2\right) \\
={}&(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)\left[1-\tau_n(\lambda_n-\lambda_0/2)\|\widehat{\boldsymbol{\alpha}}_1\|_1^2\right],
\end{aligned} \tag{4.43}$$

where the last equality is due to the definition of C_2 in (3.26). By (4.38) and (4.40),

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1 \geq{}& \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha}_1 + \tau_n\|\boldsymbol{\alpha}_1\|_{\lambda_n}^2} \geq \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1 - \frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2}{\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 + \frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2 + \tau_n\|\boldsymbol{\alpha}_1\|_1^2} \\
\geq{}& \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1 - \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1}{c_1}\frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2}{\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 + \frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2 + \tau_n\|\boldsymbol{\alpha}_1\|_1^2} = \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1\left[1-\frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2\right]}{1+\frac{1}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2 + \tau_n\|\boldsymbol{\alpha}_1\|_1^2},
\end{aligned} \tag{4.44}$$

which together with (4.43) leads to

$$\tau_n(\lambda_n-\lambda_0/2)\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \leq \frac{\left(1+\frac{1+c_1^{-1}}{C_2}\right)\tau_n\|\boldsymbol{\alpha}_1\|_1^2}{1+\left(1+\frac{1}{C_2}\right)\tau_n\|\boldsymbol{\alpha}_1\|_1^2} = \frac{(1+\frac{\lambda_0}{2})\tau_n\|\boldsymbol{\alpha}_1\|_1^2}{1+\left(1+\frac{1}{C_2}\right)\tau_n\|\boldsymbol{\alpha}_1\|_1^2}. \tag{4.45}$$

By (3.25) and the conditions in the theorem, when n is large enough, we have $\lambda_0 < \lambda_n < 1$

and $\tau_n\|\boldsymbol{\alpha}_1\|_1^2 = C\|\boldsymbol{\alpha}_1\|_1^2 s_n \to 0$. Therefore, for all n large enough, by (4.45), we have

$$\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \le 6\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0, \tag{4.46}$$

which together with (4.43) give

$$\frac{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1} \le 1 - \tau_n(\lambda_n - \lambda_0/2)\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \le 1 - \frac{1}{2}\lambda_0\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \le 1. \tag{4.47}$$

On the other hand, (4.44) implies

$$\begin{aligned}\frac{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1} &\ge \frac{1 - \frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2}{1 + (\frac{1}{C_2} + 1)\|\boldsymbol{\alpha}_1\|_1^2} \ge \left(1 - \frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\alpha}_1\|_1^2\right)\left(1 - (1 + \frac{1}{C_2})\tau_n\|\boldsymbol{\alpha}_1\|_1^2\right) \\ &\ge 1 - \left(1 + \frac{1 + c_1^{-1}}{C_2}\right)\tau_n\|\boldsymbol{\alpha}_1\|_1^2 = 1 - (1 + \lambda_0/2)\,\tau_n\|\boldsymbol{\alpha}_1\|_1^2 \ge 1 - 3\tau_n\|\boldsymbol{\alpha}_1\|_1^2/2,\end{aligned}$$

which together with (4.47) leads to

$$\left|\frac{\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1}{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1} - 1\right| \le 3\tau_n\|\boldsymbol{\alpha}_1\|_1^2/2 = 3C\|\boldsymbol{\alpha}_1\|_1^2 s_n/2. \tag{4.48}$$

It follows from (4.40) and (4.46),

$$\begin{aligned}|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1| &\le \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \le \frac{(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)}{c_1}\frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \\ \le&(\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)\frac{6Cc_1^{-1}}{\lambda_0 C_2}\|\boldsymbol{\alpha}_1\|_1^2 s_n,\end{aligned}$$

which together with (4.48) imply

$$|\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1| \le (\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)C_3\|\boldsymbol{\alpha}_1\|_1^2 s_n, \tag{4.49}$$

where $C_3 = 3C/2 + 6Cc_1^{-1}/(\lambda_0 C_2)$. Recall that $\widehat{\boldsymbol{\gamma}}_k = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\alpha}}_k$, $1 \le k \le K-1$, defined in Section 3.3. Let $\widehat{\boldsymbol{\gamma}}_1 = d_1\boldsymbol{\gamma}_1 + d_2\boldsymbol{\gamma}_2 + \cdots + d_{K-1}\boldsymbol{\gamma}_{K-1} + \hat{c}\widehat{\boldsymbol{\beta}}$ be the orthogonal expansion of $\widehat{\boldsymbol{\gamma}}_1$, where $\widehat{\boldsymbol{\beta}}$ is an vector orthogonal to each of $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K-1}$, with $\|\widehat{\boldsymbol{\beta}}\|_2 = 1$. Because $\boldsymbol{\Xi}$ has only $K-1$ nonzero eigenvalues with the corresponding eigenvectors, $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K-1}$, we have

$\boldsymbol{\Xi}\widehat{\boldsymbol{\beta}} = \mathbf{0}$. Then

$$\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1 = \widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_1 = d_1^2\lambda_1(\boldsymbol{\Xi}) + d_2^2\lambda_2(\boldsymbol{\Xi}) + \cdots + d_{K-1}^2\lambda_{K-1}(\boldsymbol{\Xi}).$$

By (4.49) and (4.41),

$$\begin{aligned}
&\lambda_1(\boldsymbol{\Xi})C_3\|\boldsymbol{\alpha}_1\|_1^2 s_n = (\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1)C_3\|\boldsymbol{\alpha}_1\|_1^2 s_n \geq |\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1| \\
=&\left|d_1^2\lambda_1(\boldsymbol{\Xi}) + d_2^2\lambda_2(\boldsymbol{\Xi}) + \cdots + d_{K-1}^2\lambda_{K-1}(\boldsymbol{\Xi}) - \lambda_1(\boldsymbol{\Xi})\right| \\
\geq&\left|d_1^2 - 1\right|\lambda_1(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})\sum_{i=2}^{K-1} d_i^2 \\
\geq&\left|d_1^2 - 1\right|\lambda_1(\boldsymbol{\Xi}) - \left|d_1^2 - 1\right|\lambda_2(\boldsymbol{\Xi}) - (d_1^2 - 1)\lambda_2(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})\sum_{i=2}^{K-1} d_i^2 \\
=&\left|d_1^2 - 1\right|[\lambda_1(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})] - \lambda_2(\boldsymbol{\Xi})\left[\sum_{i=1}^{K-1} d_i^2 - 1\right] \\
\geq&\left|d_1^2 - 1\right|[\lambda_1(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})] - \lambda_2(\boldsymbol{\Xi})\left(\|\widehat{\boldsymbol{\gamma}}_1\|_2^2 - 1\right) \\
\geq&\left|d_1^2 - 1\right|[\lambda_1(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})] - \lambda_1(\boldsymbol{\Xi})\left|\|\widehat{\boldsymbol{\gamma}}_1\|_2^2 - 1\right|.
\end{aligned} \tag{4.50}$$

By (4.46) and (4.40),

$$\begin{aligned}
&\left|\|\widehat{\boldsymbol{\gamma}}_1\|_2^2 - 1\right| = \left|\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1 - 1\right| = \left|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1 - 1\right| \leq \left|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_1 - 1\right| + \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \\
\leq&\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_{\lambda_n}^2 + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \leq (1 + \frac{1}{C_2})\tau_n\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 = 6(1 + C_2^{-1})C\|\boldsymbol{\alpha}_1\|_1^2 s_n/\lambda_0.
\end{aligned} \tag{4.51}$$

Then by (4.50), (4.51) and Condition 2 (b),

$$\left|d_1^2 - 1\right| \leq \left(\frac{\lambda_1(\boldsymbol{\Xi}) - \lambda_2(\boldsymbol{\Xi})}{\lambda_1(\boldsymbol{\Xi})}\right)^{-1}[C_3 + 6(1 + C_2^{-1})C/\lambda_0]\|\boldsymbol{\alpha}_1\|_1^2 s_n \leq C_4\|\boldsymbol{\alpha}_1\|_1^2 s_n, \tag{4.52}$$

where $C_4 = c_2^{-1}[C_3 + 6(1 + C_2^{-1})C/\lambda_0]$. Since $\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\gamma}_1 = d_1 > 0$, by (4.52),

$$\left|\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\gamma}_1 - \|\boldsymbol{\gamma}_1\|_2^2\right| = \left|d_1 - 1\right| \leq \left|d_1 - 1\right|(d_1 + 1) = \left|d_1^2 - 1\right| \leq C_4\|\boldsymbol{\alpha}_1\|_1^2 s_n. \tag{4.53}$$

Now by combining (4.51) and (4.53), we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1\|_2^2=&\left|\|\widehat{\boldsymbol{\gamma}}_1\|_2^2-2\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\gamma}_1+\|\boldsymbol{\gamma}_1\|_2^2\right|\\ \leq&\left|\|\widehat{\boldsymbol{\gamma}}_1\|_2^2-\|\boldsymbol{\gamma}_1\|_2^2\right|+\left|-2\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\gamma}_1+2\|\boldsymbol{\gamma}_1\|_2^2\right|\\ =&\left|\|\widehat{\boldsymbol{\gamma}}_1\|_2^2-1\right|+2\left|\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\boldsymbol{\gamma}_1-\|\boldsymbol{\gamma}_1\|_2^2\right|\\ \leq&C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n,\end{aligned}$$

where $C_5=6(1+C_2^{-1})C/\lambda_0+2C_4$. Moreover, by Condition 1, it follows from the above inequality

$$\begin{aligned}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1\|_2^2=&(\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1)^{\mathrm{T}}(\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1)=(\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1)\\ \leq&\|\boldsymbol{\Sigma}^{-1}\|\|\widehat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_1\|_2^2\leq c_0C_5\|\boldsymbol{\alpha}_1\|_1^2 s_n.\end{aligned}$$

We have proved the theorem.

□

4.7 Proof of Theorem 3.3.5

Theorem (Theorem 3.3.5)**.** *Suppose that $K=2$ and Conditions 1-2 hold. Then the misclassification rate of the optimal rule* (3.5) *and the conditional misclassification rate of our sparse LDA rule as given in Section 3.2.1 are*

$$\begin{aligned}R_{OPT}&=\Phi\left(-\frac{\boldsymbol{\delta}^{\mathsf{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathsf{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right),\\ R(\mathbf{X})&=\frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathsf{T}}\widehat{\mathbf{D}}(2\boldsymbol{\mu}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)}{2\|\widehat{\boldsymbol{\delta}}^{\mathsf{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right)+\frac{1}{2}\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathsf{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2-2\boldsymbol{\mu}_1)}{2\|\widehat{\boldsymbol{\delta}}^{\mathsf{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right),\end{aligned}\tag{4.54}$$

respectively, where Φ is the cumulative distribution function of the standard normal distribution, $\boldsymbol{\delta}=\boldsymbol{\mu}_2-\boldsymbol{\mu}_1$ and $\widehat{\boldsymbol{\delta}}=\bar{\mathbf{x}}_2-\bar{\mathbf{x}}_1$. Moreover, if $s_n\to 0$ and $\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n\to 0$ as $n,p\to\infty$,

our method is asymptotically optimal and we have

$$\frac{R(\mathbf{X})}{R_{OPT}} - 1 \leq O_p\left(\lambda_1(\mathbf{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n\right). \tag{4.55}$$

Proof of Theorem 3.3.5. When $K = 2$, a new observation $\mathbf{x}$ is assigned to Class 1 by the optimal rule (3.5) if and only if $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{D}[\mathbf{x} - (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2] < 0$. Hence,

$$\begin{aligned} & P\left(\mathbf{x} \text{ is assigned to Class } 1 | \mathbf{x} \in \text{ Class } 2\right) \\ =& P\left((\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{D}[\mathbf{x} - (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2] < 0 | \mathbf{x} \in \text{ Class } 2\right) \\ =& P\left(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}[\mathbf{x} - \boldsymbol{\mu}_2] < -\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2} | \mathbf{x} \in \text{ Class } 2\right) \\ =& P\left(\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\mathbf{Z} < -\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2}\right), \end{aligned} \tag{4.56}$$

where $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}[\mathbf{x} - \boldsymbol{\mu}_2] \sim N(\mathbf{0}, \mathbf{I})$ given $\mathbf{x}$ in Class 2. Hence, the above probability is equal to $\Phi(-\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2})$. The same result is true for

$$P\left(\mathbf{x} \text{ is assigned to Class } 2 | \mathbf{x} \in \text{ Class } 1\right).$$

Hence,

$$\begin{aligned} R_{OPT} =& \frac{1}{2}P\left(\mathbf{x} \text{ is assigned to Class } 1 | \mathbf{x} \in \text{ Class } 2\right) \\ & + \frac{1}{2}P\left(\mathbf{x} \text{ is assigned to Class } 2 | \mathbf{x} \in \text{ Class } 1\right) \\ =& \Phi\left(-\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right). \end{aligned}$$

On the other hand, a new observation $\mathbf{x}$ is assigned to Class 1 by sparse LDA rule (3.15) if and only if $(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^{\mathrm{T}}\widehat{\mathbf{D}}[\mathbf{x} - (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2] < 0$. A similar argument as in (4.56) leads to

$$P_{\cdot|\mathbf{X}}\left(\mathbf{x} \text{ is assigned to Class } 1 | \mathbf{x} \in \text{ Class } 2\right) = \Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right),$$

where $P_{\cdot|\mathbf{X}}$ means the probability given the training sample $\mathbf{X}$. Similarly,

$$P_{\cdot|\mathbf{X}}\left(\mathbf{x} \text{ is assigned to Class } 2|\mathbf{x} \in \text{ Class } 1\right) = \Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2 - 2\boldsymbol{\mu}_1]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right).$$

Then (4.54) follows. We will use the following inequality (see page 850 of Shorack and Wellner [39]):

$$(1 - \frac{1}{x^2})\phi(x) \leq x[1 - \Phi(x)] = x\Phi(-x) \leq \phi(x), \quad \forall x > 0, \tag{4.57}$$

where ϕ is the density function of the standard normal distribution. Therefore, if $x > \sqrt{2}$,

$$\Phi(-x) \geq (1 - \frac{1}{x^2})\phi(x)x^{-1} \geq \frac{1}{2}\phi(x)x^{-1}, \tag{4.58}$$

if $0 < x \leq \sqrt{2}$, $\Phi(-x) \geq \Phi(-\sqrt{2}) \geq \Phi(-\sqrt{2})\phi(0)^{-1}\phi(x)$. Hence, we have for any $x > 0$,

$$\Phi(-x) \geq \frac{C_1}{1 + 2C_1 x}\phi(x), \quad \text{where} \quad C_1 = \Phi(-\sqrt{2})\phi(0)^{-1} \tag{4.59}$$

By (4.59), for any $x > 0$ and ϵ with $x + \epsilon > 0$ (ϵ can be negative or positive),

$$\begin{aligned}
\left|\frac{\Phi(-(x+\epsilon))}{\Phi(-x)} - 1\right| =& \frac{|\Phi(-(x+\epsilon)) - \Phi(-x)|}{\Phi(-x)} = \frac{\left|\int_{-x}^{-(x+\epsilon)} \phi(y)dy\right|}{\Phi(-x)} \\
=& \frac{|-\epsilon\phi(-(x+\widetilde{\epsilon}))|}{\Phi(-x)} \leq \frac{(1+2C_1x)|\epsilon|\phi(-(x+\widetilde{\epsilon}))}{C_1\phi(x)} \\
=& \frac{(1+2C_1x)|\epsilon|}{C_1}e^{-\frac{(x+\widetilde{\epsilon})^2 - x^2}{2}} = \frac{(1+2C_1x)|\epsilon|}{C_1}e^{-\frac{2x\widetilde{\epsilon}+\widetilde{\epsilon}^2}{2}} \\
\leq& \frac{(1+2C_1x)|\epsilon|}{C_1}e^{x|\epsilon|},
\end{aligned} \tag{4.60}$$

where $\widetilde{\epsilon}$ is a number between 0 and ϵ. We will apply (4.60) to

$$x = \frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}, \quad \epsilon = \frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}. \tag{4.61}$$

By Lemma 10, we can choose a constant $\widetilde{C}$ such that

$$P\left(\max_{1\leq j\leq K}\|\bar{\mathbf{x}}_j-\boldsymbol{\mu}_j\|_\infty>\widetilde{C}\sqrt{\frac{K\log p}{n}}\right)\leq p^{-1},$$

Define

$$\widetilde{\Omega}_n=\left\{\max_{1\leq j\leq K}\|\bar{\mathbf{x}}_j-\boldsymbol{\mu}_j\|_\infty\leq\widetilde{C}\sqrt{\frac{K\log p}{n}}=\widetilde{C}s_n\right\},$$

then

$$[P\left(\widetilde{\Omega}_n\right)\geq 1-p^{-1}. \tag{4.62}$$

In the rest of the proof, we only consider the elements in $\Omega_n\bigcap\widetilde{\Omega}_n$ which has a probability greater than $1-3p^{-1}$ by (3.27) and (4.62). Note that by (3.2), we have $\boldsymbol{\mu}_1=-\boldsymbol{\mu}_2$. Because

$$\begin{aligned}&\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2]=[(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)-(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)+2\boldsymbol{\mu}_2]^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2-(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)-(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)]\\ =&4\boldsymbol{\mu}_2^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\mu}_2-(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)+(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)-4(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\mu}_2,\end{aligned}$$

we have

$$\begin{aligned}\left|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2-\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2]-\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}\right|\leq&\left|4\boldsymbol{\mu}_2^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\mu}_2-\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}\right|+(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_2-\boldsymbol{\mu}_2)\\ &+(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)+4\left|(\bar{\mathbf{x}}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\mu}_2\right|\\ =&I+II+III+IV.\end{aligned} \tag{4.63}$$

We estimate each of the four terms. Because $\boldsymbol{\delta}=\boldsymbol{\mu}_2-\boldsymbol{\mu}_1=2\boldsymbol{\mu}_2$, by (4.49), the first term

$$\begin{aligned}I=&\left|4\boldsymbol{\mu}_2^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\mu}_2-\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}\right|=\left|4\boldsymbol{\mu}_2^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\mu}_2-4\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\mu}_2\right|\\ =&\left|4\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\mu}_2\boldsymbol{\mu}_2^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1-4\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\mu}_2\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\alpha}_1\right|=4\left|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{B}\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1\right|\\ \leq&4C_3\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1\|\boldsymbol{\alpha}_1\|_1^2s_n=4C_3\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2s_n,\end{aligned} \tag{4.64}$$

and we have

$$\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}=4\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_1=4\lambda_1(\boldsymbol{\Xi}). \tag{4.65}$$

For the second term, by the definition of $\widetilde{\Omega}_n$ in (4.62) and Theorem 3.3.4

$$\begin{aligned} II =&(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2)^{\mathrm{T}}\widehat{\mathbf{D}}(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2) = \left|(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2)^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1\right|^2 \\ \leq&\|\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2\|_\infty^2\|\widehat{\boldsymbol{\alpha}}_1\|_1^2 \\ \leq&\widetilde{C}^2 s_n^2 6\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0. \end{aligned} \tag{4.66}$$

The same bound for the third term. For the last one, by Condition 1 (b),

$$\begin{aligned} IV =&4\left|(\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1)^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1\right|\left|\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\mu}_2\right| \\ \leq&4\left[\max_{1\leq i\leq K}\|\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i\|_\infty\right]\|\widehat{\boldsymbol{\alpha}}_1\|_1\|\widehat{\boldsymbol{\alpha}}_1\|_1\|\boldsymbol{\mu}_2\|_\infty \\ \leq&24c_0\widetilde{C}s_n\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0. \end{aligned} \tag{4.67}$$

By (4.63)-(4.67), $s_n \to 0$ and $\lambda_1(\boldsymbol{\Xi}) \geq c_1$ (see Condition 2), we have

$$\left|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[2\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2] - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}\right| \leq C_4\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n, \tag{4.68}$$

where C_4 is a constant independent of n and p.

Next, by (4.51) and $\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 1$,

$$\begin{aligned} &\left|\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2^2 - \|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2^2\right| = \left|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\boldsymbol{\delta}\right| \\ =&\left|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\delta}\right| = |\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_1 - 1|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} + |\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}| \\ =&|\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_1 - 1|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} + |\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}| \\ \leq&6(1 + C_2^{-1})C\lambda_0^{-1}\|\boldsymbol{\alpha}_1\|_1^2 s_n\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} + |\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}|. \end{aligned} \tag{4.69}$$

By a similar argument as those for (4.68), we can show that

$$\begin{aligned} &|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}| \leq C_5\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n, \\ &\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}} = (1 + o(1))\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta} = (1 + o(1))4\lambda_1(\boldsymbol{\Xi}), \end{aligned} \tag{4.70}$$

where the last equality is due to (4.65) and C_5 is a constant independent of n and p. By

Lemma 8 and (4.65), we have $\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2 = \sqrt{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}} = \sqrt{4\lambda_1(\boldsymbol{\Xi})}$ which together with (4.69), (4.70) give

$$\begin{aligned}
&\left|\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2^2 - \|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2^2\right| \le \widetilde{C_5}\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n,\\
&\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2 = \sqrt{4\lambda_1(\boldsymbol{\Xi})} + o(1),\\
&\text{and}\\
&\left|\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right|\\
=&|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}|\frac{\left|\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2^2 - \|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2^2\right|}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2\left(\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2 + \|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2\right)}\\
\le&C_6\sqrt{\lambda_1(\boldsymbol{\Xi})}\|\boldsymbol{\alpha}_1\|_1^2 s_n,
\end{aligned}$$

which together with (4.68) imply

$$\begin{aligned}
|\epsilon| =&\left|\frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[\widehat{\boldsymbol{\delta}} - 2(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2)]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right|\\
\le&\frac{|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[\widehat{\boldsymbol{\delta}} - 2(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2)] - \boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}|}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2} + \left|\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right|\\
\le&C_7\sqrt{\lambda_1(\boldsymbol{\Xi})}\|\boldsymbol{\alpha}_1\|_1^2 s_n, \qquad (4.71)
\end{aligned}$$

where $\widetilde{C_5}$, C_6 and C_7 are constants independent of n and p. By (4.60), (4.61) and (4.71) and noting that $x = \sqrt{\lambda_1(\boldsymbol{\Xi})} \ge \sqrt{c_1}$, we have $|x\epsilon| \le C_7\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n = o(1)$ and hence

$$\begin{aligned}
&\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[\widehat{\boldsymbol{\delta}} - 2(\bar{\mathbf{x}}_2 - \boldsymbol{\mu}_2)]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right)/\Phi\left(-\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right) - 1 = \frac{\Phi(-(x+\epsilon))}{\Phi(-x)} - 1\\
\le&\frac{(1+2C_1x)|\epsilon|}{C_1}e^{x|\epsilon|} \le C_8\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n
\end{aligned}$$

where C_8 is a constant independent of n and p. Similarly, we have

$$\Phi\left(-\frac{\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}[\widehat{\boldsymbol{\delta}} + 2(\bar{\mathbf{x}}_1 - \boldsymbol{\mu}_1)]}{2\|\widehat{\boldsymbol{\delta}}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\Sigma}^{1/2}\|_2}\right)/\Phi\left(-\frac{\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\delta}}{2\|\boldsymbol{\delta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Sigma}^{1/2}\|_2}\right) - 1 \le C_9\lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\alpha}_1\|_1^2 s_n.$$

where C_9 is a constant independent of n and p. Therefore, the above two inequalities together

with (4.54) give (4.55).

□

4.8 Proof of Theorem 3.3.7

Theorem (Theorem 3.3.7)**.** *Suppose that Conditions 1-2 hold. Let the tuning parameter in the optimization problem* (3.12)*,* $\kappa_n = \widetilde{C}\lambda_1(\boldsymbol{\Xi})\Lambda_p s_n$*, where* $\widetilde{C}$ *is a constant large enough and independent of* n *and* p*. For any* $1 \le i \le K-1$*, let* $\mathbf{Q}_i$ *and* $\widehat{\mathbf{Q}}_i$ *be the orthogonal projection matrices onto the following subspaces of* $\mathbb{R}^p$*, respectively,*

$$\mathbf{W}_i = \text{span}\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_i\}, \quad \widehat{\mathbf{W}}_i = \text{span}\{\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \cdots, \widehat{\boldsymbol{\xi}}_i\}, \tag{4.72}$$

where $\boldsymbol{\xi}_i = \mathbf{B}\boldsymbol{\alpha}_i = \lambda_i(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_i$*. If* $s_n \to 0$ *and* $\Lambda_p^2 s_n \to 0$ *as* $n, p \to \infty$*, then for each* $1 \le i \le K-1$*, there exist constants* $D_{i,1}$*,* $D_{i,2}$ *and* $D_{i,3}$ *independent of* n *and* p *such that in* Ω_n*,*

$$\|\widehat{\boldsymbol{\alpha}}_i\|_1 \le D_{i,1}\Lambda_p, \quad \|\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i\|_2^2 \le D_{i,2}\Lambda_p^2 s_n, \quad \|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|^2 \le D_{i,3}\Lambda_p^2 s_n. \tag{4.73}$$

Hence, for each $1 \le i \le K-1$*,* $\widehat{\boldsymbol{\alpha}}_i$ *is a consistent estimate of* $\boldsymbol{\alpha}_i$*, and the projection matrix* $\widehat{\mathbf{Q}}_i$ *is a consistent estimate of* $\mathbf{Q}_i$*.*

Proof of Theorem 3.3.7. Due to the constraints of the optimization problems (3.10) and (3.14) and the definitions of $\mathbf{W}_i$ and $\widehat{\mathbf{W}}_i$ in (4.72), for any $1 \le i \le K-1$ and $j < i$, we have

$$\begin{aligned} &\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_i = 1, \quad \widehat{\boldsymbol{\alpha}}_i^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_i + \tau_n\|\widehat{\boldsymbol{\alpha}}_i\|_{\lambda_n} = 1, \\ &\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = 0, \quad \boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\xi}_j = 0, \quad \widehat{\boldsymbol{\alpha}}_i^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_j = 0. \end{aligned} \tag{4.74}$$

Then by the definitions of $\boldsymbol{\gamma}_k$ and $\widehat{\boldsymbol{\gamma}}_k$ in (3.23) and (3.24), for any $1 \le i \le K-1$ and $j < i$, we have

$$\boldsymbol{\gamma}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_i, \quad \|\boldsymbol{\gamma}_i\|_2 = 1, \quad \boldsymbol{\gamma}_i^{\mathrm{T}}\boldsymbol{\gamma}_j = 0, \quad \widehat{\boldsymbol{\gamma}}_i = \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\alpha}}_i, \quad \widehat{\boldsymbol{\gamma}}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\xi}}_j = 0. \tag{4.75}$$

For any $1 \le i \le K-1$, in addition to $\mathbf{W}_i$ and $\widehat{\mathbf{W}}_i$, we define the following two subspaces of

$\mathbb{R}^p$,

$$\mathbf{V}_i = \text{span}\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_i\}, \quad \widehat{\mathbf{V}}_i = \text{span}\{\widehat{\boldsymbol{\zeta}}_1, \widehat{\boldsymbol{\zeta}}_2, \cdots, \widehat{\boldsymbol{\zeta}}_i\},$$

where $\widehat{\boldsymbol{\zeta}}_i = \lambda_i(\boldsymbol{\Xi})^{-1}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\xi}}_i$. Let $\mathbf{P}_i$ and $\widehat{\mathbf{P}}_i$ be the orthogonal projection matrices onto $\mathbf{V}_i$ and $\widehat{\mathbf{V}}_i$, respectively. By (4.74) and (4.75) and the definitions of $\mathbf{W}_i$ and $\widehat{\mathbf{W}}_i$ in (4.72),

$$\boldsymbol{\gamma}_k \in \mathbf{V}_i^{\perp}, \quad \boldsymbol{\alpha}_k \in \mathbf{W}_i^{\perp}, \quad \widehat{\boldsymbol{\alpha}}_k \in \widehat{\mathbf{W}}_i^{\perp}, \quad \widehat{\boldsymbol{\gamma}}_k \in \widehat{\mathbf{V}}_i^{\perp}, \quad \text{for any } k > i, \tag{4.76}$$

where $\mathbf{V}_i^{\perp}$, $\widehat{\mathbf{V}}_i^{\perp}$, $\mathbf{W}_i^{\perp}$ and $\widehat{\mathbf{W}}_i^{\perp}$ are orthogonal complementary subspaces of $\mathbf{V}_i$, $\widehat{\mathbf{V}}_i$, $\mathbf{W}_i$ and $\widehat{\mathbf{W}}_i$, respectively. We will prove that in the event Ω_n (defined in (3.27)), for each $1 \le i \le K-1$, there exist constants, $C_{i,1}$, $C_{i,2}$, $C_{i,3}$, $C_{i,4}$, $C_{i,5}$ and $C_{i,6}$ independent of n and p such that

$$\begin{aligned}
&\|\widehat{\boldsymbol{\alpha}}_i\|_1 \le C_{i,1}\Lambda_p, \quad \|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2^2 \le C_{i,2}\Lambda_p^2 s_n, \quad \|\mathbf{P}_i - \widehat{\mathbf{P}}_i\|^2 \le C_{i,3}\Lambda_p^2 s_n, \\
&\|\mathbf{Q}_i - \widehat{\mathbf{Q}}_i\|^2 \le C_{i,4}\Lambda_p^2 s_n, \quad \|\widehat{\boldsymbol{\xi}}_i\|_1 \le C_{i,5}\lambda_1(\boldsymbol{\Xi})\Lambda_p, \quad \|\widehat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\|_2^2 \le C_{i,6}\lambda_1(\boldsymbol{\Xi})^2\Lambda_p^2 s_n,
\end{aligned} \tag{4.77}$$

as n is large enough. We proceed by induction. When $i = 1$, the first two inequalities in (4.77) follow from (4.35) in Theorem 3.3.4 by setting $C_{1,1} = 6/\lambda_0$ and $C_{1,2} = C_5$, and the last two inequalities follow from the following lemma by setting $C_{1,5} = C_7$ and $C_{1,6} = C_6^2$ in (4.78).

Lemma 11. *Under the conditions of the theorem, we have, in Ω_n,*

$$\|\widehat{\boldsymbol{\xi}}_1 - \boldsymbol{\xi}_1\|_2 = \|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2 \le C_6\lambda_1(\boldsymbol{\Xi})\sqrt{\Lambda_p^2 s_n}, \quad \|\widehat{\boldsymbol{\xi}}_1\|_1 \le C_7\lambda_1(\boldsymbol{\Xi})\Lambda_p. \tag{4.78}$$

where C_6 and C_7 are constants independent of p and n.

On the other hand, since $\|\boldsymbol{\gamma}_1\|_2 = 1$, we have $\mathbf{P}_1 = \boldsymbol{\gamma}_1\boldsymbol{\gamma}_1^{\mathrm{T}}$ and $\widehat{\mathbf{P}}_1 = \widehat{\boldsymbol{\zeta}}_1\widehat{\boldsymbol{\zeta}}_1^{\mathrm{T}}/\|\widehat{\boldsymbol{\zeta}}_1\|_2^2$. Then

$$\begin{aligned}\|\mathbf{P}_1 - \widehat{\mathbf{P}}_1\| =& \left\|\boldsymbol{\gamma}_1\boldsymbol{\gamma}_1^{\mathrm{T}} - \frac{1}{\|\widehat{\boldsymbol{\zeta}}_1\|_2^2}\widehat{\boldsymbol{\zeta}}_1\widehat{\boldsymbol{\zeta}}_1^{\mathrm{T}}\right\| \\ \leq& \|(\boldsymbol{\gamma}_1 - \widehat{\boldsymbol{\zeta}}_1)\boldsymbol{\gamma}_1^{\mathrm{T}}\| + \|\widehat{\boldsymbol{\zeta}}_1(\boldsymbol{\gamma}_1 - \widehat{\boldsymbol{\zeta}}_1)^{\mathrm{T}}\| + \left\|\left(1 - \frac{1}{\|\widehat{\boldsymbol{\zeta}}_1\|_2^2}\right)\widehat{\boldsymbol{\zeta}}_1\widehat{\boldsymbol{\zeta}}_1^{\mathrm{T}}\right\| \\ \leq& \|\boldsymbol{\gamma}_1 - \widehat{\boldsymbol{\zeta}}_1\|_2(\|\boldsymbol{\gamma}_1\|_2 + \|\widehat{\boldsymbol{\zeta}}_1\|_2) + \left|1 - \|\widehat{\boldsymbol{\zeta}}_1\|_2^2\right|. \end{aligned} \tag{4.79}$$

Note that by Condition 1 (b), $\|\boldsymbol{\Sigma}^{-1/2}\| = \lambda_{max}(\boldsymbol{\Sigma}^{-1/2}) = \lambda_{min}(\boldsymbol{\Sigma})^{-1/2} \leq c_0^{1/2}$. Then by (4.78),

$$\begin{aligned}\|\boldsymbol{\gamma}_1 - \widehat{\boldsymbol{\zeta}}_1\|_2 =& \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_1 - \lambda_1(\boldsymbol{\Xi})^{-1}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\xi}}_1\|_2 \\ \leq& \lambda_1(\boldsymbol{\Xi})^{-1}\|\boldsymbol{\Sigma}^{-1/2}\|\|\lambda_i(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \widehat{\boldsymbol{\xi}}_1\|_2 \\ =& \lambda_1(\boldsymbol{\Xi})^{-1}\|\boldsymbol{\Sigma}^{-1/2}\|\|\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\boldsymbol{\xi}}_1\|_2 \\ \leq& c_0^{1/2}C_6\sqrt{\Lambda_p^2 s_n}. \end{aligned} \tag{4.80}$$

Therefore, $\|\widehat{\boldsymbol{\zeta}}_1\|_2 \leq 1 + c_0^{1/2}C_6\sqrt{\Lambda_p^2 s_n}$ and

$$\begin{aligned}\left|1 - \|\widehat{\boldsymbol{\zeta}}_1\|_2^2\right| =& \left|1 - \|\widehat{\boldsymbol{\zeta}}_1\|_2\right|\left(1 + \|\widehat{\boldsymbol{\zeta}}_1\|_2\right) = \left|\|\boldsymbol{\gamma}_1\|_2 - \|\widehat{\boldsymbol{\zeta}}_1\|_2\right|\left(1 + \|\widehat{\boldsymbol{\zeta}}_1\|_2\right) \\ \leq& \|\boldsymbol{\gamma}_1 - \widehat{\boldsymbol{\zeta}}_1\|_2(2 + c_0^{1/2}C_6\sqrt{\Lambda_p^2 s_n}) \\ \leq& c_0^{1/2}C_6\sqrt{\Lambda_p^2 s_n}(2 + c_0^{1/2}C_6\sqrt{\Lambda_p^2 s_n}). \end{aligned} \tag{4.81}$$

Since $\Lambda_p^2 s_n \to 0$, as n is large enough, by (4.79)-(4.81), we can find $C_{1,3}$ large enough and independent of n and p such that $\|\mathbf{P}_1 - \widehat{\mathbf{P}}_1\|^2 \leq C_{1,3}\Lambda_p^2 s_n$. Note that

$$\begin{aligned}&\|\boldsymbol{\xi}_1\|_2^2 = \|\mathbf{B}\boldsymbol{\alpha}_1\|_2^2 = \|\lambda_1(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_1\|_2^2 = \lambda_1(\boldsymbol{\Xi})^2\boldsymbol{\gamma}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\gamma}_1, \\ &\|\mathbf{Q}_1 - \widehat{\mathbf{Q}}_1\| = \left\|\frac{1}{\|\boldsymbol{\xi}_1\|_2^2}\boldsymbol{\xi}_1\boldsymbol{\xi}_1^{\mathrm{T}} - \frac{1}{\|\widehat{\boldsymbol{\xi}}_1\|_2^2}\widehat{\boldsymbol{\xi}}_1\widehat{\boldsymbol{\xi}}_1^{\mathrm{T}}\right\|, \end{aligned} \tag{4.82}$$

Therefore, we have $c_0^{-1}\lambda_1(\boldsymbol{\Xi})^2 \leq \|\boldsymbol{\xi}_1\|_2^2 \leq c_0\lambda_1(\boldsymbol{\Xi})^2$ and by (4.78), (4.82) and the same argument as in (4.79)-(4.81), we can find a constant $C_{1,4}$ independent of n and p such that $\|\mathbf{Q}_1 - \widehat{\mathbf{Q}}_1\|^2 \leq C_{1,4}\Lambda_p^2 s_n$. Hence, (4.77) is true for $i = 1$. Now let $1 < k \leq K - 1$. We will show that under the assumption that all the inequalities (4.77) are true for all $1 \leq i \leq k - 1$

and all large enough n, they are also true for k and all large enough n. Because the proof is long and technical, we summarize the results in the following Lemma and provide the proof in Section 4.11.

Lemma 12. *In Ω_n, suppose that* (4.77) *is true for all $1 \leq i \leq k-1$ and all large enough n. Then* (4.77) *is also true for k and all large enough n.*

Hence, it follows from 12 that the inequalities in (4.77) are true for all $1 \leq k \leq K-1$. Based on (4.77), in order to prove the theorem, we only need to show

$$\|\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i\|_2^2 \leq \|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\gamma}_i\|_2^2 \leq \|\boldsymbol{\Sigma}^{-1}\|\|\widehat{\boldsymbol{\gamma}}_i - \boldsymbol{\gamma}_i\|_2^2 \leq c_0 C_{i,2}\Lambda_p^2 s_n.$$

Then we can obtain (4.73) by setting $D_{i,1} = C_{i,1}$, $D_{i,2} = c_0 C_{i,2}$ and $D_{i,3} = C_{i,4}$.

□

4.9 Proof of Theorem 3.3.8

Theorem (Theorem 3.3.8)**.** *Suppose that Conditions 1 and 2 hold and the general classification rule T in* (3.37) *satisfies: $\widehat{\mathbf{a}}_{ji} = -\widehat{\mathbf{a}}_{ij}$ and $\widehat{\mathbf{b}}_{ji} = \widehat{\mathbf{b}}_{ij}$. Let $\{\delta_n : n \geq 1\}$ be a sequence of nonrandom positive numbers with $\delta_n \to 0$ and $\lambda_{max}(\boldsymbol{\Delta})\delta_n \to 0$ as $n \to \infty$. For any $1 \leq j \neq i \leq K$, let*

$$\mathbf{a}_{ji} = t_{ji}\widehat{\mathbf{a}}_{ji} + (\mathbf{a}_{ji})_\perp$$

be an orthogonal decomposition of $\widehat{\mathbf{a}}_{ji}$, where $t_{ji}\widehat{\mathbf{a}}_{ji}$ is the orthogonal projection of $\mathbf{a}_{ji}$ along the direction of $\widehat{\mathbf{a}}_{ji}$, t_{ji} is a real number, and $(\mathbf{a}_{ji})_\perp$ is orthogonal to $t_{ji}\widehat{\mathbf{a}}_{ji}$. Let

$$\widehat{d}_{ji} = \widehat{\mathbf{a}}_{ji}^\top \boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i)\ , \quad d_{ji} = \mathbf{a}_{ji}^\top \boldsymbol{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i) = \frac{1}{2}\|\mathbf{a}_{ji}\|_2^2. \tag{4.83}$$

If the following conditions are satisfied,

$$\|\mathbf{a}_{ji}\|_2^2 - \|\widehat{\mathbf{a}}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 O_p(\delta_n), \quad t_{ji} = 1 + O_p(\delta_n), \quad d_{ji} - \widehat{d}_{ji} = \|\widehat{\mathbf{a}}_{ji}\|_2^2 O_p(\delta_n), \tag{4.84}$$

where $O_p(\delta_n)$ are uniform for all $1 \le j \ne i \le K$, then we have

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \le O_p\left(K^2\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[\{\lambda_{max}(\mathbf{\Delta})\delta_n)\}^{-1}\right]}\right). \tag{4.85}$$

Proof of Theorem 3.3.8. Given a new observation $\mathbf{x}$, $T_{OPT}(\mathbf{x})$ and $T(\mathbf{x})$ denote the classes to which $\mathbf{x}$ is assigned by the rules T_{OPT} and T, respectively. We use $P_{\cdot|\mathbf{X}}$ to denote the conditional probability given the training sample $\mathbf{X}$.

$$\begin{aligned}
&R_T(\mathbf{X}) - R_{OPT} = (1 - R_{OPT}) - (1 - R_T(\mathbf{X})) \qquad (4.86)\\
=&\sum_{i=1}^{K} P(T_{OPT}(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class})P(\mathbf{x} \in \text{the } i\text{th class})\\
&- \sum_{i=1}^{K} P_{\cdot|\mathbf{X}}(T(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class})P(\mathbf{x} \in \text{the } i\text{th class})\\
=&\frac{1}{K}\sum_{i=1}^{K}\Big[P(T_{OPT}(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class}) - P_{\cdot|\mathbf{X}}(T(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class})\Big]\\
=&\frac{1}{K}\sum_{i=1}^{K}\Big[\sum_{j\ne i} P_{\cdot|\mathbf{X}}(T_{OPT}(\mathbf{x}) = i, T(\mathbf{x}) = j|\mathbf{x} \in \text{the } i\text{th class})\\
&+ P_{\cdot|\mathbf{X}}(T_{OPT}(\mathbf{x}) = i, T(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class}) - P_{\cdot|\mathbf{X}}(T(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class})\Big]\\
\le&\frac{1}{K}\sum_{i=1}^{K}\sum_{j\ne i} P_{\cdot|\mathbf{X}}(T_{OPT}(\mathbf{x}) = i, T(\mathbf{x}) = j|\mathbf{x} \in \text{the } i\text{th class})\\
=&\frac{1}{K}\sum_{i=1}^{K}\sum_{j\ne i} P_{\cdot|\mathbf{X}}(T(\mathbf{x}) = j, T_{OPT}(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class}).
\end{aligned}$$

We use $P_i(\cdot)$ to denote the conditional probability $P_{\cdot|\mathbf{X}}(\cdot|\mathbf{x} \in \text{the } i\text{th class})$. Then

$$\begin{aligned}
&P_{\cdot|\mathbf{X}}\left(T(\mathbf{x}) = j, T_{OPT}(\mathbf{x}) = i|\mathbf{x} \in \text{the } i\text{th class}\right) = P_i\left(T(\mathbf{x}) = j, T_{OPT}(\mathbf{x}) = i\right) \qquad (4.87)\\
=&P_i\left(\widehat{\mathbf{a}}_{kj}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \widehat{\mathbf{b}}_{kj}) < 0, \forall k \ne j, \text{ and } \mathbf{a}_{li}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{b}_{li}) < 0, \forall l \ne i\right)\\
\le&P_i\left(\widehat{\mathbf{a}}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \widehat{\mathbf{b}}_{ij}) < 0, \text{ and } \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{b}_{ji}) < 0\right)\\
=&P_i\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \widehat{\mathbf{b}}_{ji}) > 0, \text{ and } \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \mathbf{b}_{ji}) < 0\right)\\
=&P_{\mathbf{Z}}\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{Z} > \widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i), \text{ and } \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} < \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i)\right),
\end{aligned}$$

where $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i) \sim N(\mathbf{0}, \mathbf{I}_p)$ and independent of the training sample $\mathbf{X}$, $P_{\mathbf{Z}}$ is the probability

measure with respect to $\mathbf{Z}$ given $\mathbf{X}$, and in the fourth line, we use $\widehat{\mathbf{a}}_{ij} = -\widehat{\mathbf{a}}_{ji}$.

To calculate the probability $P_{\mathbf{Z}}\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{Z} > \widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i), \quad \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} < \mathbf{a}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i)\right)$, we note that it is equal to $P_{\mathbf{Z}}\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{Z} > \widehat{d}_{ji}, \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} < d_{ji}\right)$ by the definitions of d_{ji} and $\widehat{d}_{ji}$. First, we note that by the definition (3.36) of $\mathbf{a}_{ji}$ and Lemma 8,

$$\begin{aligned}
&\|\mathbf{a}_{ji}\|_2^2 = \|\boldsymbol{\Sigma}^{1/2}\mathbf{D}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)\|_2^2 = \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)\|_2^2 = \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)\|_2^2 \\
=&(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)
\end{aligned}$$

which together with Lemma 9 give

$$2Kc_1 \leq \|\mathbf{a}_{ji}\|_2^2 \leq 2\lambda_{max}(\boldsymbol{\Delta}) \tag{4.88}$$

Moreover, we have

$$d_{ji} = \mathbf{a}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i) = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)/2 = \frac{1}{2}\|\mathbf{a}_{ji}\|_2^2. \tag{4.89}$$

Since the subscript ij is fixed during the calculation, for simplicity, we omit it in the following. We also omit the subscript $\mathbf{Z}$ in $P_{\mathbf{Z}}$. Note the orthogonal decomposition $\mathbf{a} = t\widehat{\mathbf{a}} + \mathbf{a}_\perp$ and the relationship $d = \frac{1}{2}\|\mathbf{a}\|_2^2$ by (4.89). By the conditions in (4.141),

$$\begin{aligned}
&\|\mathbf{a}\|_2^2 - \|\widehat{\mathbf{a}}\|_2^2 = \|\mathbf{a}\|_2^2 O_p(\delta_n), \quad t = 1 + O_p(\delta_n), \quad d - \widehat{d} = \|\widehat{\mathbf{a}}\|_2^2 O_p(\delta_n), \\
\text{we have} \quad &\|\mathbf{a}_\perp\|_2^2 = \|\mathbf{a}\|_2^2 - t^2\|\widehat{\mathbf{a}}\|_2^2 = \|\widehat{\mathbf{a}}\|_2^2 O_p(\delta_n), \quad \widehat{d} = \|\widehat{\mathbf{a}}\|_2^2(\frac{1}{2} + O_p(\delta_n)).
\end{aligned} \tag{4.90}$$

We first assume that $\mathbf{a}_\perp \neq \mathbf{0}$. Define

$$\mathbf{W} = \frac{\widehat{\mathbf{a}}^{\mathrm{T}}}{\|\widehat{\mathbf{a}}\|_2}\mathbf{Z} \sim N(0,1), \quad \mathbf{V} = -\frac{\mathbf{a}_\perp^{\mathrm{T}}}{\|\mathbf{a}_\perp\|_2}\mathbf{Z} \sim N(0,1),$$

where the distributions are conditional on the training sample $\mathbf{X}$. Since $(\mathbf{W}, \mathbf{V})$ is jointly normal and $\widehat{\mathbf{a}}$ and $\mathbf{a}_\perp$ are orthogonal, $\mathbf{W}$ and $\mathbf{V}$ are uncorrelated and hence independent. Let ϕ and Φ are the density and cumulative distribution functions of $N(0,1)$, respectively. Define

$$\eta = \frac{|t\widehat{d} - d|}{t} + \frac{\|\mathbf{a}_\perp\|_2}{t}\sqrt{\log\left[(\|\mathbf{a}\|_2^2\delta_n)^{-1}\right]}. \tag{4.91}$$

Then

$$
\begin{aligned}
&P\left(\widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{Z} > \widehat{d}, \mathbf{a}^{\mathrm{T}}\mathbf{Z} < d\right) = P\left(\widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{Z} > \widehat{d}, (t\widehat{\mathbf{a}} + \mathbf{a}_\perp)^{\mathrm{T}}\mathbf{Z} < d\right) \\
=&P\left(\mathbf{W} > \frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}, \quad t\|\widehat{\mathbf{a}}\|_2\mathbf{W} - \|\mathbf{a}_\perp\|_2\mathbf{V} < d\right) \\
=&\int_{\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}}^{\infty} \phi(w) P\left(\mathbf{V} > \frac{t\|\widehat{\mathbf{a}}\|_2 w - d}{\|\mathbf{a}_\perp\|_2}\right) dw = \int_{\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}}^{\infty} \phi(w)\left[1 - \Phi\left(\frac{t\|\widehat{\mathbf{a}}\|_2 w - d}{\|\mathbf{a}_\perp\|_2}\right)\right] dw \\
=&\int_{\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}}^{\frac{\widehat{d}+\eta}{\|\widehat{\mathbf{a}}\|_2}} + \int_{\frac{\widehat{d}+\eta}{\|\widehat{\mathbf{a}}\|_2}}^{\infty} \phi(w)\left[1 - \Phi\left(\frac{t\|\widehat{\mathbf{a}}\|_2 w - d}{\|\mathbf{a}_\perp\|_2}\right)\right] dw \\
\leq&\int_{\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}}^{\frac{\widehat{d}+\eta}{\|\widehat{\mathbf{a}}\|_2}} \phi(w) dw + \int_{\frac{\widehat{d}+\eta}{\|\widehat{\mathbf{a}}\|_2}}^{\infty} \phi(w)\left[1 - \Phi\left(\frac{t\|\widehat{\mathbf{a}}\|_2 \frac{\widehat{d}+\eta}{\|\widehat{\mathbf{a}}\|_2} - d}{\|\mathbf{a}_\perp\|_2}\right)\right] dw \\
\leq&\frac{\eta}{\|\widehat{\mathbf{a}}\|_2}\phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}) + \left[1 - \Phi\left(\frac{t\widehat{d} - d + t\eta}{\|\mathbf{a}_\perp\|_2}\right)\right] \int_{\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}}^{\infty} \phi(w) dw. \qquad (4.92)
\end{aligned}
$$

Since $\widehat{d}/\|\widehat{\mathbf{a}}\|_2 = \|\mathbf{a}\|_2(1/2 + O_p(\delta_n))$ and by (4.88), $\|\mathbf{a}\|_2$ is bounded below, it follows from the inequality:

$$
(1 - \frac{1}{x^2})\phi(x) \leq x[1 - \Phi(x)] \leq \phi(x), \quad \forall x > 0, \qquad (4.93)
$$

that there exists a constant $C_3 > 0$ independent of p such that with probability converging to 1,

$$
C_3\phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}) \leq \frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}[1 - \Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]. \qquad (4.94)
$$

By (4.88), (4.90),(4.93), (4.94), and the definition (4.91) of η, the right hand side of (4.92),

$$\begin{aligned}
&\frac{\eta}{\|\widehat{\mathbf{a}}\|_2}\phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}) + \left[1-\Phi\left(\frac{t\widehat{d}-d+t\eta}{\|\mathbf{a}_\perp\|_2}\right)\right][1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]\\
\leq&\frac{1}{C_3}\frac{\eta}{\|\widehat{\mathbf{a}}\|_2}\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]\\
&+\left[1-\Phi\left(\frac{t\widehat{d}-d+|t\widehat{d}-d|+\|\mathbf{a}_\perp\|_2\sqrt{\log\left[(\|\mathbf{a}\|_2^2\delta_n)^{-1}\right]}}{\|\mathbf{a}_\perp\|_2}\right)\right][1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]\\
\leq&\frac{1}{C_3}(\frac{1}{2}+O_p(\delta_n))\eta[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]+\left[1-\Phi\left(\sqrt{\log\left[(\|\mathbf{a}\|_2^2\delta_n)^{-1}\right]}\right)\right][1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]\\
\leq&\frac{1}{C_3}(\frac{1}{2}+O_p(\delta_n))\eta[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]+\left[1-\Phi\left(\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right)\right][1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]\\
\leq&\left[\frac{\eta}{C_3}(\frac{1}{2}+O_p(\delta_n))+\frac{\phi\left(\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right)}{\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}}\right][1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]. \qquad (4.95)
\end{aligned}$$

By (4.90) and the definition (4.91) of η,

$$\begin{aligned}
&\frac{\eta}{C_3}(\frac{1}{2}+O_p(\delta_n))+\frac{\phi\left(\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right)}{\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}}\\
=&\|\mathbf{a}\|_2^2O_p(\delta_n)+\sqrt{\|\mathbf{a}\|_2^2O_p(\delta_n)\log\left[(\|\mathbf{a}\|_2^2\delta_n)^{-1}\right]}+O\left(\frac{\exp\left[-\left(\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right)^2/2\right]}{\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}}\right)\\
\leq&2\lambda_{max}(\mathbf{\Delta})O_p(\delta_n)+\sqrt{2\lambda_{max}(\mathbf{\Delta})O_p(\delta_n)\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}+O\left(\frac{\sqrt{2\lambda_{max}(\mathbf{\Delta})\delta_n}}{\sqrt{\log\left[(2\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}}\right)\\
=&O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n\log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right). \qquad (4.96)
\end{aligned}$$

Next, we estimate

$$\left|[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]-[1-\Phi(\frac{d}{\|\mathbf{a}\|_2})]\right|=\int_{r_1}^{r_2}\phi(x)dx,$$

where $r_1=\min\left(\widehat{d}/\|\widehat{\mathbf{a}}\|_2, d/\|\mathbf{a}\|_2\right)$, $r_2=\max\left(\widehat{d}/\|\widehat{\mathbf{a}}\|_2, d/\|\mathbf{a}\|_2\right)$. By (4.93) and (4.90),

$$\begin{aligned}
&\int_{r_1}^{r_2}\phi(x)dx\leq(r_2-r_1)\phi(r_1)=(r_2-r_1)O(r_1[1-\Phi(r_1)])=(r_2-r_1)r_1O([1-\Phi(r_1)])\\
=&\left|\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}-\frac{d}{\|\mathbf{a}\|_2}\right|r_1O([1-\Phi(r_1)])=\left|\|\widehat{\mathbf{a}}\|_2(\frac{1}{2}+O_p(\delta_n))-\frac{1}{2}\|\mathbf{a}\|_2\right|\frac{1}{2}\|\mathbf{a}\|_2O([1-\Phi(r_1)])\\
\leq&2\lambda_{max}(\mathbf{\Delta})O_p(\delta_n)O([1-\Phi(r_1)]).
\end{aligned}$$

Therefore,

$$\left|[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]-[1-\Phi(\frac{d}{\|\mathbf{a}\|_2})]\right| \leq 2\lambda_{max}(\mathbf{\Delta})O_p(\delta_n)O([1-\Phi(\frac{d}{\|\mathbf{a}\|_2})]) = o_p(1)O([1-\Phi(\frac{d}{\|\mathbf{a}\|_2})]),$$

and hence

$$[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})] = [1-\Phi(\frac{d}{\|\mathbf{a}\|_2})](1+o_p(1)), \tag{4.97}$$

By (4.92), (4.95), (4.96) and (4.97)

$$\begin{aligned} P\left(\widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{Z} > \widehat{d}, \mathbf{a}^{\mathrm{T}}\mathbf{Z} < d\right) &\leq O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right)[1-\Phi(\frac{d}{\|\mathbf{a}\|_2})] \\ &= O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) P\left(\mathbf{a}^{\mathrm{T}}\mathbf{Z} > d\right). \end{aligned} \tag{4.98}$$

Now we consider the case of $\mathbf{a}_\perp = \mathbf{0}$. By similar arguments as those for (4.97),

$$\begin{aligned} & P\left(\widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{Z} > \widehat{d}, \mathbf{a}^{\mathrm{T}}\mathbf{Z} < d\right) = P\left(\widehat{\mathbf{a}}^{\mathrm{T}}\mathbf{Z} > \widehat{d}, (t\widehat{\mathbf{a}})^{\mathrm{T}}\mathbf{Z} < d\right) \qquad (4.99) \\ =& P\left(\mathbf{W} > \frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2}, \quad t\|\widehat{\mathbf{a}}\|_2\mathbf{W} < d\right) \leq \left|[1-\Phi(\frac{\widehat{d}}{\|\widehat{\mathbf{a}}\|_2})]-[1-\Phi(\frac{d}{t\|\widehat{\mathbf{a}}\|_2})]\right| \\ \leq& \lambda_{max}(\mathbf{\Delta})O_p(\delta_n)P\left(\mathbf{a}^{\mathrm{T}}\mathbf{Z} > d\right) \leq O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) P\left(\mathbf{a}^{\mathrm{T}}\mathbf{Z} > d\right). \end{aligned}$$

Combining (4.98) and (4.99), and note the fact that given the new observation $\mathbf{x}$ belonging to the ith class, $\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)$ have the same distribution as $\mathbf{Z}$, we have

$$\begin{aligned} & P_{\mathbf{Z}}\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{Z} > \widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji}-\boldsymbol{\mu}_i), \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} < \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{b}_{ji}-\boldsymbol{\mu}_i)\right) \qquad (4.100) \\ \leq& O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) P_{\mathbf{Z}}\left(\mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} > d_{ji}\right) \\ \leq& O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) \\ & \times P_{\mathbf{Z}}\left(\mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i) > \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{b}_{ji}-\boldsymbol{\mu}_i) | \mathbf{x} \in \text{the } i\text{th class}\right) \\ =& O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) P\left(\mathbf{a}_{ji}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\mathbf{b}_{ji}) > 0 | \mathbf{x} \in \text{the } i\text{th class}\right) \\ \leq& O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) P\left(T_{OPT}(\mathbf{x}) \notin \text{the } i\text{th class} | \mathbf{x} \in \text{the } i\text{th class}\right) \\ \leq& O_p\left(\sqrt{\lambda_{max}(\mathbf{\Delta})\delta_n \log\left[(\lambda_{max}(\mathbf{\Delta})\delta_n)^{-1}\right]}\right) KR_{OPT} \end{aligned}$$

where the O_p term is uniform for all $1 \le i \ne j \le K$. Now by (4.86), (4.87) and (4.100),

$$\begin{aligned}
&R_T(\mathbf{X}) - R_{OPT} \\
\le&\frac{1}{K}\sum_{i=1}^{K}\sum_{j\neq i}\left[P_{\mathbf{Z}}\left(\widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\mathbf{Z} < \widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i), \quad \mathbf{a}_{ji}^{\mathrm{T}}\mathbf{Z} > \mathbf{a}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathbf{b}_{ji} - \boldsymbol{\mu}_i)\right)\right], \\
\le&O_p\left(K^2\sqrt{\lambda_{max}(\boldsymbol{\Delta})\delta_n \log\left[(\lambda_{max}(\boldsymbol{\Delta})\delta_n)^{-1}\right]}\right) R_{OPT}
\end{aligned}$$

By Lemma 9, $\lambda_{max}(\boldsymbol{\Delta}) = K\lambda_{max}(\boldsymbol{\Xi}) = K\lambda_1(\boldsymbol{\Xi})$. Moreover, in this paper, we assume that K is fixed, we have

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \le O_p\left(\sqrt{\lambda_1(\boldsymbol{\Xi})\delta_n \log\left[\{\lambda_1(\boldsymbol{\Xi})\delta_n\}^{-1}\right]}\right).$$

□

4.10 Proof of Theorem 3.3.9

Theorem (Theorem 3.3.9)**.** *Suppose that Conditions 1-2 hold, $s_n \to 0$ and $\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n \to 0$ as $n, p \to \infty$. Then the classification rule* (3.15) *of our sparse Fisher's discriminant analysis method is asymptotically optimal. Moreover, we have*

$$\frac{R_T(\mathbf{X})}{R_{OPT}} - 1 \le O_p\left(\sqrt{\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n \log\left[\{\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n\}^{-1}\right]}\right). \tag{4.101}$$

Proof of Theorem 3.3.9. To apply Theorem 3.3.8, we first verify the conditions (4.141) for $\delta_n = \Lambda_p^2 s_n$. In this proof, let $\boldsymbol{\delta}_{ji} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i$ and $\widehat{\boldsymbol{\delta}}_{ji} = \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_i$ for any $1 \le i, j \le K$. We only consider the elements in $\Omega_n \bigcap \widetilde{\Omega}_n$ (see their definitions (3.27) and (4.62)). The complement of $\Omega_n \bigcap \widetilde{\Omega}_n$ has a probability less than $3p^{-1} \to 0$ as $n, p \to \infty$. Therefore, by the definition (4.62) of $\widetilde{\Omega}_n$, we have

$$\|\boldsymbol{\delta}_{ji} - \widehat{\boldsymbol{\delta}}_{ji}\|_\infty \le 2\widetilde{C}\sqrt{\frac{K\log p}{n}} = 2\widetilde{C}s_n. \tag{4.102}$$

Let $\mathbf{P}_{K-1}$ and $\widetilde{\mathbf{P}}_{K-1}$ be the orthogonal projection matrices of the two subspaces of

$$\mathbf{V}_{K-1} = \text{span}\{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_{K-1}\}, \quad \widetilde{\mathbf{V}}_{K-1} = \text{span}\{\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, \cdots, \widehat{\boldsymbol{\gamma}}_{K-1}\}.$$

Let $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{K-1}]$ and $\widehat{\boldsymbol{\Gamma}} = [\widehat{\boldsymbol{\gamma}}_1, \cdots, \widehat{\boldsymbol{\gamma}}_{K-1}]$, both of which are $p \times (K-1)$ matrices. Let $\mathbf{K} = \widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\Gamma}}$ which is a symmetric $(K-1) \times (K-1)$ matrix with the (k,l)-th entry equal to $\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_l = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_l$. Then we have

$$\mathbf{P}_{K-1} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}, \qquad \widetilde{\mathbf{P}}_{K-1} = \widehat{\boldsymbol{\Gamma}}\mathbf{K}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}, \quad \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{I}_{K-1}, \tag{4.103}$$

where $\mathbf{I}_{K-1}$ is the $K-1$ dimensional identity matrix, because $\boldsymbol{\gamma}_k$, $1 \le k \le K-1$, are orthonormal vectors. By the definition (3.16),

$$\begin{aligned} \widehat{\mathbf{K}} = (\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1})^{\mathrm{T}}\, \widehat{\boldsymbol{\Sigma}}\, (\widehat{\boldsymbol{\alpha}}_1, \cdots, \widehat{\boldsymbol{\alpha}}_{K-1}) = \widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Gamma}}, \\ \widehat{\mathbf{D}} = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}. \end{aligned} \tag{4.104}$$

We consider the first equality in (4.141). By (3.36) and Lemmas 8 and 9,

$$\begin{aligned} &\mathbf{a}_{ji} = \boldsymbol{\Sigma}^{1/2}\mathbf{D}\boldsymbol{\delta}_{ji} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_{ji} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}, \\ &2Kc_1 \le (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_{ij}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 \le 2K\lambda_1(\boldsymbol{\Xi}). \end{aligned} \tag{4.105}$$

By (4.103),

$$\begin{aligned} \mathbf{D} &= \sum_{k=1}^{K-1} \boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}} = \boldsymbol{\Sigma}^{-1/2}\sum_{k=1}^{K-1} \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{\Sigma}^{-1/2}\mathbf{P}_{K-1}\boldsymbol{\Sigma}^{-1/2}. \end{aligned} \tag{4.106}$$

Hence, by (4.105), (4.106) and Lemma 8,

$$\begin{aligned} &\mathbf{P}_{K-1}\mathbf{a}_{ji} = \mathbf{P}_{K-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\mathbf{P}_{K-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ij} = \boldsymbol{\Sigma}^{1/2}\mathbf{D}\boldsymbol{\delta}_{ji} \\ =&\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_{ji} = \mathbf{a}_{ji}. \end{aligned} \tag{4.107}$$

For any (k,l), by the definition of Ω_n and Theorem 3.3.7,

$$\begin{aligned}&|(\widehat{\mathbf{K}})_{kl}-(\mathbf{K})_{kl}|=|\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_l-\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_l|\le\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\|_\infty\|\widehat{\boldsymbol{\alpha}}_k\|_1\|\widehat{\boldsymbol{\alpha}}_l\|_1\\ \le&(\tau_n/C_2)D_{k,1}\Lambda_pD_{l,1}\Lambda_p\le h_1\Lambda_p^2s_n,\end{aligned}\tag{4.108}$$

where $h_1=(C/C_2)\max_{1\le k,l\le K-1}(D_{k,1}D_{l,1})$. Because $\widehat{\mathbf{K}}-\mathbf{K}$ is symmetric, by (4.108),

$$\|\widehat{\mathbf{K}}-\mathbf{K}\|\le\max_{1\le k\le K-1}\sum_{l=1}^{K-1}|(\widehat{\mathbf{K}})_{kl}-(\mathbf{K})_{kl}|\le(K-1)h_1\Lambda_p^2s_n=o(1).\tag{4.109}$$

Because $\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_l=1$ if $k=l$, and equal to 0 if $k\neq l$, by (4.77) in the proof of Theorem 3.3.7, as n large enough,

$$\begin{aligned}&|(\mathbf{K})_{kl}-(\mathbf{I}_{K-1})_{kl}|=|\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}_l-\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_l|=|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_l-\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\gamma}_l|\\ \le&|(\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l-\boldsymbol{\gamma}_l)|+|\boldsymbol{\gamma}_k^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_l-\boldsymbol{\gamma}_l)|+|(\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^{\mathrm{T}}\boldsymbol{\gamma}_l|\\ \le&\|\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k\|_2\|\widehat{\boldsymbol{\gamma}}_l-\boldsymbol{\gamma}_l\|_2+\|\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k\|_2+\|\widehat{\boldsymbol{\gamma}}_l-\boldsymbol{\gamma}_l\|_2\le h_2\sqrt{\Lambda_p^2s_n},\end{aligned}$$

where h_2 is a constant independent of p and n. Therefore,

$$\begin{aligned}&\|\mathbf{K}-\mathbf{I}_{K-1}\|\le\max_{1\le k\le K-1}\sum_{l=1}^{K-1}|(\widehat{\mathbf{K}})_{kl}-(\mathbf{I}_{K-1})_{kl}|\le(K-1)h_2\sqrt{\Lambda_p^2s_n}=o(1),\\ &\text{and hence }\|\mathbf{K}\|=1+o(1).\end{aligned}\tag{4.110}$$

By the Taylor's expansion,

$$\begin{aligned}&\|\mathbf{K}^{-1}\|=\|[\mathbf{I}_{K-1}-(\mathbf{I}_{K-1}-\mathbf{K})]^{-1}\|=\|\mathbf{I}_{K-1}+(\mathbf{I}_{K-1}-\mathbf{K})+(\mathbf{I}_{K-1}-\mathbf{K})^2+\cdots\|\\ \le&\|\mathbf{I}_{K-1}\|+\|\mathbf{I}_{K-1}-\mathbf{K}\|+\|\mathbf{I}_{K-1}-\mathbf{K}\|^2+\cdots=\frac{1}{1-\|\mathbf{I}_{K-1}-\mathbf{K}\|}=1+o(1).\end{aligned}\tag{4.111}$$

(4.109)-(4.111) imply that $\|\widehat{\mathbf{K}}-\mathbf{I}_{K-1}\|=o(1)$. By the same argument as in (4.111), $\|\widehat{\mathbf{K}}^{-1}\|\le$

$1+o(1)$. Then by (4.109),

$$\|\widehat{\mathbf{K}}^{-1}-\mathbf{K}^{-1}\| \le \|\widehat{\mathbf{K}}^{-1}\|\|\widehat{\mathbf{K}}-\mathbf{K}\|\|\mathbf{K}^{-1}\| \le (K-1)h_1\Lambda_p^2 s_n(1+o(1))^2 \le 2(K-1)h_1\Lambda_p^2 s_n, \tag{4.112}$$

as n is large enough. Now by (4.104),

$$\begin{aligned}\hat{\mathbf{a}}_{ji} &= \boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{D}}\hat{\boldsymbol{\delta}}_{ji} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\delta}}_{ji} = \hat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\delta}}_{ji}\\ =&\hat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji}) + \hat{\boldsymbol{\Gamma}}(\widehat{\mathbf{K}}^{-1}-\mathbf{K}^{-1})\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji} + \hat{\boldsymbol{\Gamma}}\mathbf{K}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}. \end{aligned}\tag{4.113}$$

We estimate the first term on the right hand side of (4.113). Let $\mathbf{g} = \widehat{\mathbf{K}}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji})$ and g_k denote its k-th coordinate, $1\le k\le K-1$. By (4.172), (4.102), Theorem 3.3.7 and (4.62),

$$\begin{aligned}&\|\hat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji})\|_2 = \|\hat{\boldsymbol{\Gamma}}\mathbf{g}\|_2 = \|\sum_{k=1}^{K-1}\hat{\boldsymbol{\gamma}}_k g_k\|_2 \le \sum_{k=1}^{K-1}\|\hat{\boldsymbol{\gamma}}_k\|_2|g_k| \le \|\mathbf{g}\|_1\\ \le&\sqrt{K-1}\|\mathbf{g}\|_2 \le \sqrt{K-1}\|\widehat{\mathbf{K}}^{-1}\|\|\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji})\|_2\\ \le&\sqrt{K-1}\|\widehat{\mathbf{K}}^{-1}\|\|\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji})\|_1\\ =&\sqrt{K-1}\|\widehat{\mathbf{K}}^{-1}\|\sum_{k=1}^{K-1}|\hat{\boldsymbol{\alpha}}_k^{\mathrm{T}}(\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji})| \le \sqrt{K-1}\|\widehat{\mathbf{K}}^{-1}\|\sum_{k=1}^{K-1}\|\hat{\boldsymbol{\alpha}}_k\|_1\|\hat{\boldsymbol{\delta}}_{ji}-\boldsymbol{\delta}_{ji}\|_\infty\\ \le&\sqrt{K-1}(1+o(1))(K-1)D_{k,1}\Lambda_p 2\widetilde{C}s_n = O(\Lambda_p s_n) = O(\Lambda_p^2 s_n) \le O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2, \end{aligned}\tag{4.114}$$

where the second equality in the last line is due to $\Lambda_p \ge \|\boldsymbol{\alpha}_k\|_1 \ge \|\boldsymbol{\alpha}_k\|_2 \ge c_0^{-1/2}$ and the last inequality is due to (4.105). Similarly, by (4.108) and (4.105), for the second term on the right hand side of (4.113), we have

$$\begin{aligned}&\|\hat{\boldsymbol{\Gamma}}(\widehat{\mathbf{K}}^{-1}-\mathbf{K}^{-1})\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}\|_2 \le \|\hat{\boldsymbol{\Gamma}}(\widehat{\mathbf{K}}^{-1}-\mathbf{K}^{-1})\hat{\boldsymbol{\Gamma}}^{\mathrm{T}}\|\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}\|_2\\ =&\|\sum_{1\le k,l\le K-1}(\widehat{\mathbf{K}}_{kl}-\mathbf{K}_{kl})\hat{\boldsymbol{\gamma}}_k\hat{\boldsymbol{\gamma}}_l^{\mathrm{T}}\|\|\mathbf{a}_{ji}\|_2 \le h_1\Lambda_p^2 s_n \sum_{1\le k,l\le K-1}\|\hat{\boldsymbol{\gamma}}_k\|_2\|\hat{\boldsymbol{\gamma}}_l\|_2\|\mathbf{a}_{ji}\|_2\\ =&(K-1)^2 h_1\Lambda_p^2 s_n\|\mathbf{a}_{ji}\|_2. \end{aligned}\tag{4.115}$$

As for the third term on the right hand side of (4.113), by (4.103), we have

$$\|\widehat{\mathbf{\Gamma}}\mathbf{K}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}\|_2 = \|\widetilde{\mathbf{P}}_{K-1}\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_{ji}\|_2 = \|\widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2. \tag{4.116}$$

Now by (4.113)-(4.116) and (4.107), we have

$$\begin{aligned}\|\widehat{\mathbf{a}}_{ji}\|_2^2 &= \|\widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2^2 + O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 - \|\mathbf{a}_{ji} - \widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2^2 + O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2 \\ =&\|\mathbf{a}_{ji}\|_2^2 - \|\mathbf{P}_{K-1}\mathbf{a}_{ji} - \widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2^2 + O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2. \end{aligned} \tag{4.117}$$

By the second bound in (4.77), (4.110) and (4.111),

$$\begin{aligned}\|\mathbf{P}_{K-1}\mathbf{a}_{ji} - \widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2^2 &\le \|\mathbf{P}_{K-1} - \widetilde{\mathbf{P}}_{K-1}\|^2\|\mathbf{a}_{ji}\|_2^2 \le \|\mathbf{\Gamma}\mathbf{\Gamma}^{\mathrm{T}} - \widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{K}^{-1}\widehat{\mathbf{\Gamma}}\|^2\|\mathbf{a}_{ji}\|_2^2 \\ \le&\|\mathbf{\Gamma}\mathbf{\Gamma}^{\mathrm{T}} - \widehat{\mathbf{\Gamma}}^{\mathrm{T}}\widehat{\mathbf{\Gamma}}\|^2\|\mathbf{a}_{ji}\|_2^2 + \|\widehat{\mathbf{\Gamma}}^{\mathrm{T}}(\mathbf{I}_{K-1} - \mathbf{K}^{-1})\widehat{\mathbf{\Gamma}}\|^2\|\mathbf{a}_{ji}\|_2^2 = O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2. \end{aligned} \tag{4.118}$$

Hence, by (4.117) and (4.118),

$$\|\widehat{\mathbf{a}}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 + \|\mathbf{a}_{ji}\|_2^2 O(\Lambda_p^2 s_n). \tag{4.119}$$

Therefore, the first equality in the condition (4.141) is verified. For the second one, by the orthogonal decomposition (4.89), we have $t_{ji} = \mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji}/\|\widehat{\mathbf{a}}_{ji}\|_2^2$. By (4.105) and (4.104),

$$\begin{aligned}\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji} &= \boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}}_{ji} = \boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\boldsymbol{\delta}}_{ji} \\ =&\boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\boldsymbol{\delta}}_{ji} - \boldsymbol{\delta}_{ij}) + \boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_{ij} \\ =&\boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\boldsymbol{\delta}}_{ji} - \boldsymbol{\delta}_{ij}) + \mathbf{a}_{ji}^{\mathrm{T}}\widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji} \\ =&\boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\boldsymbol{\delta}}_{ji} - \boldsymbol{\delta}_{ij}) + \mathbf{a}_{ji}^{\mathrm{T}}\widetilde{\mathbf{P}}_{K-1}\widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}. \end{aligned} \tag{4.120}$$

By (4.114), the first term in the last line of (4.120)

$$\begin{aligned}|\boldsymbol{\delta}_{ij}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\boldsymbol{\delta}}_{ji} - \boldsymbol{\delta}_{ij})| &\le \|\mathbf{\Sigma}^{-1/2}\boldsymbol{\delta}_{ij}\|_2\|\widehat{\mathbf{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\mathbf{\Gamma}}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2}(\widehat{\boldsymbol{\delta}}_{ji} - \boldsymbol{\delta}_{ji})\|_2 \\ =&\|\mathbf{a}_{ji}\|_2 O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2 = O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2 \end{aligned} \tag{4.121}$$

By (4.117) and (4.118), the second term in the last line of (4.120) is equal to

$$\|\widetilde{\mathbf{P}}_{K-1}\mathbf{a}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 - \|(\mathbf{P}_{K-1} - \widetilde{\mathbf{P}}_{K-1})\mathbf{a}_{ji}\|_2^2 = \|\mathbf{a}_{ji}\|_2^2 - O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2,$$

which together with (4.120) and (4.121) imply that

$$\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji} = \|\mathbf{a}_{ji}\|_2^2 + \|\mathbf{a}_{ji}\|_2^2 O(\Lambda_p^2 s_n), \text{ and hence } t_{ji} = \frac{\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji}}{\|\widehat{\mathbf{a}}_{ji}\|_2^2} = 1 + O(\Lambda_p^2 s_n), \tag{4.122}$$

by (4.119). The second condition in (4.141) is verified. For the last condition in (4.141), by the definitions of d_{ji} and $\widehat{d}_{ji}$, and (4.105),

$$\begin{aligned}
&d_{ji} - \widehat{d}_{ji} = \frac{1}{2}\|\mathbf{a}_{ji}\|_2^2 - \widehat{\mathbf{a}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\widehat{\mathbf{b}}_{ji} - \boldsymbol{\mu}_i) = \frac{1}{2}\|\mathbf{a}_{ji}\|_2^2 - \widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\widehat{\mathbf{D}}\left(\frac{\bar{\mathbf{x}}_j + \bar{\mathbf{x}}_i}{2} - \boldsymbol{\mu}_i\right)\\
=&\frac{1}{2}\|\mathbf{a}_{ji}\|_2^2 - \widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\widehat{\mathbf{D}}\left(\frac{\bar{\mathbf{x}}_j + \bar{\mathbf{x}}_i}{2} - \frac{\boldsymbol{\mu}_j + \boldsymbol{\mu}_i}{2} + \frac{(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)}{2}\right)\\
=&\frac{1}{2}\|\mathbf{a}_{ji}\|_2^2 - \widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\widehat{\mathbf{D}}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right) - \frac{1}{2}\widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\widehat{\mathbf{D}}\boldsymbol{\delta}_{ji}\\
=&\frac{\|\mathbf{a}_{ji}\|_2^2}{2} - \widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right) - \frac{\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji}}{2}\\
=&\frac{\|\mathbf{a}_{ji}\|_2^2}{2} - \frac{\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji}}{2} - \widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right),
\end{aligned} \tag{4.123}$$

where in the last line, we use the fact that $\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}} = \widetilde{\mathbf{P}}_{K-1}$ is a projection matrix and hence $\widetilde{\mathbf{P}}_{K-1}^2 = \widetilde{\mathbf{P}}_{K-1}$. We estimate the last term on the right hand side of (4.125),

$$\begin{aligned}
&\left|\widehat{\boldsymbol{\delta}}_{ji}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right)\right|\\
\le&\|\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\delta}}_{ji}\|_2\left\|\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right)\right\|_2\\
=&\|\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{D}}\widehat{\boldsymbol{\delta}}_{ji}\|_2\left\|\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right)\right\|_2\\
=&\|\widehat{\mathbf{a}}_{ji}\|_2\left\|\widehat{\boldsymbol{\Gamma}}\widehat{\mathbf{K}}^{-1}\widehat{\boldsymbol{\Gamma}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\left(\frac{\bar{\mathbf{x}}_i - \boldsymbol{\mu}_i}{2} + \frac{\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j}{2}\right)\right\|_2 \le \|\widehat{\mathbf{a}}_{ji}\|_2 O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2,
\end{aligned} \tag{4.124}$$

where the last inequality can be obtained by a similar argument as in (4.114). it follows from

(4.125) and (4.124) that

$$d_{ji} - \widehat{d}_{ji} = \frac{\|\mathbf{a}_{ji}\|_2^2}{2} - \frac{\mathbf{a}_{ji}^{\mathrm{T}}\widehat{\mathbf{a}}_{ji}}{2} + \|\widehat{\mathbf{a}}_{ji}\|_2 O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2 = O(\Lambda_p^2 s_n)\|\mathbf{a}_{ji}\|_2^2, \tag{4.125}$$

where the last inequality is due to (4.119) and (4.122). Hence, the thir condition in (4.141) is verified. Hence, we can apply Theorem 3.3.8 to obtain the theorem.

□

4.11 Proof of Lemmas

The rest of this section is devoted to the proof of all technical lemmas.

4.11.1 Proof of Lemma 3

Proof of Lemma 3

By (4.97), we have

$$\begin{aligned}
\int_a^b |f(a) + (t-a)f'(a)|dt =& \int_a^b |f(t) - \int_a^b G(t,s)f''(s)ds|dt \\
\leq& \int_a^b |f(t)|dt + \int_a^b |\int_a^b G(t,s)f''(s)ds|dt \\
\leq& \sqrt{b-a}\|f\| + \sqrt{b-a}\|G\|\|f''\| \\
\leq& \sqrt{b-a}\|f\|_\alpha + \sqrt{b-a}\frac{\|G\|}{\sqrt{\alpha}}\|f\|_\alpha \\
=& \sqrt{b-a}\left(1 + \frac{\|G\|}{\sqrt{\alpha}}\right)\|f\|_\alpha.
\end{aligned}$$

On the other hand, if we can show that

$$\int_a^b |f(a) + (t-a)f'(a)|dt \geq (\sqrt{2}-1)(b-a)|f(a)|, \tag{4.126}$$

then we have

$$(\sqrt{2}-1)(b-a)|f(a)| \leq \sqrt{b-a}\left(1 + \frac{\|G\|}{\sqrt{\alpha}}\right)\|f\|_\alpha.$$

The second inequality in the lemma follows from

$$\begin{aligned}\frac{(b-a)^2}{2}|f'(a)| &= \int_a^b |(t-a)f'(a)|dt \\ &\leq \int_a^b \left[|f(a)| + |f(a)+(t-a)f'(a)|\right] dt \\ &\leq (b-a)|f(a)| + \sqrt{b-a}\left(1+\frac{\|G\|}{\sqrt{\alpha}}\right)\|f\|_\alpha \\ &\leq (\sqrt{2}+2)\sqrt{b-a}\left(1+\frac{\|G\|}{\sqrt{\alpha}}\right)\|f\|_\alpha.\end{aligned}$$

Now we prove the inequality (4.126). Since if $f(a)=0$, (4.126) is trivial, without loss of generality, we assume that $f(a)>0$. We first consider the case that

$$f(a)+(t-a)f'(a)>0, \quad \forall a \leq t \leq b.$$

Let $t=b$. We have

$$f(a)+(b-a)f'(a)>0, \text{ hence } f'(a) > -\frac{f(a)}{b-a}.$$

Therefore,

$$\begin{aligned}&\int_a^b |f(a)+(t-a)f'(a)|dt = \int_a^b \left(f(a)+(t-a)f'(a)\right) dt \\ =&f(a)(b-a)+\frac{(b-a)^2}{2}f'(a) > f(a)(b-a) - \frac{(b-a)^2}{2}\frac{f(a)}{b-a} \\ =&\frac{(b-a)}{2}f(a). \end{aligned} \tag{4.127}$$

Now we consider the case that there exists $x \in [a,b]$ such that

$$f(a)+(x-a)f'(a)=0.$$

Then

$$f'(a) = -\frac{f(a)}{x-a}. \tag{4.128}$$

In this case,

$$\begin{aligned}
&\int_a^b |f(a) + (t-a)f'(a)|dt \\
=&\int_a^x (f(a) + (t-a)f'(a))\, dt - \int_x^b (f(a) + (t-a)f'(a))\, dt \\
=&(x-a)f(a) + \frac{(x-a)^2}{2}f'(a) - (b-x)f(a) - \frac{[(b-a)^2 - (x-a)^2]}{2}f'(a) \\
=&2(x-a)f(a) - (b-a)f(a) + (x-a)^2 f'(a) - \frac{(b-a)^2}{2}f'(a) \\
=&2(x-a)f(a) - (b-a)f(a) - (x-a)f(a) + \frac{(b-a)^2}{2(x-a)}f(a) \qquad \text{by (4.128)} \\
=&(x-a)f(a) + \frac{(b-a)^2}{2(x-a)}f(a) - (b-a)f(a) \\
\geq&\sqrt{2}(b-a)f(a) - (b-a)f(a) = (\sqrt{2}-1)(b-a)f(a)
\end{aligned} \tag{4.129}$$

Now comparing (4.127) and (4.129), we can obtain (4.126).

4.11.2 Proof of Lemma 4

Proof of Lemma 4

For any $\beta_1, \beta_2 \in W_2^2([a,b])$, by the definition of $\hat{R}_\alpha$

$$\begin{aligned}
&\left\langle \beta_1, \hat{R}_\alpha \beta_2 \right\rangle_\alpha = \left\langle \beta_1, \hat{\Gamma}_n \beta_2 \right\rangle \\
=&A_{00}\beta_1(a)\beta_2(a) + A_{10}\beta_1'(a)\beta_2(a) + A_{01}\beta_1(a)\beta_2'(a) + A_{11}\beta_1'(a)\beta_2'(a) \\
&+ \beta_1(a)\int_a^b \xi_0(t)\beta_2''(t)dt + \beta_2(a)\int_a^b \xi_0(t)\beta_1''(t)dt + \beta_1'(a)\int_a^b \xi_1(t)\beta_2''(t)dt \\
&+ \beta_2'(a)\int_a^b \xi_1(t)\beta_1''(t)dt + \int_a^b\int_a^b \Xi(s,t)\beta_1''(s)\beta_2''(t)dsdt,
\end{aligned}$$

where

$$A_{00} = \int_a^b\int_a^b \hat{\Gamma}_n(s,t)dsdt, \quad A_{01} = A_{10} = \int_a^b\int_a^b (s-a)\hat{\Gamma}_n(s,t)dsdt,$$

$$A_{11} = \int_a^b \int_a^b (s-a)(t-a)\hat{\Gamma}_n(s,t)dsdt, \quad \xi_0(t) = \int_a^b \int_a^b \hat{\Gamma}_n(u,v)G(v,t)dudv$$
$$\xi_1(t) = \int_a^b \int_a^b (u-a)\hat{\Gamma}_n(u,v)G(v,t)dudv,$$
$$\Xi_0(s,t) = \int_a^b \int_a^b \hat{\Gamma}_n(u,v)G(u,s)G(v,t)dudv. \tag{4.130}$$

Define discretized versions of $X_p(t), 1 \le p \le n$,

$$\begin{aligned} X_p^{(m)}(t) =& X_p(t_{(1)})\chi_{[t_{(1)}, \frac{t_{(1)}+t_{(2)}}{2})}(t) + \sum_{q=2}^{m-1} X_p(t_{(q)})\chi_{[\frac{t_{(q-1)}+t_{(q)}}{2}, \frac{t_{(q)}+t_{(q+1)}}{2})}(t) \\ &+ X_p(t_{(m)})\chi_{[\frac{t_{(m-1)}+t_{(m)}}{2}, t_{(m)})}(t), \end{aligned} \tag{4.131}$$

where χ is the indicator function. Similarly, we define $\bar{X}^{(m)}$ by replacing X_p with the sample mean function $\bar{X}$ in (4.131).

We first compute $|\tilde{A}_{00} - A_{00}|$. By the definitions of $\hat{\Sigma}_{ql}$ and Y_{pq}, for any $1 \le q, l \le m$,

$$\begin{aligned} \hat{\Sigma}_{ql} =& \frac{1}{n}\sum_{p=1}^n \left(Y_{pq} - \bar{Y}_{\cdot q}\right)\left(Y_{pl} - \bar{Y}_{\cdot l}\right) \\ =& \frac{1}{n}\sum_{p=1}^n \left(X_p(t_{(q)}) + \epsilon_{pq} - \bar{X}(t_{(q)}) - \bar{\epsilon}_{\cdot q}\right)\left(X_p(t_{(l)}) + \epsilon_{pl} - \bar{X}(t_{(l)}) - \bar{\epsilon}_{\cdot l}\right) \\ =& \frac{1}{n}\sum_{p=1}^n (X_p(t_{(q)}) - \bar{X}(t_{(q)}))(X_p(t_{(l)}) - \bar{X}(t_{(l)})) + \Pi_{ql} \\ =& \hat{\Gamma}_n(t_{(q)}, t_{(l)}) + \Pi_{ql} \end{aligned}$$

where $\bar{Y}_{\cdot q} = \frac{1}{n}\sum_{p=1}^n Y_{pq}, \bar{\epsilon}_{\cdot q} = \frac{1}{n}\sum_{p=1}^n \epsilon_{pq}$, and

$$\begin{aligned} \Pi_{ql} =& \frac{1}{n}\sum_{p=1}^n \left(X_p(t_{(q)}) - \bar{X}(t_{(q)})\right)(\epsilon_{pl} - \bar{\epsilon}_{\cdot l}) \\ &+ \frac{1}{n}\sum_{p=1}^n (\epsilon_{pq} - \bar{\epsilon}_{\cdot q})\left(X_p(t_{(l)}) - \bar{X}(t_{(l)})\right) + \frac{1}{n}\sum_{p=1}^n (\epsilon_{pq} - \bar{\epsilon}_{\cdot q})(\epsilon_{pl} - \bar{\epsilon}_{\cdot l}). \end{aligned}$$

By the definitions of $\tilde{A}_{00}$ and A_{00} in (4.7) and (4.130), and the definition (4.131),

$$\begin{aligned}
&\tilde{A}_{00} - A_{00} \qquad (4.132)\\
=&\frac{1}{n}\sum_{p=1}^{n}\left(\int_a^b (X_p^{(m)}(s) - \bar{X}^{(m)}(s))ds\right)^2 - \frac{1}{n}\sum_{p=1}^{n}\left(\int_a^b (X_p(s) - \bar{X}(s))ds\right)^2\\
&+ 2\frac{1}{n}\sum_{p=1}^{n}\int_a^b (X_p^{(m)}(s) - \bar{X}^{(m)}(s))ds(\bar{\epsilon}_{p\cdot}^{(m)} - \bar{\epsilon}_{\cdot\cdot}^{(m)}) + \frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)} - \bar{\epsilon}_{\cdot\cdot}^{(m)})^2,
\end{aligned}$$

where

$$\bar{\epsilon}_{p\cdot}^{(m)} = \sum_{l=1}^{m}\epsilon_{pl}w_l, \quad \bar{\epsilon}_{\cdot\cdot}^{(m)} = \frac{1}{n}\sum_{p=1}^{n}\bar{\epsilon}_{p\cdot}^{(m)}. \qquad (4.133)$$

We provide an upper bound for $\max_q\{t_{(q)} - t_{(q-1)}\}$ in Case 2 in the following lemma.

Lemma 13. *In Case 2 (random case), we have*

$$\begin{aligned}
\max_q\{t_{(q)} - t_{(q-1)}\} &= O_p(\frac{\log m}{m}),\\
\varpi(\max_q\{t_{(q)} - t_{(q-1)}\}) &= O_p(\varpi(\frac{3\log m}{cm})),
\end{aligned}$$

where

$$\varpi(\delta) = \sup_{s,t\in[a,b],|s-t|\le\delta}\left[\Gamma(t,t) - 2\Gamma(s,t) + \Gamma(s,s)\right], \qquad (4.134)$$

has been defined in Section 3.4 and c is the lower bound of the density function $h(t)$ of the distribution of the observation points.

Proof of Lemma 13

For any $0 < x \le b - a$, let r be the positive integer satisfying

$$\frac{b-a}{r-1} > \frac{x}{2} \ge \frac{b-a}{r}. \qquad (4.135)$$

Then the event $\{\max_q\{t_{(q)} - t_{(q-1)}\} > x\}$ is contained in the event that there is at least one

of the following r intervals having no observation time points,

$$[a, a+\frac{b-a}{r}], [a+\frac{b-a}{r}, a+2\frac{b-a}{r}], \cdots, [a+(r-1)\frac{b-a}{r}, b].$$

Hence,

$$P(\max_q\{t_{(q)}-t_{(q-1)}\}>x) \le r(1-\frac{c(b-a)}{r})^m.$$

By (4.135),

$$\begin{aligned} r(1-\frac{c(b-a)}{r})^m &\le (\frac{x+2(b-a)}{x})(1-\frac{xc(b-a)}{x+2(b-a)})^m \\ &\le \frac{3(b-a)}{x}(1-\frac{xc(b-a)}{3(b-a)})^m \le \frac{3(b-a)}{x}e^{-\frac{cmx}{3}}. \end{aligned}$$

Hence, let $x=\frac{3\log m}{cm}$,

$$P(\max_q\{t_{(q)}-t_{(q-1)}\}>\frac{3\log m}{cm}) \le \frac{c(b-a)}{\log m}. \qquad \square$$

For the last terms on the right hand side of the equality (4.132), in Case 1, by directly computing its first moment, we have that its order is $O_p(\delta_m)$. In Case 2, by Lemma 4, for any small positive number τ, there exists $M>0$ (not depending m) such that

$$P(\max_q\{t_{(q)}-t_{(q-1)}\}>M\frac{\log m}{m}) \le \frac{\tau}{2}.$$

Let $Z=\max_q\{t_{(q)}-t_{(q-1)}\}$ and $\kappa=\frac{8\sigma^2(b-a)M}{\tau}$.

$$\begin{aligned} &P(\frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})^2>\frac{\kappa\log m}{m}) \qquad (4.136)\\ \le &P(\frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)})^2>\frac{\kappa\log m}{m}, Z\le M\frac{\log m}{m}) + P(Z>M\frac{\log m}{m}) \\ = &E\left[P\left(\frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)})^2>\frac{\kappa\log m}{m}\middle| t_q, 1\le q\le m\right)\chi_{\{Z\le M\frac{\log m}{m}\}}\right]+\frac{\tau}{2} \end{aligned}$$

$$
\begin{aligned}
\leq & E\left[\left(\frac{\kappa\log m}{m}\right)^{-1} E\left(\frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)})^2\Big| t_q, 1\leq q\leq m\right)\chi_{\{Z\leq M\frac{\log m}{m}\}}\right]+\frac{\tau}{2}\\
\leq & E\left[\left(\frac{\kappa\log m}{m}\right)^{-1} 2\sigma^2\sum_{q=1}^{m}(t_{(q)}-t_{(q-1)})^2\chi_{\{Z\leq M\frac{\kappa\log m}{m}\}}\right]+\frac{\tau}{2}\\
\leq & E\left[\left(\frac{\kappa\log m}{m}\right)^{-1} 2\sigma^2(b-a)\max_q(t_{(q)}-t_{(q-1)})\chi_{\{Z\leq M\frac{\log m}{m}\}}\right]+\frac{\tau}{2}\\
\leq & E\left[\left(\frac{\kappa\log m}{m}\right)^{-1} 2\sigma^2(b-a)Z\chi_{\{Z\leq M\frac{\log m}{m}\}}\right]+\frac{\tau}{2}\\
\leq & \left(\frac{\kappa\log m}{m}\right)^{-1} 2\sigma^2(b-a)M\frac{\log m}{m}+\frac{\tau}{2}=\tau.
\end{aligned}
$$

Hence

$$
\frac{1}{n}\sum_{p=1}^{n}(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})^2\leq O_p(\frac{\log m}{m}). \tag{4.137}
$$

Now we deal with the first and second terms on the right hand side of the equality (4.132).

$$
\begin{aligned}
& \frac{1}{n}\sum_{p=1}^{n}\left(\int_a^b(X_p^{(m)}(s)-\bar{X}^{(m)}(s))ds\right)^2-\frac{1}{n}\sum_{p=1}^{n}\left(\int_a^b(X_p(s)-\bar{X}(s))ds\right)^2\\
=&\frac{2}{n}\sum_{p=1}^{n}\int_a^b\left((X_p^{(m)}(s)-X_p(s))-(\bar{X}^{(m)}(s)-\bar{X}(s))\right)ds\\
&\qquad\times\int_a^b(X_p(t)-\bar{X}(t))dt\\
&+\frac{1}{n}\sum_{p=1}^{n}\left[\int_a^b\left((X_p^{(m)}(s)-X_p(s))-(\bar{X}^{(m)}(s)-\bar{X}(s))\right)ds\right]^2.
\end{aligned}
$$

We use $E_{\cdot|T}$ to denote the conditional expectation given $\{t_q, 1\leq q\leq m\}$. Then

$$
\begin{aligned}
& E_{\cdot|T}\left[\int_a^b\left((X_p^{(m)}(s)-X_p(s))-(\bar{X}^{(m)}(s)-\bar{X}(s)\right)ds\right]^2\\
\leq & 2E_{\cdot|T}\left[\int_a^b(X_p^{(m)}(s)-X_p(s))ds\right]^2+2E_{\cdot|T}\left[\int_a^b(\bar{X}^{(m)}(s)-\bar{X}(s))ds\right]^2\\
\leq & 2(b-a)\left(\int_a^b E_{\cdot|T}\left[(X_p^{(m)}(s)-X_p(s))^2\right]ds+\int_a^b E_{\cdot|T}\left[(\bar{X}^{(m)}(s)-\bar{X}(s))^2\right]ds\right)
\end{aligned}
$$

$$
\begin{aligned}
=&4(b-a)\int_a^b E_{\cdot|T}\left[(X_p^{(m)}(s)-X_p(s))^2\right]ds\\
=&4(b-a)\Bigg(\int_{t_{(1)}}^{\frac{t_{(2)}-t_{(1)}}{2}} E_{\cdot|T}\left[(X_p(t_{(1)})-X_p(s))^2\right]ds\\
&+\sum_{q=2}^{m-1}\int_{\frac{t_{(q)}+t_{(q-1)}}{2}}^{\frac{t_{(q)}+t_{(q+1)}}{2}} E_{\cdot|T}\left[(X_p(t_{(q)})-X_p(s))^2\right]ds\\
&+\int_{\frac{t_{(m)}+t_{(m-1)}}{2}}^{t_{(m)}} E\left[(X_p(t_{(m)})-X_p(s))^2\right]ds\Bigg)\\
\leq&4(b-a)^2\varpi(\max_q\{t_{(q)}-t_{(q-1)}\}),
\end{aligned}
$$

where the third line follows from Cauchy-Schwarz inequality and the last line follows the definition of ϖ. By using the same argument as in the proof of (4.136), we have that in Case 2,

$$
\begin{aligned}
&\frac{1}{n}\sum_{p=1}^n\left[\int_a^b\left((X_p^{(m)}(s)-X_p(s))-(\bar{X}^{(m)}(s)-\bar{X}(s))\right)ds\right]^2\\
=&O_p(\varpi(\frac{3\log m}{cm})),
\end{aligned}
\tag{4.138}
$$

and by Cauchy-Schwarz inequality

$$
\begin{aligned}
&\frac{2}{n}\sum_{p=1}^n\int_a^b\left((X_p^{(m)}(s)-X_p(s))-(\bar{X}^{(m)}(s)-\bar{X}(s))\right)ds\int_a^b(X_p(t)-\bar{X}(t))dt\\
=&O_p(\sqrt{\varpi(\frac{3\log m}{cm})}).
\end{aligned}
$$

In Case 1, they are $O_p(\varpi(\delta_m))$ and $O_p(\sqrt{\varpi(\delta_m)})$ respectively.

For the third terms on the right hand side of the equality (4.132),

$$
\begin{aligned}
&\frac{1}{n}\sum_{p=1}^n\int_a^b(X_p^{(m)}(s)-\bar{X}^{(m)}(s))ds(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})\\
=&\frac{1}{n}\sum_{p=1}^n\int_a^b(X_p^{(m)}(s)-X_p(s)-\bar{X}^{(m)}(s)+\bar{X}(s))ds(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})\\
&+\frac{1}{n}\sum_{p=1}^n\int_a^b X_p(s)ds(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})-\frac{1}{n}\sum_{p=1}^n\int_a^b\bar{X}(s)ds(\bar{\epsilon}_{p\cdot}^{(m)}-\bar{\epsilon}_{\cdot\cdot}^{(m)})
\end{aligned}
$$

Note that the last term in the above equality is zero. By (4.137), (4.138) and Cauchy-Schwarz inequality, the first term on the right hand side of the above equality is less than

$$O_p(\sqrt{\frac{\log m}{m}\varpi(\frac{3\log m}{cm})})$$

in Case 2, and less than

$$O_p(\sqrt{\delta_m \varpi(\delta_m)})$$

in Case 1. One can see that the second term on the right hand side of the above equality is an average of i.i.d. random variables. Under Assumptions 1 and 4, because the random curves, observation times and measurement errors are independent, these i.i.d. random variables have means zero, variances less than $O_p(\sqrt{\delta_m})$ in Case 1 and less than $O_p(\sqrt{\frac{\log m}{m}})$ in Case 2, and uniformly bounded third moments. By the Berry-Esseen theorem and a similar argument as in (4.136), we have that the second term on the right hand side of the above equality is $O_p(\sqrt{\frac{\delta_m}{n}})$ in Case 1 and $O_p(\sqrt{\frac{\log m}{nm}})$ in Case 2. Now we have that $|\tilde{A}_{00} - A_{00}|$ is, in Case 1, less than

$$\begin{aligned} &O_p(\delta_m) + O_p(\varpi(\delta_m)) + O_p(\sqrt{\varpi(\delta_m)}) + O_p(\sqrt{\delta_m \varpi(\delta_m)}) + O_p(\sqrt{\frac{\delta_m}{n}}) \\ &\leq O_p(\sqrt{\varpi(\delta_m)}) + O_p(\delta_m) + O_p(\sqrt{\frac{\delta_m}{n}}) \end{aligned}$$

and in Case 2, less than

$$O_p(\sqrt{\varpi(\frac{3\log m}{cm})}) + O_p(\frac{\log m}{m}) + O_p(\sqrt{\frac{\log m}{nm}}).$$

Hence, by Lemma 3, in Case 1,

$$\begin{aligned} &|\tilde{A}_{00} - A_{00}||\beta_1(a)|\beta_2(a)| \\ &\leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\delta_m)}) + O_p(\delta_m) + O_p(\sqrt{\frac{\delta_m}{n}})\right] \|\beta_1\|_\alpha \|\beta_2\|_\alpha, \end{aligned} \tag{4.139}$$

and for Case 2,

$$
\begin{aligned}
&|\tilde{A}_{00} - A_{00}||\beta_1(a)|\beta_2(a)| \qquad (4.140)\\
&\leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\frac{3\log m}{cm})}) + O_p(\frac{\log m}{m}) + O_p(\sqrt{\frac{\log m}{nm}})\right]\|\beta_1\|_\alpha\|\beta_2\|_\alpha.
\end{aligned}
$$

By similar arguments, all the following terms

$$
\begin{aligned}
&|(\tilde{A}_{10} - A_{10})\beta_1'(a)\beta_2(a)|, \quad |(\tilde{A}_{01} - A_{01})\beta_1(a)\beta_2'(a)|,\\
&|(\tilde{A}_{11} - A_{11})\beta_1'(a)\beta_2'(a)|, \quad |\beta_1(a)\int_a^b(\tilde{\xi}_0(t) - \xi_0(t))\beta_2''(t)dt|,\\
&|\beta_2(a)\int_a^b(\tilde{\xi}_0(t) - \xi_0(t))\beta_1''(t)dt|, \quad |\beta_1'(a)\int_a^b(\tilde{\xi}_1(t) - \xi_1(t))\beta_2''(t)dt|\\
&|\beta_2'(a)\int_a^b(\tilde{\xi}_1(t) - \xi_1(t))\beta_1''(t)dt|, \quad |\int_a^b\int_a^b(\tilde{\Xi}(s,t) - \Xi(s,t))\beta_1''(s)\beta_2''(t)dsdt|,
\end{aligned}
$$

have the same bounds as those in (4.139) and (4.140). Hence, we have that in Case 1,

$$
\|\hat{R}_\alpha^{(m)} - \hat{R}_\alpha\|_\alpha \leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\delta_m)}) + O_p(\delta_m) + O_p(\sqrt{\frac{\delta_m}{n}})\right],
$$

and in Case 2,

$$
\|\hat{R}_\alpha^{(m)} - \hat{R}_\alpha\|_\alpha \leq \frac{1}{\alpha}\left[O_p(\sqrt{\varpi(\frac{3\log m}{cm})}) + O_p(\frac{\log m}{m}) + O_p(\sqrt{\frac{\log m}{nm}})\right].
$$

4.11.3 Proof of Lemma 6

Proof of Lemma 6

Since $\|\hat{\gamma}_j\| = 1$, $\|\hat{\tilde{\gamma}}_j\| = 1$ and $\|\gamma_j^{[\alpha_j]}\| = 1$ by their definitions, it follows from (4.16) that

$$
\alpha_j\left[\hat{\gamma}_j, \hat{\gamma}_j\right] \to 0, \quad \alpha_j\left[\hat{\tilde{\gamma}}_j, \hat{\tilde{\gamma}}_j\right] \to 0, \quad \alpha_j\left[\gamma_j^{[\alpha_j]}, \gamma_j^{[\alpha_j]}\right] \to 0, \qquad (4.141)
$$

for all $1 \le j \le k-1$. By the following condition in this theorem,

$$\frac{\max_{1\le k\le K}\alpha_k}{\min_{1\le k\le K}\alpha_k} = O_p(1).$$

it follows from (4.16) that for any $1 \le j \le k-1$,

$$\alpha_k\left[\hat{\gamma}_j, \hat{\gamma}_j\right] \to 0, \quad \alpha_k\left[\hat{\hat{\gamma}}_j, \hat{\hat{\gamma}}_j\right] \to 0, \quad \alpha_k\left[\gamma_j^{[\alpha_j]}, \gamma_j^{[\alpha_j]}\right] \to 0, \tag{4.142}$$

$$\|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\|_{\alpha_k} \to 0, \quad \|\gamma_j^{[\alpha_j]} - \gamma_j\| \to 0, \quad \|\gamma_j^{[\alpha_j]} - \gamma_j\|_{\alpha_k} \to 0, \tag{4.143}$$

from which we have $\|\hat{\gamma}_j\|_{\alpha_k} \to 1$, $\|\hat{\hat{\gamma}}_j\|_{\alpha_k} \to 1$ and $\|\gamma_j^{[\alpha_j]}\|_{\alpha_k} \to 1$.

Let $\hat{\mathbf{V}}_k = \{\gamma \in W_2^2([a,b]) | \langle \gamma, \hat{\gamma}_j\rangle = 0, 1 \le j \le k-1\}$ and $\hat{\hat{\mathbf{V}}}_k = \{\gamma \in W_2^2([a,b]) | \left\langle \gamma, \hat{\hat{\gamma}}_j\right\rangle = 0, 1 \le j \le k-1\}$. Note that $\hat{\mathbf{V}}_k$ and $\hat{\hat{\mathbf{V}}}_k$ are the orthogonal complements of $\hat{\gamma}_j$ and $\hat{\hat{\gamma}}_j$, $1 \le j \le k-1$, in L^2 inner product respectively. They are the closed subspaces in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_{\alpha_k})$. Let $\hat{P}_k$ and $\hat{\hat{P}}_k$ be the orthogonal projections onto $\hat{\mathbf{V}}_k$ and $\hat{\hat{\mathbf{V}}}_k$ respectively in $(W_2^2([a,b]), \langle\cdot,\cdot\rangle_{\alpha_k})$. Note that they are not the orthogonal projections in L^2 inner product. Now it can be see that $\{\hat{\lambda}_k, \hat{\gamma}_k\}$ and $\{\hat{\hat{\gamma}}_k, \hat{\hat{\gamma}}_k\}$ are the first eigenvalues and eigenfunctions of $\hat{P}_k\hat{R}_{\alpha_k}^{(m)}\hat{P}_k$ and $\hat{\hat{P}}_k\hat{R}_{\alpha_k}\hat{\hat{P}}_k$. Since $\|\hat{P}_k\|_{\alpha_k} = \|\hat{\hat{P}}_k\|_{\alpha_k} = 1$,

$$\begin{aligned}
&\|\hat{P}_k\hat{R}_{\alpha_k}^{(m)}\hat{P}_k - \hat{\hat{P}}_k\hat{R}_{\alpha_k}\hat{\hat{P}}_k\|_{\alpha_k} \\
=&\|\hat{P}_k\hat{R}_{\alpha_k}^{(m)}\hat{P}_k - \hat{P}_k\hat{R}_{\alpha_k}\hat{P}_k + \hat{P}_k\hat{R}_{\alpha_k}\hat{P}_k - \hat{P}_k\hat{R}_{\alpha_k}\hat{\hat{P}}_k + \hat{P}_k\hat{R}_{\alpha_k}\hat{\hat{P}}_k - \hat{\hat{P}}_k\hat{R}_{\alpha_k}\hat{\hat{P}}_k\|_{\alpha_k} \\
\le&\|\hat{R}_{\alpha_k}^{(m)} - \hat{R}_{\alpha_k}\|_{\alpha_k} + 2\|\hat{R}_{\alpha_k}\|_{\alpha_k}\|\hat{P}_k - \hat{\hat{P}}_k\|_{\alpha_k}.
\end{aligned} \tag{4.144}$$

By the definition of $\hat{R}_{\alpha_k}$, $\|\hat{R}_{\alpha_k}\|_{\alpha_k} \le \|\hat{\Gamma}_n\| = O_p(1)$. Now we compute $\|\hat{P}_k - \hat{\hat{P}}_k\|_{\alpha_k}$. For any $x \in W_2^2([a,b])$, we have the following two decompositions,

$$x = \left(x - \sum_{j=1}^{k-1}\langle x, \hat{\gamma}_j\rangle\hat{\gamma}_j\right) + \sum_{j=1}^{k-1}\langle x, \hat{\gamma}_j\rangle\hat{\gamma}_j = \left(x - \sum_{j=1}^{k-1}\left\langle x, \hat{\hat{\gamma}}_j\right\rangle\hat{\hat{\gamma}}_j\right) + \sum_{j=1}^{k-1}\left\langle x, \hat{\hat{\gamma}}_j\right\rangle\hat{\hat{\gamma}}_j.$$

Since $x - \sum_{j=1}^{k-1}\langle x, \hat{\gamma}_j\rangle\hat{\gamma}_j \in \mathbf{V}_k$, it is mapped to itself by $\hat{P}_k$. Similarly, $x - \sum_{j=1}^{k-1}\left\langle x, \hat{\hat{\gamma}}_j\right\rangle\hat{\hat{\gamma}}_j$ is

mapped to itself by $\hat{\hat{P}}_k$. For any $1 \le j \le k-1$, because

$$0 = \left\langle \hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right\rangle = \left\langle \hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right\rangle_{\alpha_k} - \alpha_k \left[\hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right],$$

we have

$$\begin{aligned}
&\left\langle \hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right\rangle_{\alpha_k} = \|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k}^2 = \alpha_k \left[\hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right] \\
=&\frac{\alpha_k}{2} \left([\hat{\gamma}_j, \hat{\gamma}_j] + \left[\hat{P}_k \hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right] - \left[\hat{\gamma}_j - \hat{P}_k \hat{\gamma}_j, \hat{\gamma}_j - \hat{P}_k \hat{\gamma}_j \right] \right) \\
\le&\frac{\alpha_k}{2} \left([\hat{\gamma}_j, \hat{\gamma}_j] + \left[\hat{P}_k \hat{\gamma}_j, \hat{P}_k \hat{\gamma}_j \right] \right) \le \frac{\alpha_k}{2} [\hat{\gamma}_j, \hat{\gamma}_j] + \frac{1}{2} \|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k}^2,
\end{aligned}$$

and hence $\|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k}^2 \le \alpha_k [\hat{\gamma}_j, \hat{\gamma}_j] \to 0$. Similarly, $\|\hat{\hat{P}}_k \hat{\gamma}_j\|_{\alpha_k}^2 \to 0$. Now for any $x \in W_2^2([a,b])$,

$$\begin{aligned}
\|\hat{P}_k x - \hat{\hat{P}}_k x\|_{\alpha_k} =& \| \left(x - \sum_{j=1}^{k-1} \langle x, \hat{\gamma}_j \rangle \hat{\gamma}_j \right) + \sum_{j=1}^{k-1} \langle x, \hat{\gamma}_j \rangle \hat{P}_k \hat{\gamma}_j \\
&- \left(x - \sum_{j=1}^{k-1} \left\langle x, \hat{\hat{\gamma}}_j \right\rangle \hat{\hat{\gamma}}_j \right) - \sum_{j=1}^{k-1} \left\langle x, \hat{\hat{\gamma}}_j \right\rangle \hat{\hat{P}}_k \hat{\hat{\gamma}}_j \|_{\alpha_k} \\
\le& \| \sum_{j=1}^{k-1} \langle x, \hat{\gamma}_j \rangle \hat{\gamma}_j - \sum_{j=1}^{k-1} \left\langle x, \hat{\hat{\gamma}}_j \right\rangle \hat{\hat{\gamma}}_j \|_{\alpha_k} + \sum_{j=1}^{k-1} | \langle x, \hat{\gamma}_j \rangle | \|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k} \\
&+ \sum_{j=1}^{k-1} | \left\langle x, \hat{\hat{\gamma}}_j \right\rangle | \|\hat{\hat{P}}_k \hat{\hat{\gamma}}_j\|_{\alpha_k} \\
\le& \sum_{j=1}^{k-1} \|x\| \|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\|_{\alpha_k} + \sum_{j=1}^{k-1} \|x\| \|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\| \|\hat{\hat{\gamma}}_j\|_{\alpha_k} \\
&+ \sum_{j=1}^{k-1} \|x\| \left[\|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k} + \|\hat{\hat{P}}_k \hat{\hat{\gamma}}_j\|_{\alpha_k} \right],
\end{aligned}$$

hence,

$$\|\hat{P}_k - \hat{\hat{P}}_k\|_{\alpha_k} \le \sum_{j=1}^{k-1} \left[\|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\|_{\alpha_k} + \|\hat{\gamma}_j - \hat{\hat{\gamma}}_j\| \|\hat{\hat{\gamma}}_j\|_{\alpha_k} + \|\hat{P}_k \hat{\gamma}_j\|_{\alpha_k} + \|\hat{\hat{P}}_k \hat{\hat{\gamma}}_j\|_{\alpha_k} \right],$$

Then by (4.144) and Lemma 4, $\|\hat{P}_k \hat{R}_{\alpha_k}^{(m)} \hat{P}_k - \hat{\hat{P}}_k \hat{R}_{\alpha_k} \hat{\hat{P}}_k\|_{\alpha_k} \to 0$. By a similar argument as the

proofs of (4.15) and (4.12), we have

$$|\hat{\lambda}_k - \hat{\hat{\lambda}}_k| \le \|\hat{P}_k \hat{R}^{(m)}_{\alpha_k} \hat{P}_k - \hat{\hat{P}}_k \hat{R}_{\alpha_k} \hat{\hat{P}}_k\|_{\alpha_k},$$
$$\|\hat{\gamma}_k - \hat{\hat{\gamma}}_k\|_{\alpha_k} \le \|\gamma_k^{[\alpha_k]}\|_{\alpha_k} O_p(\|\hat{P}_k \hat{R}^{(m)}_{\alpha_k} \hat{P}_k - \hat{\hat{P}}_k \hat{R}_{\alpha_k} \hat{\hat{P}}_k\|^{\frac{1}{2}}_{\alpha_k}).$$

The same arguments lead to similar inequalities for $|\hat{\hat{\lambda}}_k - \lambda_k^{[\alpha_k]}|$, $\|\hat{\hat{\gamma}}_k - \gamma_k^{[\alpha_k]}\|_{\alpha_k}$ and $|\lambda_k^{[\alpha_k]} - \lambda_k|$, $\|\gamma_k^{[\alpha_k]} - \gamma_k\|_{\alpha_k}$. Then the lemma follows.

4.11.4 Proof of Lemma 7

Proof of Lemma 7

For any $\beta_1, \beta_2 \in W_2^2([a,b])$, by the definition of S_α,

$$\begin{aligned}
&\langle \beta_1, S_\alpha \beta_2 \rangle_\alpha = \langle \beta_1, \Gamma \beta_2 \rangle \\
=&B_{00}\beta_1(a)\beta_2(a) + B_{10}\beta_1'(a)\beta_2(a) + B_{01}\beta_1(a)\beta_2'(a) + B_{11}\beta_1'(a)\beta_2'(a) \\
&+ \beta_1(a)\int_a^b \psi_0(t)\beta_2''(t)dt + \beta_2(a)\int_a^b \psi_0(t)\beta_1''(t)dt + \beta_1'(a)\int_a^b \psi_1(t)\beta_2''(t)dt \\
&+ \beta_2'(a)\int_a^b \psi_1(t)\beta_1''(t)dt + \int_a^b\int_a^b \Psi(s,t)\beta_1''(s)\beta_2''(t)dsdt,
\end{aligned}$$

where

$$\begin{aligned}
&B_{00} = \int_a^b\int_a^b \Gamma(s,t)dsdt, \quad B_{01} = B_{10} = \int_a^b\int_a^b (s-a)\Gamma(s,t)dsdt, \\
&B_{11} = \int_a^b\int_a^b (s-a)(t-a)\Gamma(s,t)dsdt, \quad \psi_0(t) = \int_a^b\int_a^b \Gamma(s,u)G(u,t)dsdu, \\
&\psi_1(t) = \int_a^b\int_a^b \Gamma(s,u)G(u,t)(s-a)dsdu, \Psi(t,s) = \int_a^b\int_a^b \Gamma(v,u)G(v,t)G(u,s)dsdu.
\end{aligned}$$

We first estimate $\left|\tilde{B}_{00} - B_{00}\right|$. Define

$$\hat{B}_{00} = \frac{1}{n'}\sum_{p=1}^{n} \chi_{[N_p>1]} \frac{1}{N_p(N_p-1)} \sum_{l\neq q:1}^{N_p} \frac{(Y_{pq} - \mu(t_{pq}))}{h(t_{pq})} \cdot \frac{(Y_{pl} - \mu(t_{pl}))}{h(t_{pl})}.$$

By a routine argument, it follows from (4.17) that

$$\left|\tilde{B}_{00} - \hat{B}_{00}\right| = O_p(\frac{1}{\sqrt{n}\eta_\mu}) + O_p(n^{-\frac{1}{2}}\eta_g^{-\frac{1}{2}-\epsilon} + \eta_g^{\frac{3}{4}-\epsilon}).$$

By Assumption 6, the following random variables are i.i.d. with mean zeros and finite variances,

$$Z_p = \chi_{[N_p>1]}\frac{1}{N_p(N_p-1)}\sum_{l\neq q:1}^{N_p}\left[\frac{(Y_{pq}-\mu(t_{pq}))}{h(t_{pq})}\cdot\frac{(Y_{pl}-\mu(t_{pl}))}{h(t_{pl})} - B_{00}\right], \quad 1 \le p \le n.$$

Then $\hat{B}_{00} - B_{00} = \frac{1}{n'}\sum_{p=1}^{n} Z_p$. Therefore,

$$E\left(\frac{n'}{n}(\hat{B}_{00} - B_{00})\right)^2 = \frac{1}{n}Var(Z_1).$$

Since $\frac{n'}{n} = \frac{1}{n}\sum_{p=1}^{n}\chi_{[N_p>1]} \to P(N_1 > 1) > 0$ a.s. by the strong law of large numbers, $\frac{n}{n'} \to \frac{1}{P(N_1>1)}$ a.s.. Hence, $\frac{n}{n'} = O_p(1)$. Now

$$|\hat{B}_{00} - B_{00}| = \frac{n}{n'}|\frac{n'}{n}(\hat{B}_{00} - B_{00})| = O_p(\frac{1}{\sqrt{n}}).$$

Hence,

$$|\tilde{B}_{00} - B_{00}| = O_p(\frac{1}{\sqrt{n}\eta_\mu}) + O_p(n^{-\frac{1}{2}}\eta_g^{-\frac{1}{2}-\epsilon} + \eta_g^{\frac{3}{4}-\epsilon}) + O_p(\frac{1}{\sqrt{n}}).$$

We can obtain the same bounds for other terms. Then by the same argument as in the proof of Lemma 4, we have

$$\|\hat{S}_\alpha - S_\alpha\|_\alpha \le \frac{1}{\alpha}\left[O_p(\frac{1}{\sqrt{n}\eta_\mu}) + O_p(n^{-\frac{1}{2}}\eta_g^{-\frac{1}{2}-\epsilon} + \eta_g^{\frac{3}{4}-\epsilon}) + O_p(\frac{1}{\sqrt{n}})\right].$$

4.11.5 Proof of Lemma 1

Proof of Lemma 1

We will use the Bernstein's inequality for bounded variables (see Lemma 2.2.9 in Van Der Vaart and Wellner [48] or page 855 of Shorack and Wellner [39]). Given $1 \le i \le K$, define i.i.d. random variables Z_j, $1 \le j \le n$, where $Z_j = 1$ if the jth sample observation belongs to the

i-th class, otherwise $Z_j = 0$. Hence, Z_j has the Binomial distribution with parameters 1 and $1/K$, and it mean and variance are $1/K$ and $(1-1/K)1/K$. Let $Y_j = Z_j - 1/K$. Then $EY_j = 0$ and $Var(Y_j) = 1/K(1-1/K)$. Since $-1 \le Y_i \le 1$, by the Bernstein's inequality for bounded variables,

$$P\left(|Y_1 + \cdots + Y_n| > x\right) \le 2\exp\left(-\frac{1}{2}\frac{x^2}{nVar(Y_1) + x/3}\right),$$

for any $x > 0$. For any constant $C > 0$, let $x = nC\sqrt{\frac{\log p}{Kn}}$. Then we have

$$\begin{aligned} &P\left(\left|\frac{n_i}{n} - \frac{1}{K}\right| > C\sqrt{\frac{\log p}{Kn}}\right) = P\left(|Y_1 + \cdots + Y_n| > nC\sqrt{\frac{\log p}{Kn}}\right) \\ \le& 2\exp\left(-\frac{1}{2}\frac{n^2C^2\frac{\log p}{Kn}}{\frac{n}{K}(1-\frac{1}{K}) + n\frac{C}{3}\sqrt{\frac{\log p}{Kn}}}\right) \le 2\exp\left(-\frac{1}{2}\frac{C^2\log p}{1 + \frac{C}{3}\sqrt{\frac{K\log p}{n}}}\right). \end{aligned} \tag{4.145}$$

By the conditions, $p \ge 2$, $\sqrt{K\log p/n} \le d_0$ and $K \le p+1$, (4.145) gives

$$\begin{aligned} &P\left(\max_{1\le i\le K}\left|\frac{n_i}{n} - \frac{1}{K}\right| > C\sqrt{\frac{\log p}{Kn}}\right) \le 2K\exp\left(-\frac{1}{2}\frac{C^2\log p}{1 + \frac{C}{3}\sqrt{\frac{K\log p}{n}}}\right) \\ \le& 2(p+1)\exp\left(-\frac{C^2\log p}{2 + 2Cd_0/3}\right) \le 2(p+1)p^{-C^2/(2+2Cd_0/3)} \le p^{3-C^2/(2+2Cd_0/3)}. \end{aligned} \tag{4.146}$$

For any $M > 0$, when $C \ge (M+3)(d_0+1)$, we have $C > 3$ and

$$(M+3)(2+2Cd_0/3) \le (M+3)(C + Cd_0) \le C(M+3)(d_0+1) \le C^2.$$

Then $C^2/(2+2Cd_0/3) - 3 \ge M$ and by (4.146), we have

$$P\left(\max_{1\le i\le K}\left|\frac{n_i}{n} - \frac{1}{K}\right| > C\sqrt{\frac{\log p}{Kn}}\right) \le p^{3-C^2/(2+2Cd_0/3)} \le p^{-M}.$$

4.11.6 Proof of Lemma 2

Proof of Lemma 2

Given $1 \le k \le K-1$, let $t_i = \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\alpha}_k$ for $i = 1, \ldots, K$. Then we have $t_K = -\sum_{i=1}^{K-1} t_i$ because

$\sum_{i=1}^{K} t_i = (\sum_{i=1}^{K} \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\alpha}_k = 0$ by (3.2). By (3.22),

$$\begin{aligned}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k = \frac{1}{\lambda_k(\boldsymbol{\Xi})}\mathbf{B}\boldsymbol{\alpha}_k = \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\sum_{i=1}^{K}\boldsymbol{\mu}_i(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\alpha}_k) = \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\left[\sum_{i=1}^{K-1} t_i\boldsymbol{\mu}_i + t_K\boldsymbol{\mu}_K\right] \\ = \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\left[\sum_{i=1}^{K-1} t_i\boldsymbol{\mu}_i - \sum_{i=1}^{K-1} t_i\boldsymbol{\mu}_K\right] = \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\sum_{i=1}^{K-1} t_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_K) \qquad (4.147)\end{aligned}$$

It follows from (4.147) that

$$\begin{aligned}\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 \le \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\sum_{i=1}^{K-1}|t_i|\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_K\|_1 \le \frac{1}{\lambda_k(\boldsymbol{\Xi})K}(\sum_{i=1}^{K}|t_i|)\left(\max_{1\le i\ne j\le K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right) \\ \le \frac{1}{\lambda_k(\boldsymbol{\Xi})K}\sqrt{K\sum_{i=1}^{K} t_i^2}\left(\max_{1\le i\ne j\le K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right) = \frac{1}{\lambda_k(\boldsymbol{\Xi})}\sqrt{\frac{\sum_{i=1}^{K} t_i^2}{K}}\left(\max_{1\le i\ne j\le K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right) \\ = \frac{1}{\sqrt{\lambda_k(\boldsymbol{\Xi})}}\left(\max_{1\le i\ne j\le K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right), \qquad (4.148)\end{aligned}$$

where the last equality is due to $\sum_{i=1}^{K} t_i^2/K = \sum_{i=1}^{K}(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\alpha}_k)^2/K = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k = \lambda_k(\boldsymbol{\Xi})$. By Condition 2 (c), $\lambda_k(\boldsymbol{\Xi}) \ge \lambda_{K-1}(\boldsymbol{\Xi}) \ge c_3^{-1}\lambda_1(\boldsymbol{\Xi})$ which together with (4.148) give

$$\max_{1\le k\le K-1}\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 \le \frac{\sqrt{c_3}}{\sqrt{\lambda_1(\boldsymbol{\Xi})}}\left(\max_{1\le i\ne j\le K}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1\right).$$

On the other hand, given $1 \le i \ne j \le K$, by (4.152) in the proof of Lemma 8,

$$\boldsymbol{\mu}_j - \boldsymbol{\mu}_i = \mathbf{B}\sum_{k=1}^{K-1} s_k\boldsymbol{\alpha}_k = \sum_{k=1}^{K-1} s_k\lambda_k(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_k, \qquad (4.149)$$

where s_k's are real numbers. Multiplying $\boldsymbol{\alpha}_k^{\mathrm{T}}$ on both sides of (4.149), we can obtain $\boldsymbol{\alpha}_k^{\mathrm{T}}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) = s_k\lambda_k(\boldsymbol{\Xi})$. Note that by Lemma 9 and Condition 1 (b),

$$\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|_2^2 \le c_0(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \le 2c_0\lambda_{max}(\boldsymbol{\Delta}) = 2c_0\lambda_1(\boldsymbol{\Delta}) = 2c_0K\lambda_1(\boldsymbol{\Xi}), \quad (4.150)$$

and $\|\boldsymbol{\alpha}_k\|_2^2 \le c_0 \boldsymbol{\alpha}_k^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\alpha}_k = c_0$. By (4.149) and (4.150),

$$\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|_1 \le \sum_{k=1}^{K-1} |s_k \lambda_k(\boldsymbol{\Xi})| \|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 = \sum_{k=1}^{K-1} |\boldsymbol{\alpha}_k^{\mathrm{T}}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)| \|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1$$
$$\le \sum_{k=1}^{K-1} \|\boldsymbol{\alpha}_k\|_2 \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|_2 \|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 \le (K-1)\sqrt{c_0}\sqrt{2c_0 K \lambda_1(\boldsymbol{\Xi})} \left(\max_{1 \le k \le K-1} \|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 \right). \quad (4.151)$$

Therefore,

$$\frac{1}{(K-1)c_0\sqrt{2K\lambda_1(\boldsymbol{\Xi})}} \left(\max_{1 \le i \ne j \le K} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_1 \right) \le \max_{1 \le k \le K-1} \|\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\|_1 .$$

4.11.7 Proof of Lemma 8

Proof of Lemma 8

By the definition (3.3) of $\mathbf{B}$, for any $1 \le i \ne j \le K-1$,

$$(\mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_i)^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_j) = \boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_j = K\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_j = K\lambda_j(\boldsymbol{\Xi})\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\alpha}_j = 0,$$

and by (3.2), $\mathbf{1}_K^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_i = 0$ for any $1 \le i \le K-1$. Therefore, $\{\mathbf{1}_K, \mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_1, \cdots, \mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_{K-1}\}$ forms an orthogonal basis of $\mathbb{R}^K$. For any $1 \le i \ne j \le K$, let $\mathbf{v}_{ij}$ be the vector in $\mathbb{R}^K$ with all coordinates equal to zero except the ith and the j-th coordinates which are equal to -1 and 1, respectively. Let

$$\mathbf{v}_{ij} = a\mathbf{1}_K + \sum_{k=1}^{K-1} b_k \mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_k,$$

be the orthogonal expansion of $\mathbf{v}_{ij}$, where and a and b_k are coefficients. Since $\mathbf{v}_{ij}$ is orthogonal to $\mathbf{1}_K$, we have $a = 0$. Now

$$\boldsymbol{\mu}_j - \boldsymbol{\mu}_i = \mathbf{U}\mathbf{v}_{ij} = \mathbf{U}\sum_{k=1}^{K-1} b_k \mathbf{U}^{\mathrm{T}}\boldsymbol{\alpha}_k = \mathbf{B}\mathbf{z}, \quad (4.152)$$

where $\mathbf{z} = K\sum_{k=1}^{K-1} b_k\boldsymbol{\alpha}_k$ is a linear combination of $\boldsymbol{\alpha}_k$, $1 \le k \le K-1$. By the eigen-decomposition, $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Xi} = \sum_{k=1}^{K-1}\lambda_k(\boldsymbol{\Xi})\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathrm{T}}$. Hence,

$$\mathbf{B} = \sum_{k=1}^{K-1}\lambda_k(\boldsymbol{\Xi})\boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{1/2} = \sum_{k=1}^{K-1}\lambda_k(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}.$$

Because $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B} = \lambda_k(\boldsymbol{\Xi})\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}$ and

$$\begin{aligned}\mathbf{D}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) &= \sum_{k=1}^{K-1}\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\mathbf{z} = \sum_{k=1}^{K-1}\lambda_k(\boldsymbol{\Xi})\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{z} = \boldsymbol{\Sigma}^{-1}\sum_{k=1}^{K-1}\lambda_k(\boldsymbol{\Xi})\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{z} \\ &= \boldsymbol{\Sigma}^{-1}\mathbf{B}\mathbf{z} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i). \end{aligned} \tag{4.153}$$

Hence, the lemma is proved.

4.11.8 Proof of Lemma 9

Proof of Lemma 9

Let $\Phi = \boldsymbol{\Sigma}^{-1/2}\mathbf{U}$. Then by the definitions the definition (3.3) and (3.21), we have

$$\boldsymbol{\Delta} = \Phi^{\mathrm{T}}\Phi, \quad \boldsymbol{\Xi} = \frac{1}{K}\Phi\Phi^{\mathrm{T}}.$$

By (3.2), $\mathbf{U}\mathbf{1}_K = 0$, where $\mathbf{1}_K = (1, 1, \cdots, 1)^{\mathrm{T}}$. Hence $\boldsymbol{\Delta}\mathbf{1}_K = 0$. Since $\boldsymbol{\Delta}$ is $K \times K$, the rank of $\boldsymbol{\Delta}$ is at most $K-1$ and it has at most $K-1$ nonzero eigenvalues. For any $1 \le i \le K-1$, since $\boldsymbol{\gamma}_i$ is the i-th eigenvector of $\boldsymbol{\Xi}$ with the eigenvalue $\lambda_i(\boldsymbol{\Xi})$, we have

$$\boldsymbol{\Delta}\Phi^{\mathrm{T}}\boldsymbol{\gamma}_i = \Phi^{\mathrm{T}}\Phi\Phi^{\mathrm{T}}\boldsymbol{\gamma}_i = K\Phi^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\gamma}_i = K\Phi\lambda_i(\boldsymbol{\Xi})\boldsymbol{\gamma}_i = K\lambda_i(\boldsymbol{\Xi})\Phi\boldsymbol{\gamma}_i. \tag{4.154}$$

Therefore, $\Phi^{\mathrm{T}}\boldsymbol{\gamma}_i$ is the eigenvector of $\boldsymbol{\Delta}$ with the eigenvalue $K\lambda_i(\boldsymbol{\Xi})$, $1 \le i \le K-1$. Hence, all the nonzero eigenvalues of $\boldsymbol{\Delta}$ are

$$\lambda_{K-1}(\boldsymbol{\Delta}) = K\lambda_{K-1}(\boldsymbol{\Xi}) \le \cdots \le \lambda_2(\boldsymbol{\Delta}) = K\lambda_2(\boldsymbol{\Xi}) \le \lambda_1(\boldsymbol{\Delta}) = K\lambda_1(\boldsymbol{\Xi}).$$

To prove the inequalities in the lemma, we define $\mathbf{v}_{ij}$ to be the K-vector with all coordinates are equal to zeros except the ith and jth coordinates which are equal to 1 and -1, respectively, where $1 \leq i \neq j \leq K$. Then

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \mathbf{v}_{ij}^{\mathrm{T}} \mathbf{U}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{U} \mathbf{v}_{ij} = \mathbf{v}_{ij}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{v}_{ij} \leq \|\mathbf{v}_{ij}\|_2^2 \lambda_{max}(\boldsymbol{\Delta}) = 2\lambda_{max}(\boldsymbol{\Delta}),$$

and $\mathbf{v}_{ij}^{\mathrm{T}} \mathbf{1}_K = 0$. Since $\mathbf{1}_K$ is the eigenvector of $\boldsymbol{\Delta}$ with eigenvalue zero, all the eigenvalues of $\boldsymbol{\Delta} + \lambda_{min}^{+}(\boldsymbol{\Delta}) \mathbf{1}_K \mathbf{1}_K^{\mathrm{T}} / K$ are not less than $\lambda_{min}^{+}(\boldsymbol{\Delta})$, where $\lambda_{min}^{+}(\boldsymbol{\Delta}) = \lambda_{K-1}(\boldsymbol{\Delta}) = K\lambda_{K-1}(\boldsymbol{\Xi})$. Hence,

$$\begin{aligned}
&(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \mathbf{v}_{ij}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{v}_{ij} = \mathbf{v}_{ij}^{\mathrm{T}} \left[\boldsymbol{\Delta} + \lambda_{min}^{+}(\boldsymbol{\Delta}) \mathbf{1}_K \mathbf{1}_K^{\mathrm{T}} / K \right] \mathbf{v}_{ij} \\
\geq &\|\mathbf{v}_{ij}\|_2^2 \lambda_{min}^{+}(\boldsymbol{\Delta}) = 2\lambda_{min}^{+}(\boldsymbol{\Delta}) = 2\lambda_{K-1}(\boldsymbol{\Delta}) = 2K\lambda_{K-1}(\boldsymbol{\Xi}) \geq 2Kc_1,
\end{aligned}$$

where the last inequality is due to Condition 2 (a).

4.11.9 Proof of Lemma 10

Proof of Lemma 10

We only consider the case that $(n_1, n_2, \cdots, n_K)$ follows a multinomial distribution. For the nonrandom case, a similar argument can prove the lemma. Let $\bar{\mathbf{x}}_j^k$ denote the k-th coordinate of the j-th sample class mean $\bar{\mathbf{x}}_j$ and σ_{kk} is the k-th diagonal element of $\boldsymbol{\Sigma}$. Since $\sqrt{n_i}(\bar{\mathbf{x}}_j^k - \boldsymbol{\mu}_j^k)/\sqrt{\sigma_{kk}}$ has a standard normal distribution, for any $C_1 > 0$,

$$\begin{aligned}
&P\left(|\bar{\mathbf{x}}_j^k - \boldsymbol{\mu}_j^k| \geq C_1 \sqrt{\frac{\sigma_{kk} \log p}{n_j}} \right) \\
=&1 - \Phi(C_1 \sqrt{\log p}) \leq \frac{\phi(C_1 \sqrt{\log p})}{C_1 \sqrt{\log p}} \\
=&\frac{1}{\sqrt{2\pi \log p} C_1 p^{C_1^2/2}},
\end{aligned} \tag{4.155}$$

where Φ and ϕ are the cumulative and density functions of the standard normal distribution

and we use the inequality $1-\Phi(x) \leq \phi(x)/x$ for any $x>0$ (see page 850 in Shorack and Wellner [39]). For any $M'>0$, let $C'=2(M'+3)$. Since $K\log p/n \to 0$, $\sqrt{K\log p/n} \leq \min\{1, 1/(2C')\}$ for all n large enough. By Lemma 1,

$$P\left(\max_{1\leq i\leq K}\left|\frac{n_i}{n}-\frac{1}{K}\right| \leq C'\sqrt{\frac{\log p}{Kn}}\right) \geq 1-p^{-M'}. \tag{4.156}$$

Since $C'\sqrt{\log p/(Kn)} = (C'\sqrt{K\log p/n})/K \leq 1/(2K)$, the inequality in the parenthesis in (4.156) implies $\min_{1\leq i\leq K} n_i \geq n/(2K)$. Hence, we have $P(\min_{1\leq i\leq K} n_i \geq n/(2K)) \geq 1-p^{-M'}$, which together with (4.155) and the inequality $|\sigma_{kk}| \leq \lambda_{max}(\mathbf{\Sigma}) \leq c_0$ (see Condition 1 (b)) leads to

$$\begin{aligned}
&P\left(|\bar{\mathbf{x}}_j^k - \boldsymbol{\mu}_j^k| \geq C_1\sqrt{\frac{2c_0K\log p}{n}}\right)\\
\leq& P\left(|\bar{\mathbf{x}}_j^k - \boldsymbol{\mu}_j^k| \geq C_1\sqrt{\frac{\sigma_{kk}\log p}{n_j}}, \quad \min_{1\leq i\leq K} n_i \geq n/(2K)\right)\\
&+P(\min_{1\leq i\leq K} n_i < n/(2K))\\
\leq& \frac{1}{\sqrt{2\pi\log p}C_1p^{C_1^2/2}} + p^{-M'}.
\end{aligned}$$

Hence, it follows from the above inequality that

$$\begin{aligned}
&P\left(\max_{1\leq j\leq K}\|\bar{\mathbf{x}}_j - \boldsymbol{\mu}_j\|_\infty > C_1\sqrt{\frac{c_0K\log p}{n}}\right)\\
\leq& \sum_{1\leq j\leq K}\sum_{1\leq k\leq p} P\left(|\bar{\mathbf{x}}_j^k - \boldsymbol{\mu}_j^k| \geq C_1\sqrt{\frac{c_0K\log p}{n}}\right)\\
\leq& \frac{Kp}{\sqrt{2\pi\log p}C_1p^{C_1^2/2}} + Kpp^{-M'}.
\end{aligned}$$

Since $K \leq p+1 \leq p^2$, for any $M>0$, we choose C_1 large enough such that the first term on the right hand side of the above inequality is less than $p^{-M}/2$. Let $M'=M+4$, then the second term on the right hand side of the above inequality is less than $p^{-M-1} \leq p^M/2$. Hence, (4.19) is true for $C \geq C_1\sqrt{c_0}$ and all n large enough.

4.11.10 Proof of Lemma 11

Proof of Lemma 11

In this proof, we only consider the element in Ω_n. Since $\widehat{\boldsymbol{\xi}}_1$ is the solution to (3.12) with $j = 1$, we have

$$\|\widehat{\boldsymbol{\xi}}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1\|_2^2 + \kappa_n\|\widehat{\boldsymbol{\xi}}_1\|_1 \leq \|\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1\|_2^2 + \kappa_n\|\mathbf{B}\boldsymbol{\alpha}_1\|_1 \tag{4.157}$$

where

$$\|\widehat{\boldsymbol{\xi}}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1\|_2^2 = \|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2^2 + 2(\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1)^T(\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1) + \|\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1\|_2^2. \tag{4.158}$$

It follows from (4.157) and (4.158) that

$$\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2^2 + 2(\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1)^T(\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1) + \kappa_n\|\widehat{\boldsymbol{\xi}}_1\|_1 \leq \kappa_n\|\mathbf{B}\boldsymbol{\alpha}_1\|_1 \tag{4.159}$$

By Theorem 3.3.4, the definition (3.27) of Ω_n, the definition (3.33) of Λ_p and the fact that $\|\mathbf{B}\| \leq \lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\Sigma}\| \leq \lambda_1(\boldsymbol{\Xi})c_0$,

$$\begin{aligned}
&2(\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1)^T(\mathbf{B}\boldsymbol{\alpha}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1)\\
=&2(\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1)^T(\mathbf{B}\boldsymbol{\alpha}_1 - \mathbf{B}\widehat{\boldsymbol{\alpha}}_1) + 2(\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1)^T(\mathbf{B}\widehat{\boldsymbol{\alpha}}_1 - \widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_1)\\
\geq& -2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2\|\mathbf{B}\boldsymbol{\alpha}_1 - \mathbf{B}\widehat{\boldsymbol{\alpha}}_1\|_2 - 2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_1\|\mathbf{B} - \widehat{\mathbf{B}}\|_\infty\|\widehat{\boldsymbol{\alpha}}_1\|_1\\
\geq& -2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2\|\mathbf{B}\boldsymbol{\alpha}_1 - \mathbf{B}\widehat{\boldsymbol{\alpha}}_1\|_2 - 2(\|\widehat{\boldsymbol{\xi}}_1\|_1 + \|\mathbf{B}\boldsymbol{\alpha}_1\|_1)\|\mathbf{B} - \widehat{\mathbf{B}}\|_\infty\|\widehat{\boldsymbol{\alpha}}_1\|_1\\
\geq& -2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2\|\mathbf{B}\boldsymbol{\alpha}_1 - \mathbf{B}\widehat{\boldsymbol{\alpha}}_1\|_2 - 2(\|\widehat{\boldsymbol{\xi}}_1\|_1 + \lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_1\|_1)\|\mathbf{B} - \widehat{\mathbf{B}}\|_\infty\sqrt{6\|\boldsymbol{\alpha}_1\|_1^2/\lambda_0}\\
\geq& -2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2\|\mathbf{B}\|\|\boldsymbol{\alpha}_1 - \widehat{\boldsymbol{\alpha}}_1\|_2 - 2(\|\widehat{\boldsymbol{\xi}}_1\|_1 + \lambda_1(\boldsymbol{\Xi})\Lambda_p)(\tau_n/C_2)(\sqrt{6\Lambda_p^2/\lambda_0})\\
\geq& -2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2\lambda_1(\boldsymbol{\Xi})c_0\sqrt{C_5c_0}\sqrt{\Lambda_p^2 s_n} - 2(C/C_2)\sqrt{6/\lambda_0}\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n\\
&-2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n\|\widehat{\boldsymbol{\xi}}_1\|_1
\end{aligned}$$

which together with (4.159) lead to

$$\begin{aligned}
&\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2^2 - 2\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2 \lambda_1(\boldsymbol{\Xi}) c_0 \sqrt{C_5 c_0}\sqrt{\Lambda_p^2 s_n} - 2(C/C_2)\sqrt{6/\lambda_0}\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n \\
&- 2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n \|\widehat{\boldsymbol{\xi}}_1\|_1 + \kappa_n \|\widehat{\boldsymbol{\xi}}_1\|_1 \leq \kappa_n \|\mathbf{B}\boldsymbol{\alpha}_1\|_1 = \kappa_n \lambda_1(\boldsymbol{\Xi})\|\boldsymbol{\Sigma}\boldsymbol{\alpha}_1\|_1 \\
\leq &\kappa_n \lambda_1(\boldsymbol{\Xi})\Lambda_n = \widetilde{C}\lambda_1(\boldsymbol{\Xi})^2 \Lambda_p^2 s_n. \qquad (4.160)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\left(\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2 - \lambda_1(\boldsymbol{\Xi}) c_0 \sqrt{C_5 c_0}\sqrt{\Lambda_p^2 s_n}\right)^2 + \left(\kappa_n - 2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n\right)\|\widehat{\boldsymbol{\xi}}_1\|_1 \\
\leq &\left(\widetilde{C}\lambda_1(\boldsymbol{\Xi})^2 + 2(C/C_2)\sqrt{6/\lambda_0}\lambda_1(\boldsymbol{\Xi}) + \lambda_1(\boldsymbol{\Xi})^2(C_5 c_0^3)\right)\Lambda_p^2 s_n \\
\leq &\left(\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5 c_0^3)\right)\lambda_1(\boldsymbol{\Xi})^2 \Lambda_p^2 s_n \qquad (4.161)
\end{aligned}$$

where the last inequality is due to $\lambda_1(\boldsymbol{\Xi}) \geq c_1$ by Condition 2 (a). Then it follows from (4.161) that

$$\begin{aligned}
&\left(\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2 - \lambda_1(\boldsymbol{\Xi}) c_0 \sqrt{C_5 c_0}\sqrt{\Lambda_p^2 s_n}\right)^2 \\
\leq &\left(\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5 c_0^3)\right)\lambda_1(\boldsymbol{\Xi})^2 \Lambda_p^2 s_n
\end{aligned}$$

which implies that

$$\|\widehat{\boldsymbol{\xi}}_1 - \mathbf{B}\boldsymbol{\alpha}_1\|_2 \leq \left(\sqrt{\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5 c_0^3)} + c_0\sqrt{C_5 c_0}\right)\lambda_1(\boldsymbol{\Xi})\sqrt{\Lambda_p^2 s_n}. \qquad (4.162)$$

It also follows from (4.161) that

$$\left(\kappa_n - 2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n\right)\|\widehat{\boldsymbol{\xi}}_1\|_1 \leq \left(\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5 c_0^3)\right)\lambda_1(\boldsymbol{\Xi})^2 \Lambda_p^2 s_n. \qquad (4.163)$$

We take $\widetilde{C} > 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1}$. Then

$$\begin{aligned}
&\kappa_n - 2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n = \widetilde{C}\lambda_1(\boldsymbol{\Xi})\Lambda_p s_n - 2(C/C_2)\sqrt{6/\lambda_0}\Lambda_p s_n \\
&\geq \widetilde{C}\lambda_1(\boldsymbol{\Xi})\Lambda_p s_n - 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1}\lambda_1(\boldsymbol{\Xi})\Lambda_p s_n = \left(\widetilde{C} - 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1}\right)\lambda_1(\boldsymbol{\Xi})\Lambda_p s_n
\end{aligned}$$

which together with (4.163) lead to

$$\|\widehat{\boldsymbol{\xi}}_1\|_1 \leq \left[\left(\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5c_0^3)\right) / \left(\widetilde{C} - 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1}\right)\right]\lambda_1(\boldsymbol{\Xi})\Lambda_p. \quad (4.164)$$

Therefore, the lemma follows from (4.162) and (4.164) with

$$\begin{aligned}
C_6 &= \left(\sqrt{\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5c_0^3)} + c_0\sqrt{C_5c_0}\right), \\
C_7 &= \left(\widetilde{C} + 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1} + (C_5c_0^3)\right) / \left(\widetilde{C} - 2(C/C_2)\sqrt{6/\lambda_0}c_1^{-1}\right).
\end{aligned}$$

4.11.11 Proof of Lemma 12

Proof of Lemma 12

We only consider elements in the event Ω_n. Because $\boldsymbol{\gamma}_k$ is the k-th eigenvector of $\boldsymbol{\Xi}$, it is the solution to

$$\max_{\boldsymbol{\gamma}\in\mathbf{V}_{k-1}^{\perp}} \frac{\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\gamma}}{\|\boldsymbol{\gamma}\|_2^2}.$$

Since the projection $(\mathbf{I} - \mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k \in \mathbf{V}_{k-1}^{\perp}$ and $\|(\mathbf{I} - \mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k\|_2 \leq \|\widehat{\boldsymbol{\gamma}}_k\|_2$, we have

$$\frac{\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{k-1})\boldsymbol{\Xi}(\mathbf{I} - \mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k}{\|\widehat{\boldsymbol{\gamma}}_k\|_2^2} \leq \frac{\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{k-1})\boldsymbol{\Xi}(\mathbf{I} - \mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k}{\|(\mathbf{I} - \mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k\|_2^2} \leq \frac{\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\Xi}\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^{\mathrm{T}}\boldsymbol{\gamma}_k} = \lambda_k(\boldsymbol{\Xi}). \quad (4.165)$$

It can be seen that $\mathbf{P}_{k-1} = \sum_{i=1}^{k-1} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{\mathrm{T}}$, hence

$$\boldsymbol{\Xi}\mathbf{P}_{k-1} = \sum_{i=1}^{k-1} \boldsymbol{\Xi}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^{\mathrm{T}} = \sum_{i=1}^{k-1} \lambda_i(\boldsymbol{\Xi})\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^{\mathrm{T}}, \qquad \mathbf{P}_{k-1}\boldsymbol{\Xi}\mathbf{P}_{k-1} = \boldsymbol{\Xi}\mathbf{P}_{k-1}. \tag{4.166}$$

By (4.166),

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{k-1})\boldsymbol{\Xi}(\mathbf{I}-\mathbf{P}_{k-1})\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - 2\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\mathbf{P}_{k-1}\boldsymbol{\Xi}\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k \\
=&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - 2\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k \\
=&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - \sum_{i=1}^{k-1}\lambda_i(\boldsymbol{\Xi})(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2,
\end{aligned}$$

which combined with (4.165) lead to

$$\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - \sum_{i=1}^{k-1}\lambda_i(\boldsymbol{\Xi})(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 \le \lambda_k(\boldsymbol{\Xi})\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 = \lambda_k(\boldsymbol{\Xi})\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k.$$

Then we have

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k \le \lambda_k(\boldsymbol{\Xi})\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k + \sum_{i=1}^{k-1}\lambda_i(\boldsymbol{\Xi})(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 \le \lambda_k(\boldsymbol{\Xi})\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k + \lambda_1(\boldsymbol{\Xi})\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 \\
\le&\lambda_k(\boldsymbol{\Xi})\left[\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2\right],
\end{aligned} \tag{4.167}$$

where the last inequality is due to Condition 2(b). Because $\widehat{\boldsymbol{\alpha}}_k$ is the solution to

$$\max_{\boldsymbol{\alpha}\in\widehat{\mathbf{W}}_{k-1}^{\perp}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau_n\|\boldsymbol{\alpha}\|_{\lambda_n}^2}, \tag{4.168}$$

and noting that $(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k \in \widehat{\mathbf{W}}_{k-1}^{\perp}$, $\widehat{\boldsymbol{\alpha}}_k = \boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k$ and $\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k + \tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_{\lambda_n}^2 = 1$, we have

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{B}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_k = \frac{\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\mathbf{B}}\widehat{\boldsymbol{\alpha}}_k}{\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k + \tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_{\lambda_n}^2} \\
\ge&\frac{\boldsymbol{\alpha}_k^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\widehat{\mathbf{B}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k}{\boldsymbol{\alpha}_k^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\widehat{\boldsymbol{\Sigma}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k + \tau_n\|(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k\|_{\lambda_n}^2},
\end{aligned} \tag{4.169}$$

By (4.167) and the definition of Ω_n, the left hand side of (4.169)

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{B}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k \le \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k + \|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2\\
\le&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k + \frac{1}{C_2}\tau_n\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2 \le \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k + \frac{\lambda_k(\boldsymbol{\Xi})}{c_1}\frac{1}{C_2}\tau_n\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2\\
\le&\lambda_k(\boldsymbol{\Xi})\left[\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 + \frac{c_1^{-1}}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2\right] \qquad (4.170)
\end{aligned}$$

where we use $\lambda_k(\boldsymbol{\Xi}) \ge c_1$. Now

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k - \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k \qquad (4.171)\\
=&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k \le \|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\|_\infty\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2 + \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k\\
\le&\frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 + \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 = 1-\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_{\lambda_n} + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2\\
\le&1-\lambda_n\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1 + \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 = 1-(\lambda_n-1/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2.
\end{aligned}$$

Because as n is large enough, $\lambda_n - 1/C_2 \ge \lambda_n - (1+c_1^{-1})/C_2 = \lambda_n - \lambda_0/2 > \lambda_0/2$, (4.171) gives $\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 \le 1$. On the other hand,

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k - \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\\
=&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k + \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k \ge \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k - \|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\|_\infty\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2\\
\ge&\widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}_k - \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 = 1-\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_{\lambda_n} - \frac{1}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 \ge 1-(1+1/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2,
\end{aligned}$$

which together with (4.171) lead to

$$1-(1+1/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 \le \|\widehat{\boldsymbol{\gamma}}_k\|_2^2 \le 1, \qquad (4.172)$$

as n is large enough. It follows from (4.170) and (4.171) that

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{B}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k \le \lambda_k(\boldsymbol{\Xi})\left[1-(\lambda_n-1/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 + \frac{c_1^{-1}}{C_2}\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2\right]\\
=&\lambda_k(\boldsymbol{\Xi})\left[1-(\lambda_n-(1+c_1^{-1})/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2\right]\\
=&\lambda_k(\boldsymbol{\Xi})\left[1-(\lambda_n-\lambda_0/2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2\right] \qquad (4.173)
\end{aligned}$$

where we use $(1+c_1^{-1})/C_2 = \lambda_0/2$ by the definition (3.26) of C_2. Next, we calculate the right hand side of (4.169). Let $\boldsymbol{\beta}_k = (\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k$.

$$\begin{aligned}
&\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\beta}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\mathbf{B}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\mathbf{B}\boldsymbol{\alpha}_k + \boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\mathbf{B}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\\
\ge&\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\mathbf{B}\boldsymbol{\alpha}_k = \lambda_k(\boldsymbol{\Xi})\left[\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\right]. \qquad (4.174)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\Sigma}(\mathbf{I}-\widehat{\mathbf{Q}}_{k-1})\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + \boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\\
\le&\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + \|\boldsymbol{\Sigma}\|\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 \le \boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2.
\end{aligned} \qquad (4.175)$$

Then by the definition (3.27) of Ω_n, (4.174), (4.175) and Condition 2 (a), the right hand side of (4.169) is equal to

$$\begin{aligned}
&\frac{\boldsymbol{\beta}_k^{\mathrm{T}}\widehat{\mathbf{B}}\boldsymbol{\beta}_k}{\boldsymbol{\beta}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k + \tau_n\|\boldsymbol{\beta}_k\|_{\lambda_n}^2} \ge \frac{\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\beta}_k - \|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k + \|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_\infty\|\boldsymbol{\beta}_k\|_1^2 + \tau_n\|\boldsymbol{\beta}_k\|_1^2}\\
\ge&\frac{\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{B}\boldsymbol{\beta}_k - \frac{1}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}_k + \frac{1}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2 + \tau_n\|\boldsymbol{\beta}_k\|_1^2}\\
\ge&\frac{\lambda_k(\boldsymbol{\Xi})\left[\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\right] - \lambda_k(\boldsymbol{\Xi})\frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2}. \qquad (4.176)
\end{aligned}$$

Now by (4.169), (4.173) and (4.176),

$$\begin{aligned}
&\frac{\lambda_k(\boldsymbol{\Xi})\left[\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\right] - \lambda_k(\boldsymbol{\Xi})\frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} \\
\leq&\lambda_k(\boldsymbol{\Xi})\left[1 - (\lambda_n - \lambda_0/2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2\right],
\end{aligned}$$

which, by a simple calculation, leads to

$$\begin{aligned}
&(\lambda_n - \lambda_0/2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 \\
\leq&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2 + \frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 \\
=&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+\lambda_0/2)\tau_n\|\boldsymbol{\beta}_k\|_1^2}{1 - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 \\
\leq&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+\lambda_0/2)\tau_n\|\boldsymbol{\beta}_k\|_1^2}{1 - 2c_0^{3/2}\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2} + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2, \qquad (4.177)
\end{aligned}$$

where we use

$$\|\boldsymbol{\alpha}_i\|_2^2 \leq \|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_i\|_2^2 = \|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\gamma}_k\|_2^2 = \|\boldsymbol{\Sigma}^{-1}\| \leq c_0. \qquad (4.178)$$

We will estimate terms on the right hand side of (4.177). By (4.76), we have

$$\widehat{\mathbf{P}}_{k-1}\widehat{\boldsymbol{\gamma}}_k = \mathbf{0}, \quad \mathbf{Q}_{k-1}\boldsymbol{\alpha}_k = \mathbf{0}. \qquad (4.179)$$

Since we assume that (4.77) is true for all $1 \leq i \leq k-1$ and all n large enough, by (4.172), (4.178) and (4.179), we have

$$\begin{aligned}
&\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2 = \|\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k\|_2^2 = \|\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k - \widehat{\mathbf{P}}_{k-1}\widehat{\boldsymbol{\gamma}}_k\|_2^2 \leq \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\|^2\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 \\
\leq&\|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\|^2 \leq C_{k-1,3}\Lambda_p^2 s_n \\
\text{and } &\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 = \|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k - \mathbf{Q}_{k-1}\boldsymbol{\alpha}_k\|_2^2 \leq \|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|^2\|\boldsymbol{\alpha}_k\|_2^2, \\
&\leq c_0\|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|^2 \leq c_0 C_{k-1,4}\Lambda_p^2 s_n. \qquad (4.180)
\end{aligned}$$

Next, we estimate $\|\boldsymbol{\beta}_k\|_1$. Let $\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k = \sum_{i=1}^{k-1} t_i\widehat{\boldsymbol{\xi}}_i$, where $\mathbf{t} = (t_1, \cdots, t_{k-1})$ is the coefficient vector. By (4.77),

$$\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_1 \le \sum_{i=1}^{k-1} |t_i| \max_{1\le i\le k-1} \|\widehat{\boldsymbol{\xi}}_i\|_1 \le \|\mathbf{t}\|_1 \left(\max_{1\le i\le k-1} C_{i,5}\right) \lambda_1(\boldsymbol{\Xi})\Lambda_p. \tag{4.181}$$

To find an upper bound for $\|\mathbf{t}\|_1$, we multiply $\boldsymbol{\alpha}_j^{\mathrm{T}}$, $1 \le j \le k-1$, on both sides of $\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k = \sum_{i=1}^{k-1} t_i\widehat{\boldsymbol{\xi}}_i = \sum_{i=1}^{k-1} t_i\boldsymbol{\xi}_i - \sum_{i=1}^{k-1} t_i(\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i)$, then by (4.77) and (4.178), we have

$$\begin{aligned}
|\boldsymbol{\alpha}_j^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k| &= \left|\sum_{i=1}^{k-1} t_i\boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{\xi}_i - \sum_{i=1}^{k-1} t_i\boldsymbol{\alpha}_j^{\mathrm{T}}(\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i)\right| = \left|\lambda_j(\boldsymbol{\Xi})t_j - \sum_{i=1}^{k-1} t_i\boldsymbol{\alpha}_j^{\mathrm{T}}(\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i)\right| \\
&\ge \lambda_j(\boldsymbol{\Xi})|t_j| - \sum_{i=1}^{k-1} |t_i|\|\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i\|_2\|\boldsymbol{\alpha}_j\|_2 \ge \lambda_j(\boldsymbol{\Xi})|t_j| - c_0^{1/2}\sum_{i=1}^{k-1} |t_i|\|\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i\|_2 \\
&\ge \lambda_j(\boldsymbol{\Xi})|t_j| - c_0^{1/2}\sum_{i=1}^{k-1} |t_i| \left(\max_{1\le i\le k-1} \|\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i\|_2\right) \\
&\ge \lambda_j(\boldsymbol{\Xi})|t_j| - c_0^{1/2}\|\mathbf{t}\|_1 \left(\max_{1\le i\le k-1} \sqrt{C_{i,6}}\right) \lambda_1(\boldsymbol{\Xi})\sqrt{\Lambda_p^2 s_n} \\
&\ge \lambda_1(\boldsymbol{\Xi}) \left[c_3^{-1}|t_j| - c_0^{1/2}\|\mathbf{t}\|_1 \left(\max_{1\le i\le k-1} \sqrt{C_{i,6}}\right) \sqrt{\Lambda_p^2 s_n}\right],
\end{aligned} \tag{4.182}$$

where the last inequality is due to Condition 2 (c). On the other hand, by (4.179) and (4.178), $|\boldsymbol{\alpha}_j^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k| = |\boldsymbol{\alpha}_j^{\mathrm{T}}(\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1})\boldsymbol{\alpha}_k| \le \|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|\|\boldsymbol{\alpha}_j\|_2\|\boldsymbol{\alpha}_k\|_2 \le c_0\|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|$ which together with (4.182) leads to

$$\begin{aligned}
&\lambda_1(\boldsymbol{\Xi})c_3^{-1}\|\mathbf{t}\|_1 = \lambda_1(\boldsymbol{\Xi})c_3^{-1}\sum_{j=1}^{k-1} |t_j| \\
\le &\sum_{j=1}^{k-1} |\boldsymbol{\alpha}_j^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k| + (k-1)c_0^{1/2}\|\mathbf{t}\|_1 \left(\max_{1\le i\le k-1} \sqrt{C_{i,6}}\right) \lambda_1(\boldsymbol{\Xi})\sqrt{\Lambda_p^2 s_n} \\
\le &(k-1)c_0\|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\| + (k-1)c_0^{1/2}\|\mathbf{t}\|_1 \left(\max_{1\le i\le k-1} \sqrt{C_{i,6}}\right) \lambda_1(\boldsymbol{\Xi})\sqrt{\Lambda_p^2 s_n}
\end{aligned}$$

By solving the above inequality, we obtain

$$\|\mathbf{t}\|_1 \le \left[c_3^{-1} - (k-1)c_0^{1/2} \left(\max_{1\le i\le k-1} \sqrt{C_{i,6}}\right) \sqrt{\Lambda_p^2 s_n}\right]^{-1} \lambda_1(\boldsymbol{\Xi})^{-1}(k-1)c_0\|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|,$$

which together (4.181) imply $\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_1 \le O(1)\|\widehat{\mathbf{Q}}_{k-1} - \mathbf{Q}_{k-1}\|\Lambda_p = o(1)\Lambda_p$ by (4.77).

Therefore,

$$\|\boldsymbol{\beta}_k\|_1 = \|\boldsymbol{\alpha}_k - \widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_1 \le \|\boldsymbol{\alpha}_k\|_1 + \|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_1 = (1+o(1))\Lambda_p \le 2\Lambda_p, \tag{4.183}$$

as n is large enough. By (4.177), (4.180) and (4.183), and noting that $\lambda_n > \lambda_0$ as n is large enough, we have

$$\|\widehat{\boldsymbol{\alpha}}_k\|_1 \le C_{k,1}\Lambda_p, \tag{4.184}$$

for all n large enough, where $C_{k,1}$ is a constant only depending $C_{i,j}$, $1 \le i \le k-1$ and $1 \le j \le 6$, λ_0, C, C_2, $\widetilde{C}$ and the constants in Conditions 1 and 2. By (4.167), (4.172) and (4.180),

$$\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k \le \lambda_k(\boldsymbol{\Xi})\left[\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k + c_3\sum_{i=1}^{k-1}(\boldsymbol{\gamma}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k)^2\right] \le \lambda_k(\boldsymbol{\Xi})\left[1 + c_3C_{k-1,3}\Lambda_p^2 s_n\right]. \tag{4.185}$$

By (4.169) and (4.176),

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k = \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k \ge \widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\widehat{\mathbf{B}}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k - \|\widehat{\mathbf{B}}-\mathbf{B}\|_\infty\|\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\gamma}}_k\|_1^2 \\
\ge&\frac{\lambda_k(\boldsymbol{\Xi})\left[\boldsymbol{\alpha}_k\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k\right] - \lambda_k(\boldsymbol{\Xi})\frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\alpha}_k\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} - \tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2/C_2.
\end{aligned} \tag{4.186}$$

By the similar arguments as in (4.177) and (4.184), we have

$$\begin{aligned}
&\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k/\lambda_k(\boldsymbol{\Xi}) - 1 \\
\ge&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2 + \frac{c_1^{-1}}{C_2}\tau_n\|\boldsymbol{\beta}_k\|_1^2}{\boldsymbol{\alpha}_k^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} - \tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2/(C_2\lambda_k(\boldsymbol{\Xi})) \\
=&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+\lambda_0/2)\tau_n\|\boldsymbol{\beta}_k\|_1^2}{1 - 2\boldsymbol{\alpha}_k^{\mathrm{T}}\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\Sigma}\boldsymbol{\alpha}_k + c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+1/C_2)\tau_n\|\boldsymbol{\beta}_k\|_1^2} - \tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2/(C_2\lambda_k(\boldsymbol{\Xi})) \\
\ge&\frac{c_0\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2^2 + (1+\lambda_0/2)\tau_n\|\boldsymbol{\beta}_k\|_1^2}{1 - 2c_0^{3/2}\|\widehat{\mathbf{Q}}_{k-1}\boldsymbol{\alpha}_k\|_2} - CC_2^{-1}C_{k,1}^2\lambda_k(\boldsymbol{\Xi})^{-1}\Lambda_p^2 s_n
\end{aligned}$$

which together with (4.180) and (4.183) imply that as n is large enough,

$$\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k/\lambda_k(\boldsymbol{\Xi}) - 1 \geq -C_8\Lambda_p^2 s_n, \tag{4.187}$$

where C_8 is a constant independent of n and p. Combining (4.185) and (4.187), we obtain

$$|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - \lambda_k(\boldsymbol{\Xi})| \leq C_9\lambda_k(\boldsymbol{\Xi})\Lambda_p^2 s_n, \tag{4.188}$$

where $C_9 = \max(C_8, c_3 C_{k-1,3})$. Let

$$\widehat{\boldsymbol{\gamma}}_k = d_1\boldsymbol{\gamma}_1 + d_2\boldsymbol{\gamma}_2 + \cdots + d_{K-1}\boldsymbol{\gamma}_{K-1} + \widehat{c}\widehat{\boldsymbol{\beta}} \tag{4.189}$$

be the orthogonal expansion of $\widehat{\boldsymbol{\gamma}}_k$, where $\widehat{\boldsymbol{\beta}}$ is an vector orthogonal to each of $\boldsymbol{\gamma}_{K-1}, \cdots, \boldsymbol{\gamma}_1$, with $\|\widehat{\boldsymbol{\beta}}\|_2 = 1$.

$$\begin{aligned}
&|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\Xi}\widehat{\boldsymbol{\gamma}}_k - \lambda_k(\boldsymbol{\Xi})| \\
=&\left|d_1^2\lambda_1(\boldsymbol{\Xi}) + d_2^2\lambda_2(\boldsymbol{\Xi}) + \cdots + d_{K-1}^2\lambda_{K-1}(\boldsymbol{\Xi}) - \lambda_k(\boldsymbol{\Xi})\right| \\
\geq&\left|d_k^2 - 1\right|\left|\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})\sum_{i=k+1}^{K-1} d_i^2 - \lambda_1(\boldsymbol{\Xi})\sum_{i=1}^{k-1} d_i^2\right. \\
\geq&\left|d_k^2 - 1\right|\left|\lambda_k(\boldsymbol{\Xi}) - \left|d_k^2 - 1\right|\right|\lambda_{k+1}(\boldsymbol{\Xi}) - (d_k^2 - 1)\lambda_{k+1}(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})\sum_{i=k+1}^{K-1} d_i^2 - \lambda_{k+1}(\boldsymbol{\Xi})\sum_{i=1}^{k-1} d_i^2\right. \\
&- (\lambda_1(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi}))\sum_{i=1}^{k-1} d_i^2 \\
=&\left|d_k^2 - 1\right|\left|[\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})] - \lambda_{k+1}(\boldsymbol{\Xi})\left[\sum_{i=1}^{K-1} d_i^2 - 1\right] - (\lambda_1(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi}))\sum_{i=1}^{k-1} d_i^2\right. \\
\geq&\left|d_k^2 - 1\right|\left|[\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})] - \lambda_{k+1}(\boldsymbol{\Xi})\left[\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right] - \lambda_1(\boldsymbol{\Xi})\sum_{i=1}^{k-1} d_i^2\right. \\
=&\left|d_k^2 - 1\right|\left|[\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})] - \lambda_{k+1}(\boldsymbol{\Xi})\left[\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right] - \lambda_1(\boldsymbol{\Xi})\|\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k\|_2^2\right. \\
=&\left|d_k^2 - 1\right|\left|[\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})] - \lambda_{k+1}(\boldsymbol{\Xi})\left[\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right] - \lambda_1(\boldsymbol{\Xi})\|\mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k - \mathbf{P}_{k-1}\widehat{\boldsymbol{\gamma}}_k\|_2^2\right. \\
\geq&\left|d_k^2 - 1\right|\left|[\lambda_k(\boldsymbol{\Xi}) - \lambda_{k+1}(\boldsymbol{\Xi})] - \lambda_{k+1}(\boldsymbol{\Xi})\left[\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right] - \lambda_1(\boldsymbol{\Xi})\|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\|^2\right.,
\end{aligned} \tag{4.190}$$

where by (4.172), (4.184) and (4.77), as n is large enough,

$$\left|\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right| = \left|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_k - 1\right| \leq (1 + 1/C_2)\tau_n\|\widehat{\boldsymbol{\alpha}}_k\|_1^2 \leq (1 + 1/C_2)CC_{k,1}\Lambda_p^2 s_n. \tag{4.191}$$

Hence by (4.180),(4.188),(4.190), (4.191), (4.77) and Condition 2,

$$\left|d_k^2 - 1\right| \leq C_{10}\Lambda_p^2 s_n, \tag{4.192}$$

where C_{10} is a constant independent of n and p. Since $\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\gamma}_k = d_k > 0$, by the orthogonal decomposition (4.189), (4.191) and (4.192)

$$\left|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\gamma}_k - \|\boldsymbol{\gamma}_k\|_2^2\right| = \left|d_k - 1\right| \leq \left|d_k - 1\right|(d_k + 1) = \left|d_k^2 - 1\right| \leq C_{10}\Lambda_p^2 s_n. \tag{4.193}$$

and

$$\begin{aligned}\|\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\|_2^2 &= \left|\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 2\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\gamma}_k + \|\boldsymbol{\gamma}_k\|_2^2\right| \leq \left|\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - \|\boldsymbol{\gamma}_k\|_2^2\right| + \left| - 2\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\gamma}_k + 2\|\boldsymbol{\gamma}_k\|_2^2\right| \\ &= \left|\|\widehat{\boldsymbol{\gamma}}_k\|_2^2 - 1\right| + 2\left|\widehat{\boldsymbol{\gamma}}_k^{\mathrm{T}}\boldsymbol{\gamma}_k - \|\boldsymbol{\gamma}_k\|_2^2\right| \leq C_{k,2}\Lambda_p^2 s_n,\end{aligned} \tag{4.194}$$

where $C_{k,2} = (1 + 1/C_2)CC_{k,1} + 2C_{10}$. By (4.184), a similar argument as in the proof of Lemma 11 leads to

$$\|\widehat{\boldsymbol{\xi}}_k\|_1 \leq C_{k,5}\lambda_1(\boldsymbol{\Xi})\Lambda_p, \quad \|\widehat{\boldsymbol{\xi}}_k - \boldsymbol{\xi}_k\|_2^2 \leq C_{k,6}\lambda_1(\boldsymbol{\Xi})\Lambda_p^2 s_n \tag{4.195}$$

where $C_{k,5}$ and $C_{k,6}$ are constants independent of n and p. Now we estimate $\|\widehat{\mathbf{P}}_k - \mathbf{P}_k\|$ and $\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|$. Let $\widehat{\mathbf{w}}_k = (\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k / \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2$. Then it is easy to show that $\widehat{\mathbf{P}}_k = \widehat{\mathbf{w}}_k\widehat{\mathbf{w}}_k^{\mathrm{T}} + \widehat{\mathbf{P}}_{k-1}$ and $\mathbf{P}_k = \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathrm{T}} + \mathbf{P}_{k-1}$. Hence,

$$\begin{aligned}&\|\widehat{\mathbf{P}}_k - \mathbf{P}_k\| = \|\widehat{\mathbf{w}}_k\widehat{\mathbf{w}}_k^{\mathrm{T}} + \widehat{\mathbf{P}}_{k-1} - \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^{\mathrm{T}} - \mathbf{P}_{k-1}\| \leq \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\| \\ &+ 2\|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2\|\boldsymbol{\gamma}_k\|_2 + \|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2^2 = \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\| + 2\|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2 + \|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2^2,\end{aligned} \tag{4.196}$$

where

$$\begin{aligned}
\|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2 &= \left\| \frac{(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k}{\|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2} - \boldsymbol{\gamma}_k \right\|_2 \\
&\leq \left\| \frac{(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k}{\|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2} - (\mathbf{I} - \widehat{\mathbf{P}}_{k-1}) \right\|_2 + \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1}) - \boldsymbol{\gamma}_k\|_2 \\
&\leq \left| 1 - \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2 \right| + \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k\|_2. \qquad (4.197)
\end{aligned}$$

By (4.180) , (4.195) and (4.77)

$$\begin{aligned}
&\|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k\|_2 = \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})(\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k) - \widehat{\mathbf{P}}_{k-1}\boldsymbol{\gamma}_k\|_2 \\
=&\|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})(\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k) - (\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1})\boldsymbol{\gamma}_k\|_2 \leq \|\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k\|_2 + \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\| \\
=&\|\lambda_k(\boldsymbol{\Xi})^{-1}\boldsymbol{\Sigma}^{-1/2}\widehat{\boldsymbol{\xi}}_k - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}_k\|_2 + \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\| \\
\leq&\lambda_k(\boldsymbol{\Xi})^{-1}\|\boldsymbol{\Sigma}^{-1/2}\|\|\widehat{\boldsymbol{\xi}}_k - \mathbf{B}\boldsymbol{\alpha}_k\|_2 + \|\widehat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1}\| \leq \sqrt{C_{11}\Lambda_p^2 s_n}, \qquad (4.198)
\end{aligned}$$

where C_{11} is a constant independent of n and p, and we use Condition 2 (c). Hence

$$\left| \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2 - 1 \right| = \left| \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k\|_2 - \|\boldsymbol{\gamma}_k\|_2 \right| \leq \|(\mathbf{I} - \widehat{\mathbf{P}}_{k-1})\widehat{\boldsymbol{\zeta}}_k - \boldsymbol{\gamma}_k\|_2 \leq \sqrt{C_{11}\Lambda_p^2 s_n}. \qquad (4.199)$$

Therefore, by (4.197)-(4.199), $\|\widehat{\mathbf{w}}_k - \boldsymbol{\gamma}_k\|_2 = 2\sqrt{C_{11}\Lambda_p^2 s_n}$ which together with (4.196) imply that $\|\widehat{\mathbf{P}}_k - \mathbf{P}_k\| \leq \sqrt{C_{k,3}\Lambda_p^2 s_n}$, where $C_{k,3}$ is a constant independent of n and p. Let $\widehat{\mathbf{v}}_k = (\mathbf{I} - \widehat{\mathbf{Q}}_{k-1})\widehat{\boldsymbol{\xi}}_k / \|(\mathbf{I} - \widehat{\mathbf{Q}}_{k-1})\widehat{\boldsymbol{\xi}}_k\|_2$. Then it is easy to show that $\widehat{\mathbf{Q}}_k = \widehat{\mathbf{v}}_k\widehat{\mathbf{v}}_k^{\mathrm{T}} + \widehat{\mathbf{Q}}_{k-1}$ and $\mathbf{Q}_k = \boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\mathrm{T}}/\|\boldsymbol{\xi}_k\|_2^2 + \mathbf{Q}_{k-1}$. A similar argument leads to $\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\| \leq \sqrt{C_{k,4}\Lambda_p^2 s_n}$, where $C_{k,4}$ is a constant independent of n and p. We have proved the lemma.

Bibliography

[1] Kerem Altun and Billur Barshan. Human activity recognition using inertial/magnetic sensor units. In *Human Behavior Understanding*, pages 38–51. Springer, 2010.

[2] Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43 (10):3605–3620, 2010.

[3] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis, Third Edition*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471360919.

[4] K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[5] Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, page bxt075, 2013.

[6] Michal Benko, Wolfgang Härdle, and Alois Kneip. Common functional principal components. *Ann. Statist.*, 37:1–34, 2009.

[7] 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*, pages 989–1010, 2004.

[8] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

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

[10] M. A. Chrenek and et al. Analysis of the rpe sheet in the rd10 retinal degeneration model, in retinal degenerative diseases. *in press*, 2011.

[11] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersboll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[12] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a random vector function: some application to statistical inference. *J. Multivariate Anal.*, 12:136–154, 1982.

[13] S. Dudoit, J. Fridlyand, , and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[14] Neilson Dunford and Jacob T. Schwartz. *Linear Operators, Spectral Theory, Self Adjoint Operators in Hilbert Space, Part 2*. Wiley-Interscience, 1988.

[15] 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.

[16] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, 2006.

[17] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[18] Paul Glasserman. *Monte Carlo Methods in Financial Engineering. 1 edition.* Springer, 2003.

[19] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its applications in microarrays. *Biostatistics*, 8:86–100, 2007.

[20] P. Hall, H. Müller, and J. Wang. Properties of principal component methods for functional and longitudinal data. *The Annals of Statistics*, 34:1493–1517, 2006.

[21] Wolfgang Härdle and Léopold Simar. *Applied multivariate statistical analysis, Third Edition*. Springer, 2012.

[22] Jianhua Z. Huang, Haipeng Shen, and Andreas Buja. Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statististics*, 2(1): 678–695, 2008.

[23] Gareth M. James, Trevor J. Hastie, and Catherine A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(1):587–602, 2000.

[24] Y. Jiang and et al. Retinal pigment epithelium morphology distinguishes age and phenotype of the eye. *in preparation*, 2011.

[25] M. W. Kadous. Temporal classification: Extending the classification paradigm to multivariate time series. *PhD Thesis (draft), School of Computer Science and Engineering, University of New South Wales*, 2002.

[26] W. Krzanowski, P. Jonathan, W. McCarthy, , and M. Thomas. Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society*, 44:101–115, 1995.

[27] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3.Folge, Band 23. A Series of Modern Surveys in Mathematics, Springer, 2006.

[28] H. Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32:223–240, 2005.

[29] J. Peng and D. Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. in press. *Journal of Computational and Graphical Statistics*, 2009.

[30] Maurice Priestley. *Spectral Analysis and Time Series. Volumes I and II.* Academic Press, 1983.

[31] Xin Qi and Hongyu Zhao. Some theoretical properties of Silverman's method for smoothed functional principal component analysis. *Journal of Multivariate Analysis*, 102:741–767, 2011.

[32] Xin Qi, Ruiyan Luo, Raymond J. Carroll, and Hongyu Zhao. Sparse regression by projection and sparse discriminant analysis. *Journal of Computational and Graphical Statistics*, 24(2):416–438, 2015.

[33] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis. 2nd Edition.* New York: Springer, 2005.

[34] James O. Ramsay, Giles Hooker, and Spencer B. Graves. *Functional Data Analysis with R and Matlab*. Springer, New York, 2009.

[35] John A. Rice and Colin O. Wu. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57(1):253–259, 2001.

[36] Frigyes Riesz and Bela Sz.-Nagy. *Functional Analysis*. Dover Publications, 1990.

[37] J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, 39(5):1241–1265, 2011.

[38] Minggao Shi, Robert E. Weiss, and Jeremy M. G. Taylor. An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(1):151–163, 1996.

[39] Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*, volume 59. Siam, 2009.

[40] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 10:795–810, 1982.

[41] Bernard W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996.

[42] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[43] R.F. Spaide and et al. Serous detachment of the retina. *Retina*, 23(6):820–846, 2003.

[44] Joan G. Staniswalis and J. Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(1):1403–1418, 1998.

[45] Olaf Strauss. The retinal pigment epithelium in visual function. *Physiol. Rev.*, 85: 845–881, 2005.

[46] Christine Thomas-Agnan. Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13:21–32, 1996.

[47] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99:6567–6572, 2002.

[48] Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[49] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer; 4th Edition, 2002.

[50] Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[51] Hans F. Weinberger. *Variational Methods for Eigenvalue Approximation, 2nd Edition (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Society for Industrial Mathematics, 1987.

[52] D. Witten and R. Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society, Ser. B*, 73:753–772, 2011.

[53] P. Xu, G. Brock, and R. Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics and Data Analysis*, 53:1674–1687, 2009.

[54] F. Yao, H. G. Müller, and J. L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.

[55] Keith M. Zinn and Michael F. Marmor. *The Retinal Pigment Epithelium.* Harvard University Press, 1979.