Electronic Journal of Statistics
Vol. 16 (2022) 5378–5436
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2066

Varying coefficient linear discriminant analysis for dynamic data

Yajie Bao and Yuyang Liu

School of Mathematical Sciences,
Shanghai Jiao Tong University,
200240 Shanghai, China
e-mail: baoyajie2019stat@sjtu.edu.cn; d0408x@sjtu.edu.cn

Abstract: Linear discriminant analysis (LDA) is an important classification tool in statistics and machine learning. This paper investigates the varying coefficient LDA model for dynamic data, with Bayes' discriminant direction being a function of some exposure variable to address the heterogeneity. We propose a new least-square estimation method based on the B-spline approximation. The data-driven discriminant procedure is more computationally efficient than the dynamic linear programming rule [21]. We also establish the convergence rates for the corresponding estimation error bound and the excess misclassification risk. The estimation error in L_2 distance is optimal for the low-dimensional regime and is near optimal for the high-dimensional regime. Numerical experiments on synthetic data and real data both corroborate the superiority of our proposed classification method.

MSC2020 subject classifications: Primary 62H30; secondary 62G05.
Keywords and phrases: Bayes' rule, B-spline, least-square classification, group lasso.

Received November 2021.

Contents

1. Introduction

Classification is one of the most essential topics in statistics and machine learning, and widely applied in many scientific and industrial fields. Consider a pair of random variables $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^p$ is the covariate and $Y \in \{0, 1\}$ is the label. If $Y = 1$, the covariate $\boldsymbol{X}$ follows the p-dimensional multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, otherwise $\boldsymbol{X}$ is distributed as $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. We assume the prior probabilities of two classes are equal, that is $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = 1/2$. For a new random covariate $\boldsymbol{X}_{\text{new}}$, we aim to predict its unknown label Y_{new} according to some discriminant rule. If we know the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ in advance, let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, then the well-known Bayes' linear discriminant rule is given by

$$\psi(\boldsymbol{X}_{\text{new}}) = \mathbb{I}\left((\boldsymbol{X}_{\text{new}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0\right) \tag{1.1}$$

where $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is called Bayes' discriminant direction. In real data analysis, a data-driven Bayes' classification rule is given by plugging sample means $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_2$ and pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ in (1.1), which is asymptotically optimal when the dimensionality p is fixed [1].

Driven by contemporary measurement technologies, high-dimensional data sets have been broadly collected in classification problems. Classical LDA has been proved to perform poorly (no better than random guessing) in the high-dimensional setting, especially when the dimension is much larger than the sample size [3]. To address high-dimensional issue, the sparsity assumption is introduced to LDA. Several proposed methods assumed that both $\mathbf{\Sigma}^{-1}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ have sparse structures. For example, [30] used the thresholding procedure to estimate $\mathbf{\Sigma}^{-1}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ separately, then constructed a plug-in sparse Bayes' linear discriminant rule. Similar regularized methods can also be found in [16, 35, 34], etc. In addition, some works only assumed Bayes' linear discriminant direction $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is sparse. [4] proposed the linear programming discriminant (LPD) rule by directly estimating the product $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ through constrained ℓ_1 minimization. Recently, [6] proposed an adaptive LPD procedure that achieved the minimax optimal convergence rate of estimation error and excess misclassification risk in high-dimensional case. [25] estimated the sparse discriminant direction via a sparse penalized least squares formulation. [24] studied high-dimensional sparse semiparametric discriminant analysis and relaxed the Gaussian assumption. For multiclass problem, [23] proposed a sparse discriminant procedure by estimating all discriminant directions simultaneously.

Heterogeneous data is widespread in many modern scientific fields, such as finance, biology, and astronomy [12]. The prevalent statistical approach to address the heterogeneity is imposing the dynamic or varying coefficient assumption, where the population means and covariance matrix may vary with some observable exposure variable. In specific, [9, 10, 32] investigated the dynamic covariance model in the high-dimensional regime. Under the dynamic setting, Bayes' discriminant direction is a function of the exposure variable. Consequently, classical plug-in Bayes' discriminant rule will deteriorate in analyzing non-static data, and thus leads to unsatisfactory performance. To address the dynamic data, [21] proposed the dynamic linear programming discriminant (DLPD) rule by assuming $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\mathbf{\Sigma}$ are functions of some q-dimensional random covariate $\boldsymbol{U}$. To estimate the sparse Fisher's linear discriminant direction function $\boldsymbol{\beta}^*(\boldsymbol{u}) = \mathbf{\Sigma}^{-1}(\boldsymbol{u})(\boldsymbol{\mu}_1(\boldsymbol{u}) - \boldsymbol{\mu}_2(\boldsymbol{u}))$ given $\boldsymbol{U} = \boldsymbol{u}$, they first used the Nadaraya-Watson method to obtain estimators $\widehat{\boldsymbol{\mu}}_1(\boldsymbol{u})$, $\widehat{\boldsymbol{\mu}}_2(\boldsymbol{u})$ and $\widehat{\mathbf{\Sigma}}(\boldsymbol{u})$. Then they estimated $\boldsymbol{\beta}^*(\boldsymbol{u})$ using the linear programming approach [4, 8]:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{u}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \|\boldsymbol{\beta}\|_1 \text{ subject to } |\widehat{\mathbf{\Sigma}}(\boldsymbol{u})\boldsymbol{\beta} - (\widehat{\boldsymbol{\mu}}_1(\boldsymbol{u}) - \widehat{\boldsymbol{\mu}}_2(\boldsymbol{u}))\,|_\infty \le \lambda_n \right\}, \tag{1.2}$$

where λ_n is a tuning parameter. However, this classification procedure is computationally expensive for large scale prediction problem. For each new observation $(\boldsymbol{X}_{\text{new}}, \boldsymbol{U}_{\text{new}})$, DLPD method needs to re-estimate $\widehat{\boldsymbol{\mu}}_1(\boldsymbol{U}_{\text{new}})$, $\widehat{\boldsymbol{\mu}}_2(\boldsymbol{U}_{\text{new}})$ and $\widehat{\mathbf{\Sigma}}(\boldsymbol{U}_{\text{new}})$ and re-solve the corresponding large scale linear programming (1.2). In addition, the support set of discriminant direction $\boldsymbol{\beta}^*(\boldsymbol{u})$ decides which variable contributes to classification but (1.2) can not provide a invariant support set since it is a point-wise estimator. In some real applications, the varying support set of discriminant direction in DLPD method may lack interpretability.

The dynamic discriminant analysis shares the same semi-parametric spirit with the classical varying coefficient model [17], where the unknown parameters are assumed to be a smooth function of the exposure variable. In the past decades, the varying coefficient method has been applied to a variety of statistical models, such as linear regression model [19, 15], generalized linear model [7, 14], quantile regression [18, 33] and support vector machine [22], etc. Motivated by the least square form of Bayes' discriminant direction, we propose a new estimation method for the discriminant direction function based on B-spline approximation, which can be applied in the classification for dynamic data. In high-dimensional regime, we can estimate the approximation coefficient by solving a penalized least square problem. The computational drawback of the DLPD rule [21] is circumvented in our developed varying coefficient discriminant procedure. For each new observation, we only need to re-compute the B-spline basis vector. Hence it has a significant computational advantage over the DLPD rule. In the high-dimensional case, the support set of our proposed estimator is irrelevant with the value of exposure variable, which is indeed helpful to select important features contributing to classification.

The remainder of this paper is organized as follows. In section 2, we propose a new discriminant direction function and its varying coefficient estimators in both low-dimensional and high-dimensional regimes. In section 3, we establish the upper bounds for the estimation error and uniform excess misclassification risk for our proposed varying coefficient LDA procedure. In section 5 and 6, we verify the performance of our method through simulations on synthetic data and real data respectively.

Notations We define some notations that will be used throughout the paper. For two real positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists some positive constant m such that $a_n \leq m b_n$. And we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a real-valued vector $\boldsymbol{x} \in \mathbb{R}^p$, we use $\|\boldsymbol{x}\|_1 = \sum_{j=1}^p |x_j|$, $\|\boldsymbol{x}\|_2 = (\sum_{j=1}^p |x_j|^2)^{1/2}$ and $|\boldsymbol{x}|_\infty = \max_{1 \leq j \leq p} |x_j|$ to denote the ℓ_1, ℓ_2 and ℓ_∞ norm respectively. For a subset $S \subseteq \{1, 2, ..., p\}$, we use $\boldsymbol{x}_S$ to denote the sub-vector $(x_j : j \in S)$. Specially, for a vector $\boldsymbol{b} \in \mathbb{R}^{pq}$, we write sub-vector $\boldsymbol{b}_{(j)} = (b_{(j-1)q+1}, \cdots, b_{jq})^\top$ for $j = 1, 2, ..., p$. And for a subset $S \subseteq \{1, 2, ..., p\}$, we use $\boldsymbol{b}_{(S)}$ to denote the group sub-vector $(\boldsymbol{b}_{(j)} : j \in S)$. For a real-valued matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$, ℓ_2 (spectral) norm is defined by $\|\mathbf{A}\|_2 = \sup_{\|\boldsymbol{x}\|_2=1, \|\boldsymbol{y}\|_2=1} |\boldsymbol{x}^\top \mathbf{A} \boldsymbol{y}|$, the maximal entry in absolute value is denoted by $|\mathbf{A}|_\infty = \max_{i,j} |A_{ij}|$. For two subsets $S \subseteq \{1, 2, ..., p\}$ and $T \subseteq \{1, 2, ..., q\}$, we write sub-matrix $\mathbf{A}_{ST} = (A_{ij})$ for $i \in S$ and $j \in T$. We use $\mathbf{A} \otimes \boldsymbol{B}$ to denote the Kronecker product on two matrices $\mathbf{A}$ and $\boldsymbol{B}$ with proper sizes. Specially, for a matrix $\mathbf{A} \in \mathbb{R}^{pq \times pq}$ and two subsets $S, T \subseteq \{1, 2, ..., p\}$, we write the group sub-matrix as $\mathbf{A}_{(ST)} = (A_{ij})$ for $i \in \{(k-1)q + 1, ..., kq : k \in S\}$ and $j \in \{(k-1)q + 1, ..., kq : k \in T\}$. For a sequence of real random variables X_n, we write $X_n = O_{\mathbb{P}}(a_n)$ if for any $\epsilon > 0$, there exists some constant $C > 0$ such that $\mathbb{P}(|X_n| > C a_n) < \epsilon$.

2. Varying coefficient LDA via B-spline approximation

In this section, we provide a detailed description of the varying coefficient linear discriminant rule. Given the univariate exposure variable $U = u \in [0,1]$, we assume $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_1(u), \boldsymbol{\Sigma}(u))$ if $Y = 1$ and $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_2(u), \boldsymbol{\Sigma}(u))$ if $Y = 0$, then Bayes' discriminant direction is $\boldsymbol{\beta}^*(u) = \boldsymbol{\Sigma}^{-1}(u)(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))$. We also denote the pooled mean as $\boldsymbol{\mu}(u) = \pi_1\boldsymbol{\mu}_1(u) + \pi_2\boldsymbol{\mu}_2(u)$, where $\pi_1 = \mathbb{P}(Y = 1)$ and $\pi_2 = \mathbb{P}(Y = 0)$. To introduce our new discriminant direction function, we define a new response variable as $Z = \pi_2$ if $Y = 1$ and $Z = -\pi_1$ if $Y = 0$. In addition, the exposure variable U is assumed to be independent with the label Y. Motivated by the least square form of the plug-in Bayes' discriminant direction $\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$ in static setting [1, 25], we propose a new discriminant direction function $\boldsymbol{\theta}^*(U) = (\theta_1^*(U), \cdots, \theta_p^*(U))^\top$ as the minimizer of the following population least square problem

$$\min_{\theta_j(U) \in \mathcal{L}^2(\mathcal{P})} \mathbb{E}\left[\left(Z - \sum_{j=1}^{p} \theta_j(U)(X_j - \mu_j(U))\right)^2 \Big| U\right], \tag{2.1}$$

where $\mathcal{P}$ is the joint distribution of $(\boldsymbol{X}, Z, U)$ and $\mathcal{L}^2(\mathcal{P})$ denotes the L^2 space under measure $\mathcal{P}$. It is worthwhile noting that the representation (2.1) is similar to the approximation of coefficient function in the varying coefficient linear model. Then the discriminant direction function $\boldsymbol{\theta}^*(U)$ satisfies

$$\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}(U))(Z - (\boldsymbol{X} - \boldsymbol{\mu}(U))^\top \boldsymbol{\theta}^*(U))|U\right] = \mathbf{0}.$$

A further computation gives rise to the following closed form,

$$\boldsymbol{\theta}^*(U) = \pi_1\pi_2\boldsymbol{\Sigma}^{-1}(U)(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))[1 - (\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))^\top \boldsymbol{\theta}^*(U)].$$

If the population covariance matrix $\boldsymbol{\Sigma}(U)$ is positive definite and $\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U) \neq 0$, we are guaranteed that $(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))^\top \boldsymbol{\theta}^*(U) \in (0, 1)$. As a consequence, Bayes' discriminant direction function satisfies that

$$\boldsymbol{\beta}^*(U) = \boldsymbol{\Sigma}^{-1}(U)(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U)) = c^*(U)\boldsymbol{\theta}^*(U), \tag{2.2}$$

where $c^*(U) = 1/[\pi_1\pi_2 - \pi_1\pi_2(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))^\top \boldsymbol{\theta}^*(U)]$. For the equal-prior case ($\pi_1 = \pi_2$), given any new observation $(\boldsymbol{X}_{\text{new}}, U_{\text{new}})$, we define the oracle varying coefficient discriminant rule as

$$\psi(\boldsymbol{X}_{\text{new}}, U_{\text{new}}) = \mathbb{I}\left((\boldsymbol{X}_{\text{new}} - \boldsymbol{\mu}(U_{\text{new}}))^\top \boldsymbol{\theta}^*(U_{\text{new}}) \geq 0\right). \tag{2.3}$$

Recall that $c^*(u) > 0$, the classification result of (2.3) is consistent with using Bayes' discriminant direction $\boldsymbol{\beta}^*(U_{\text{new}})$.

2.1. Approximation of discriminant direction function

Let $\boldsymbol{B}(\cdot) = (B_1(\cdot), ..., B_{L_n}(\cdot))$ be the scaled B-spline basis of the polynomial splines space, which satisfies that $B_k(\cdot) \geq 0$ and $\sum_{k=1}^{L_n} B_k(\cdot) = \sqrt{L_n}$. According to the B-spline approximation theory [11], under some regular conditions, each coordinate of discriminant direction $\boldsymbol{\theta}^*(u)$ can be approximated by $\theta_j^*(u) \approx \boldsymbol{\gamma}_{(j)}^\top \boldsymbol{B}(u)$, where $\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}$ is the approximation coefficient. If $\boldsymbol{\mu}_1(u)$ and $\boldsymbol{\mu}_2(u)$ are known, the "best" approximation coefficients in population form are defined as

$$(\widetilde{\boldsymbol{\gamma}}_{(1)}, \cdots, \widetilde{\boldsymbol{\gamma}}_{(p)}) = \arg \min_{\substack{\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}, \\ 1 \leq j \leq p}} \mathbb{E}\left[\left(Z - \sum_{j=1}^{p}(X_j - \mu_j(U))\boldsymbol{\gamma}_j^\top \boldsymbol{B}(U)\right)^2\right]. \quad (2.4)$$

Let $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}_{(1)}^\top, \cdots, \widetilde{\boldsymbol{\gamma}}_{(p)}^\top)^\top$ and $\widetilde{\boldsymbol{B}}(U) = (\boldsymbol{X} - \boldsymbol{\mu}(U)) \otimes \boldsymbol{B}(U)$, it is easy to show that

$$\widetilde{\boldsymbol{\gamma}} = \left(\mathbb{E}[\widetilde{\boldsymbol{B}}(U)\widetilde{\boldsymbol{B}}(U)^\top]\right)^{-1} \mathbb{E}[\widetilde{\boldsymbol{B}}(U)Z].$$

For any $u \in [0, 1]$, we may write the approximated discriminant direction as

$$\widetilde{\boldsymbol{\theta}}(u) = \left(\widetilde{\boldsymbol{\gamma}}_{(1)}^\top \boldsymbol{B}(u), \cdots, \widetilde{\boldsymbol{\gamma}}_{(p)}^\top \boldsymbol{B}(u)\right)^\top.$$

Therefore, the data-driven discriminant procedure boils down to estimate the approximation coefficient $\widetilde{\boldsymbol{\gamma}}$ and the mean functions $\boldsymbol{\mu}_1(u)$, $\boldsymbol{\mu}_2(u)$ based on collected samples. In the following subsections, we consider the equal-prior case, that is $\pi_1 = \pi_2 = 1/2$. And we provide the extension of our method to unbalanced case in Section 4.1.

2.2. Data-driven discriminant procedure

Let $\{(\boldsymbol{X}_i, U_i, Y_i) : i = 1, 2, ..., 2n\}$ be an i.i.d. sample set. We denote the pseudo response variable by $Z_i = \mathbb{I}(Y_i = 1) - \frac{1}{2}$ for $i = 1, 2, ..., 2n$ and denote the value of B-spline basis taken at U_i by $\boldsymbol{B}_i = (B_1(U_i), ..., B_{L_n}(U_i))^\top$. Without loss of generality, we assume the sample size of the two classes is equal. The sample index sets of two classes are $\mathcal{I}_1 = \{i : Y_i = 1\}$ and $\mathcal{I}_2 = \{i : Y_i = 0\}$ with $|\mathcal{I}_1| = |\mathcal{I}_2| = n$.

2.2.1. Classical low-dimensional regime

To construct the sample form of problem (2.4), we start with estimating the mean functions $\boldsymbol{\mu}_1(u)$ and $\boldsymbol{\mu}_2(u)$. By the B-spline theory, we may estimate the mean functions by $\widehat{\boldsymbol{\mu}}_l(u) = (\widehat{\boldsymbol{\alpha}}_{l1}^\top \boldsymbol{B}(u), \cdots, \widehat{\boldsymbol{\alpha}}_{lp}^\top \boldsymbol{B}(u))^\top$ for $l = 1, 2$, where

$$\widehat{\boldsymbol{\alpha}}_{lj} = \left(\sum_{i \in \mathcal{I}_l} \boldsymbol{B}_i \boldsymbol{B}_i^\top\right)^{-1} \sum_{i \in \mathcal{I}_l} \boldsymbol{B}_i X_{ij}, \quad \text{for } j = 1, 2, ..., p.$$

Let $\widehat{\boldsymbol{\mu}}(u) = (\widehat{\boldsymbol{\mu}}_1(u) + \widehat{\boldsymbol{\mu}}_2(u))/2$, the estimators $(\widehat{\boldsymbol{\gamma}}_{(1)}, \cdots, \widehat{\boldsymbol{\gamma}}_{(p)})$ can be obtained by solving the following least-square problem

$$\min_{\substack{\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}, \\ 1 \le j \le p}} \frac{1}{2n} \sum_{i=1}^{2n} \left(Z_i - \sum_{j=1}^{p} (X_{ij} - \widehat{\mu}_j(U_i)) \boldsymbol{B}_i^\top \boldsymbol{\gamma}_{(j)} \right)^2 . \tag{2.5}$$

With slightly abusing notations, we denote $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}_{(1)}^\top, \cdots, \widehat{\boldsymbol{\gamma}}_{(p)}^\top)^\top$ and $\widetilde{\boldsymbol{B}}_i = (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}(U_i)) \otimes \boldsymbol{B}_i$. In low-dimensional regime, the problem (2.5) has a closed form solution

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i \widetilde{\boldsymbol{B}}_i^\top \right)^{-1} \sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i Z_i. \tag{2.6}$$

2.2.2. Sparse high-dimensional regime

In the high-dimensional case, we assume Bayes' discriminant function $\boldsymbol{\beta}^*(u)$ is sparse with the support set $S := \{j : \mathbb{E}[|\beta_j^*(U)|^2] > 0\}$ and $|S| = s$. Without loss of generality, let $S = \{1, ..., s\}$.

Since $\boldsymbol{\theta}^*(u)$ has the same support set with $\boldsymbol{\beta}^*(u)$, the "best" coefficients for approximating $\theta_j^*(u)$ for $j \in S$ are defined as

$$(\widetilde{\boldsymbol{\gamma}}_{(1)}, \cdots, \widetilde{\boldsymbol{\gamma}}_{(s)}) = \arg \min_{\substack{\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}, \\ 1 \le j \le s}} \mathbb{E} \left[\left(Z - \sum_{j=1}^{s} (X_j - \mu_j(U)) \boldsymbol{\gamma}_{(j)}^\top \boldsymbol{B} \right)^2 \right] . \tag{2.7}$$

Consequently, for any $u \in [0, 1]$, we shall approximate the discriminant direction function $\boldsymbol{\theta}^*(u)$ by

$$\widetilde{\boldsymbol{\theta}}(u) = \left(\widetilde{\boldsymbol{\gamma}}_1^\top \boldsymbol{B}(u), \cdots, \widetilde{\boldsymbol{\gamma}}_s^\top \boldsymbol{B}(u), 0, \cdots, 0 \right)^\top .$$

Let $\mathbf{D} = \mathbb{E}[\widetilde{\boldsymbol{B}}(U) \widetilde{\boldsymbol{B}}(U)^\top]$ and $\boldsymbol{b} = \mathbb{E}[\widetilde{\boldsymbol{B}}(U) Z]$, the approximation coefficient vector can be equivalently written as $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}_1^\top, \cdots, \widetilde{\boldsymbol{\gamma}}_s^\top, \mathbf{0}^\top, \cdots, \mathbf{0}^\top)^\top = (\widetilde{\boldsymbol{\gamma}}_{(S)}^\top, \widetilde{\boldsymbol{\gamma}}_{(S^c)}^\top)^\top$ where $\widetilde{\boldsymbol{\gamma}}_{(S)} = \mathbf{D}_{(SS)}^{-1} \boldsymbol{b}_{(S)}$ and $\widetilde{\boldsymbol{\gamma}}_{(S^c)} = \mathbf{0}_{(p-s)L_n}$. It means that the estimator of approximation coefficient $\widetilde{\boldsymbol{\gamma}}$ should have *group sparsity* structure. Therefore, we add the group lasso penalty [38] to the objective function in (2.5), and then obtain the estimators $(\widehat{\boldsymbol{\gamma}}_{(1)}, \cdots, \widehat{\boldsymbol{\gamma}}_{(p)})$ by solving

$$\min_{\substack{\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}, \\ 1 \le j \le p}} \frac{1}{2n} \sum_{i=1}^{2n} \left(Z_i - \sum_{j=1}^{p} (X_{ij} - \widehat{\mu}_j(U_i)) \boldsymbol{B}(U_i)^\top \boldsymbol{\gamma}_{(j)} \right)^2 + \lambda_n \sum_{j=1}^{p} \|\boldsymbol{\gamma}_{(j)}\|_2, \tag{2.8}$$

where λ_n is a tuning parameter. After some simplifications, the problem (2.8) is equivalent to the following quadratic programming form

$$\widehat{\boldsymbol{\gamma}} = \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^{pL_n}} \left\{ \frac{1}{2} \boldsymbol{\gamma}^\top \mathbf{D}_n \boldsymbol{\gamma} - \boldsymbol{b}_n^\top \boldsymbol{\gamma} + \lambda_n \sum_{j=1}^{p} \|\boldsymbol{\gamma}_{(j)}\|_2 \right\}, \tag{2.9}$$

where

$$\mathbf{D}_n = \frac{1}{2n}\sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i \widetilde{\boldsymbol{B}}_i^\top, \quad \boldsymbol{b}_n = \frac{1}{2n}\sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i Z_i.$$

The problem (2.9) can be efficiently solved by several well studied optimization methods, such as group coordinate descent algorithm and iterative shrinkage thresholding algorithm (ISTA) [2]. We provide a detailed description about ISTA to solve (2.9) in Appendix E.

2.2.3. Discriminant rule and asymptotic optimality

After obtaining $\widehat{\boldsymbol{\mu}}(u)$ and $\widehat{\boldsymbol{\gamma}}$, the estimator of discriminant direction function is given by

$$\widehat{\boldsymbol{\theta}}(u) = \left(\widehat{\boldsymbol{\gamma}}_{(1)}^\top \boldsymbol{B}(u), \cdots, \widehat{\boldsymbol{\gamma}}_{(p)}^\top \boldsymbol{B}(u)\right)^\top. \tag{2.10}$$

For any new observation $(\boldsymbol{X}_{\text{new}}, U_{\text{new}})$, the data-driven varying coefficient linear discriminant rule is

$$\widehat{\psi}(\boldsymbol{X}_{\text{new}}, U_{\text{new}}) = \mathbb{I}\left((\boldsymbol{X}_{\text{new}} - \widehat{\boldsymbol{\mu}}(U_{\text{new}}))^\top \widehat{\boldsymbol{\theta}}(U_{\text{new}}) \geq 0\right). \tag{2.11}$$

For any $u \in [0,1]$, the optimal misclassification risk of oracle rule (2.3) is $R(u) = \Phi(-\Delta(u)/2)$, where $\Delta(u) = \sqrt{(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))^\top \boldsymbol{\Sigma}^{-1}(u)(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))}$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Given the samples and $u \in [0,1]$, the conditional misclassification risk of data-driven rule (2.11) is

$$R_n(u) := \frac{1}{2}\Phi\left(\frac{(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}_1(u))^\top \widehat{\boldsymbol{\theta}}(u)}{\sqrt{\widehat{\boldsymbol{\theta}}^\top(u)\boldsymbol{\Sigma}(u)\widehat{\boldsymbol{\theta}}(u)}}\right) + \frac{1}{2}\bar{\Phi}\left(\frac{(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}_2(u))^\top \widehat{\boldsymbol{\theta}}(u)}{\sqrt{\widehat{\boldsymbol{\theta}}^\top(u)\boldsymbol{\Sigma}(u)\widehat{\boldsymbol{\theta}}(u)}}\right),$$

where $\widehat{\boldsymbol{\mu}}(u) = (\widehat{\boldsymbol{\mu}}_1(u) + \widehat{\boldsymbol{\mu}}_2(u))/2$ and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$. Through utilizing the technique developed in [6], we have the following proposition to provide an upper bound for the excess misclassification risk.

Proposition 2.1. *Suppose that for any $u \in [0,1]$, $\|\boldsymbol{\Sigma}(u)\|_2$ is uniformly upper bounded from infinity and $\Delta(u)$ is uniformly lower bounded away from zero. In addition, if $\|\widehat{\boldsymbol{\mu}}_1(u) - \boldsymbol{\mu}_1(u)\|_2 = o(1)$, $\|\widehat{\boldsymbol{\mu}}_2(u) - \boldsymbol{\mu}_2(u)\|_2 = o(1)$ and $\|\widehat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}^*(u)\|_2 = o(1)$ for any $u \in [0,1]$, we have*

$$|R_n(u) - R(u)| \lesssim \|\widehat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}^*(u)\|_2^2 + |\left(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u)\right)^\top \boldsymbol{\beta}^*(u)|^2. \tag{2.12}$$

3. Theoretical results

In this section, we will present the estimation error bounds and the convergence rates of excess misclassification risk of our proposed varying coefficient LDA procedure in both low-dimensional regime and high-dimensional regime. Specially, for two function vectors $\boldsymbol{\nu}(\cdot) = (\nu_1(\cdot), ..., \nu_m(\cdot))^\top$ and $\boldsymbol{\xi}(\cdot) = (\xi_1(\cdot), ..., \xi_m(\cdot))^\top$ mapping from $[0,1]$ to $\mathbb{R}^p$, we define the L_2 distance between $\boldsymbol{\nu}(\cdot)$ and $\boldsymbol{\xi}(\cdot)$ as

$$\|\boldsymbol{\nu} - \boldsymbol{\xi}\|_{\mathrm{L}_2} = \left(\int_0^1 \|\boldsymbol{\nu}(u) - \boldsymbol{\xi}(u)\|_2^2 du \right)^{\frac{1}{2}}.$$

3.1. Classical low-dimensional regime

Before presenting the convergence rates of our proposed estimator, we introduce the following necessary technical assumptions for the clarity of ensuing theoretical results.

(**C**1) There exist two constants $0 < \lambda_0 \le \lambda_1 < \infty$ such that for any $u \in [0,1]$

$$\lambda_0 \le \lambda_{\min}(\boldsymbol{\Sigma}(u)) \le \lambda_{\max}(\boldsymbol{\Sigma}(u)) \le \lambda_1,$$

where $\lambda_{\min}(\boldsymbol{\Sigma}(u))$ and $\lambda_{\max}(\boldsymbol{\Sigma}(u))$ are respectively the minimum and maximum eigenvalues of $\boldsymbol{\Sigma}(u)$.

(**C**2) The density function h of the exopsure U satisfies that $0 < D_1 \le h(u) \le D_2 < \infty$ for two positive constants D_1 and D_2 and any $u \in [0,1]$.

(**C**3) Each entry of functions $\boldsymbol{\mu}_1(u)$, $\boldsymbol{\mu}_2(u)$ and $\boldsymbol{\Sigma}(u)^{-1}(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))$ belongs to the following function space

$$\mathcal{W}^d([0,1]) := \Big\{ f : [0,1] \to \mathbb{R}, \ \sup_x |f^{(\ell)}(x)| \le D \text{ for } \ell = 0, 1, ..., t \text{ and } \sup_{x,x'} |f^{(t)}(x) - f^{(t)}(x')| \le L|x - x'|^r \Big\}$$

where $d = r + t \ge 1$ and $f^{(s)}$ denotes the s-th derivative of function f and $f^{(0)} = f$.

(**C**4) Assume $\sup_{u \in [0,1]} \max\{\|\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u)\|_2, \|\boldsymbol{\theta}^*(u)\|_2\} = \delta_p \le M$ for some large constant M. In addition, $p = o(n^{(2d-1)/(2d+1)})$.

Assumption (**C**1) is very common in high-dimensional linear discriminant analysis literature [25, 4, 21]. Assumptions (**C**2) and (**C**3) are regular conditions in B-spline approximation theory, similar assumptions also appear in [37, 13]. For the simplicity of convergence rates, we assume $\|\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u)\|_2$ and $\|\boldsymbol{\theta}^*(u)\|_2$ are both uniformly bounded in (**C**4). The condition on the dimensionality ensures that $L_n\sqrt{p \log n / n} = o(1)$ under the optimal length of B-spline basis $L_n \asymp n^{1/(2d+1)}$, which guarantees the optimality of our proposed estimator.

Note that L_2 error of our proposed estimator can be decomposed into two parts: the approximation error $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{L}_2}$ and the estimation error $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\mathrm{L}_2}$. Our first result shows that the approximation error shrinks as the length of spline basis vector L_n grows, which also attains the optimal convergence rate of classical B-spine approximation error (see [20, 29]). The proof of Theorem 3.1 is given in Appendix B.2.

Theorem 3.1. *Assume the assumptions* $(\mathbf{C}1)$-$(\mathbf{C}4)$ *hold, then the approximation error in* L_2 *distance is bounded by*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{L}_2} \lesssim \sqrt{p} L_n^{-d}. \tag{3.1}$$

The following theorem provides the upper bound of estimation error for the discriminant direction function estimator (2.10). Compared with the analysis in the varying coefficient linear model, the theoretical development in this paper is more challenging. The reason is two-fold:

- There is no direct relation between the pseudo-response variable Z_i and the covariate $\boldsymbol{X}_i$. The empirical processes in the proof are established upon a fine-grained decomposition to $\mathbf{D}_n - \mathbf{D}$ (see Appendix C.3).
- The estimator for approximation coefficient $\widehat{\boldsymbol{\gamma}}$ in (2.5) involves the mean function estimators $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ computed from the same samples. We utilize chaining technique to establish several concentration inequalities on the operator norm of matrices and ℓ_2 norm of matrix-vector-products (see Lemma C.4-C.5).

The proof of Theorem 3.2 is deferred to Appendix B.3.

Theorem 3.2. *Assume conditions* $(\mathbf{C}1)$-$(\mathbf{C}4)$ *hold. Let* $a_n = \sqrt{L_n \log n / n} + L_n^{-d}$*, the estimation error in* L_2 *distance is bounded by*

$$\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\mathrm{L}_2} = O_{\mathbb{P}}\left(\sqrt{\frac{pL_n \log n}{n}} + a_n p L_n \sqrt{\frac{\log n}{n}} + \sqrt{p} L_n^{-d}\right). \tag{3.2}$$

Remark 3.1. *Together with the approximation error in Theorem 3.1 and assumption* $(\mathbf{C}4)$*, if we take the length of B-spline vector as* $L_n \asymp (n/\log n)^{\frac{1}{2d+1}}$*, it is easy to see that the* L_2 *error can be bounded by*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{L}_2} = O_{\mathbb{P}}\left(\sqrt{p}\left(\frac{\log n}{n}\right)^{\frac{d}{2d+1}}\right). \tag{3.3}$$

According to [31], the minimax convergence rate for one-dimensional function in function space $\mathcal{W}^d([0,1])$ *is* $n^{-d/(2d+1)}$*. Apparently, our proposed estimation procedure is optimal up to a logarithmic factor.*

From assumptions $(\mathbf{C}1)$ and $(\mathbf{C}4)$, we know $|\left(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u)\right)^\top \boldsymbol{\beta}^*(u)|^2 \lesssim \|\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u)\|_2^2$. In Proposition A.1, we establish the uniform bound for the mean function estimator, that is

$$\sup_{u\in[0,1]} \|\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u)\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{pL_n \log n}{n}} + \sqrt{p} L_n^{-d}\right).$$

In addition, we also have $\sup_{u\in[0,1]} \|\widehat{\boldsymbol{\theta}}(u) - \widetilde{\boldsymbol{\theta}}(u)\|_2 = O_{\mathbb{P}}(L_n \sqrt{p \log n/n})$ since $\|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_2$ shares the same bound with (3.2) (see Appendix B.3) and $\|\boldsymbol{B}(u)\|_2 \leq \sqrt{L_n}$. In conjunction with (3.3), we can obatin the L_2 bound of the excess misclassification risk in the following corollary.

Corollary 3.1. *Under the same settings of Theorem 3.2, we assume* $\Delta(u) \geq c > 0$ *for some constant* c *and take* $L_n \asymp (n/\log n)^{\frac{1}{2d+1}}$, *then it holds*

$$\|R_n - R\|_{\mathrm{L}_2} = O_{\mathbb{P}}\left(p\left(\frac{\log n}{n}\right)^{\frac{2d}{2d+1}}\right).$$

3.2. Sparse high-dimensional regime

The following assumption plays a similar role as the condition (**C**4) in low-dimensional regime.

(**C**5) Assume $\sup_{u\in[0,1]} \max\{\|(\boldsymbol{\mu}_1(u)-\boldsymbol{\mu}_2(u)_S\|_2, \|\boldsymbol{\theta}^*(u)\|_2\} = \delta_s \leq M$ for some large constant M. In addition, $s = o(n^{(2d-1)/4(d+1)})$.

The approximation error bound under sparse setting is presented in the following theorem, which can be easily obtained by tracing the proof of Theorem 3.1 since $\theta_j^*(\cdot) = \widetilde{\theta}_j(\cdot) = 0$ for $j \in S^c$.

Theorem 3.3. *Assume the assumptions* (**C**1)-(**C**3) *and* (**C**5) *hold, then the approximation error in high-dimensional case is*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{L}_2} \lesssim \sqrt{s}L_n^{-d}.$$

Below we provide the estimation error bound for the group-sparse estimator in (2.9), and the proof is deferred to Appendix B.4.

Theorem 3.4. *Assume conditions* (**C**1), (**C**2), (**C**3) *and* (**C**5) *hold, let* $a_n = \sqrt{L_n \log n/n} + L_n^{-d}$, *for any* $\vartheta > 0$, *if we take*

$$\lambda_n \geq C\left(\sqrt{\frac{L_n \log p}{n}} + a_n L_n s\sqrt{\frac{\log p}{n}} + \sqrt{s}L_n^{-d}\right) \tag{3.4}$$

for some sufficiently large positive constant C, *then*

$$\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\|_{\mathrm{L}_2} \lesssim \sqrt{s}\lambda_n$$

holds with probability at least $1 - L_n p^{-\vartheta} - L_n p^{-\vartheta s L_n}$.

Remark 3.2. *To interpret the orders in (3.4), we introduce the crucial quantity in the proof of Theorem 3.4:* $\max_{1\leq j\leq p} \|(\mathbf{D}_n)_{(jS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - (\boldsymbol{b}_n)_{(j)}\|_2$, *which can be bounded by*

$$\begin{aligned}\|(\mathbf{D}_n)_{(jS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - (\boldsymbol{b}_n)_{(j)}\|_2 \leq \|(\mathbf{D}_n - \mathbf{D})_{(jS)}\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2 + \|(\boldsymbol{b}_n - \boldsymbol{b})_{(j)}\|_2 \\ + \|\mathbf{D}_{(jS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - \boldsymbol{b}_{(j)}\|_2.\end{aligned} \tag{3.5}$$

For any $1 \leq j \leq p$, *the first two terms in (3.5) can be bounded by* $\sqrt{L_n \log p/n} + a_n L_n s\sqrt{\log p/n}$ *through concentration. For* $j \in S$, $\mathbf{D}_{(jS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - \boldsymbol{b}_{(j)} = \mathbf{0}$ *holds due to the definition of* $\widetilde{\boldsymbol{\gamma}}$ *in (2.7). For* $j \notin S$, *despite the fact* $\mathbf{D}_{(S^cS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - \boldsymbol{b}_{(S^c)} \neq \mathbf{0}$, *we can still show that it is bounded by* $\|(\boldsymbol{\theta}^*(u) - \widetilde{\boldsymbol{\theta}}(u))_S\|_2$ *(see Appendix B.4), which is exactly the last term* $\sqrt{s}L_n^{-d}$ *in (3.4).*

If we set the length of B-spline basis vector as $L_n \asymp (ns/\log p)^{\frac{1}{2d+1}}$, the L_2 error of the group-sparse estimator $\widehat{\boldsymbol{\theta}}(\cdot)$ will be

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\mathrm{L}_2} = O_{\mathbb{P}}\left(s^{\frac{d+1}{2d+1}}\left(\frac{\log p}{n}\right)^{\frac{d}{2d+1}}\right). \tag{3.6}$$

Compared with the oracle minimax rate $\sqrt{s}n^{-\frac{2d}{2d+1}}$, there is an additional factor $s^{1/(2d+1)}$ in (3.6) due to the bias $\mathbf{D}_{(S^cS)}\widetilde{\boldsymbol{\gamma}}_{(S)} - \boldsymbol{b}_{(S^c)}$. To obtain the convergence rate for the excess misclassification risk, it suffices to control the upper bound of $|(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u))^\top \boldsymbol{\beta}^*(u)|^2$. Recall the fact $\boldsymbol{\Sigma}(u)\boldsymbol{\beta}^*(u) = \boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u)$, then simple algebra shows that $\boldsymbol{\beta}^*_S(u) = (\boldsymbol{\Sigma}_{SS}(u))^{-1}(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))_S$. Combining with condition (**C**5), we have $|(\widehat{\boldsymbol{\mu}}(u)-\boldsymbol{\mu}(u))^\top\boldsymbol{\beta}^*(u)|^2 \lesssim \|(\widehat{\boldsymbol{\mu}}(u)-\boldsymbol{\mu}(u)_S\|_2^2$. Then the following corollary is a direct result of (3.6) and Proposition A.1.

Corollary 3.2. *With the same conditions and choice of λ_n in Theorem 3.4, if we take $L_n \asymp (ns/\log p)^{\frac{1}{2d+1}}$, the excess misclassification risk of $\widehat{\boldsymbol{\theta}}$ satisfies that*

$$\|R_n - R\|_{\mathrm{L}_2} = O_{\mathbb{P}}\left(s^{\frac{2d+2}{2d+1}}\left(\frac{\log p}{n}\right)^{\frac{2d}{2d+1}}\right).$$

4. Extensions

This section will generalize our approach to more general classification problems in dynamic data.

4.1. Binary classification with unequal prior

For general static binary classification problem, Bayes' discriminant rule is given by

$$\psi(\boldsymbol{X}) = \mathbb{I}\left((\boldsymbol{X} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log\frac{\pi_1}{\pi_2} \geq 0\right) \tag{4.1}$$

where $\pi_1 = \mathbb{P}(Y = 1), \pi_2 = \mathbb{P}(Y = 0)$ and $\boldsymbol{\mu} = \pi_1\boldsymbol{\mu}_1 + \pi_2\boldsymbol{\mu}_2$. In varying coefficient regime, according to (2.2), (4.1) can be generalized to the following form

$$\psi(\boldsymbol{X}, U) = \mathbb{I}\left((\boldsymbol{X} - \boldsymbol{\mu}(U))^\top c^*(U)\boldsymbol{\theta}^*(U) + \log\frac{\pi_1}{\pi_2} \geq 0\right), \tag{4.2}$$

where $c^*(U) = 1/[\pi_1\pi_2 - \pi_1\pi_2(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))^\top\boldsymbol{\theta}^*(U)]$. The prior probabilities can be estimated by $\widehat{\pi}_1 = \sum_{i=1}^N \mathbb{I}(Y_i = 1)/N$ and $\widehat{\pi}_2 = \sum_{i=1}^N \mathbb{I}(Y_i = 0)/N$, where N is the total sample size. To estimate $\boldsymbol{\theta}^*(u)$, we only need to set $Z_i = \widehat{\pi}_2$ if $Y_i = 1$ and $Z_i = -\widehat{\pi}_1$ if $Y_i = 0$ in (2.8). As a consequence, for any new observation $(\boldsymbol{X}_{\mathrm{new}}, U_{\mathrm{new}})$, we can perform varying coefficient discriminant rule by plugging in corresponding estimators into (4.2).

4.2. Multivariate exposure variable

For multivariate $\boldsymbol{U} \in \mathbb{R}^m$, we may consider the following single-index extension. Specially, given $\boldsymbol{U} = \boldsymbol{u}$, we assume the covariate $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_1(\boldsymbol{u}^\top \boldsymbol{\varphi}^*), \boldsymbol{\Sigma}(\boldsymbol{u}^\top \boldsymbol{\varphi}^*))$ if $Y = 1$ and $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_2(\boldsymbol{u}^\top \boldsymbol{\varphi}^*), \boldsymbol{\Sigma}(\boldsymbol{u}^\top \boldsymbol{\varphi}^*))$ if $Y = 0$. Then Bayes's discriminant direction is also a function of $\boldsymbol{u}^\top \boldsymbol{\varphi}^*$, that is $\theta_j^*(\boldsymbol{u}) = g_j^*(\boldsymbol{u}^\top \boldsymbol{\varphi}^*)$ for $j = 1, ..., p$, where $g_j^*(\cdot)$ is a smooth univariate function. Similar to (2.1), g_j^* s are defined as the solutions of the following least-square problem

$$\min_{g_j \in \mathcal{L}^2(\mathcal{P}), j=1,\dots,p} \mathbb{E}\left[\left(Z - \sum_{j=1}^p g_j(\mathbf{U}^\top \boldsymbol{\varphi}^*)(X_j - \mu_j(\mathbf{U}^\top \boldsymbol{\varphi}^*))\right)^2 \Bigg| \mathbf{U}\right].$$

If $\boldsymbol{\varphi}^*$ is known, we can approximate the function $g_j^*(\mathbf{U}^\top \boldsymbol{\varphi}^*)$ by $\boldsymbol{\gamma}_{(j)}^\top \boldsymbol{B}(\mathbf{U}^\top \boldsymbol{\varphi}^*)$. And the optimal approximation coefficients $(\widetilde{\boldsymbol{\gamma}}_{(1)}, ..., \widetilde{\boldsymbol{\gamma}}_{(p)})$ are defined as

$$\arg \min_{\boldsymbol{\gamma}_{(j)} \in \mathbb{R}^{L_n}, j=1,\dots,p} \mathbb{E}\left[\left(Z - \sum_{j=1}^p (X_j - \mu_j(\mathbf{U}^\top \boldsymbol{\varphi}^*)) \boldsymbol{\gamma}_{(j)}^\top \boldsymbol{B}(\mathbf{U}^\top \boldsymbol{\varphi}^*)\right)^2\right].$$

As for the initial estimator of $\boldsymbol{\varphi}$, according to our assumption, we can equivalently write the covariate $\boldsymbol{X}_i$ as the form of the standard single index model

$$\begin{aligned} \boldsymbol{X}_i &= \boldsymbol{\mu}_1(\boldsymbol{U}_i^\top \boldsymbol{\varphi}^*) + \Big(\boldsymbol{\Sigma}(\boldsymbol{U}_i^\top \boldsymbol{\varphi}^*)\Big)^{1/2} \boldsymbol{\epsilon}_i \quad \text{if} \quad Y_i = 1, \\ \boldsymbol{X}_i &= \boldsymbol{\mu}_2(\boldsymbol{U}_i^\top \boldsymbol{\varphi}^*) + \Big(\boldsymbol{\Sigma}(\boldsymbol{U}_i^\top \boldsymbol{\varphi}^*)\Big)^{1/2} \boldsymbol{\epsilon}_i \quad \text{if} \quad Y_i = 0, \end{aligned}$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. We may utilize the method proposed in [36] to obtain the estimator of $\boldsymbol{\varphi}^*$, denoted by $\widehat{\boldsymbol{\varphi}}$. By plugging in $\widehat{\boldsymbol{\varphi}}$, the estimators of univariate functions g_j^* s can be estimated by the B-spline procedure in our paper.

5. Numerical experiments

This section investigates the numerical performance of the proposed varying coefficient discriminant procedure. In our simulation study, we only consider the balanced case where the sample sizes of the two classes are equal.

The exposure variable U_i for $i = 1, 2, ..., 2n$ are generated independently from uniform distribution on $[0, 1]$ in the following experiments. After generating U_i, we sample the covariate $\boldsymbol{X}_i$ with $Y_i = 1$ from $\mathcal{N}(\boldsymbol{\mu}_1(U_i), \boldsymbol{\Sigma}(U_i))$ for $i = 1, ..., n$ and sample $\boldsymbol{X}_i$ with $Y_i = 0$ from $\mathcal{N}(\boldsymbol{\mu}_2(U_i), \boldsymbol{\Sigma}(U_i))$ for $i = n+1, ..., 2n$, where $\boldsymbol{\mu}_1(u) = \mathbf{0}$ and $\boldsymbol{\mu}_2(u) = \boldsymbol{\Sigma}(u)\boldsymbol{\beta}(u)$. Several combinations of $\boldsymbol{\beta}(u)$ and $\boldsymbol{\Sigma}(u)$ are considered in our simulation. Each entry of Bayes' discriminant direction function take values as:

- Direction 1: $\beta_j^{(1)}(u) = 1$ for $1 \le j \le p$ (or s);

- Direction 2: $\beta_j^{(2)}(u) = u$ for $1 \leq j \leq p$ (or s);
- Direction 3: $\beta_j^{(2)}(u) = \sin(4u)$ for $1 \leq j \leq p$ (or s);
- Direction 4: $\beta_j^{(4)}(u) = e^u$ for $1 \leq j \leq p$ (or s).

In high-dimensional case, we set $\beta_j(\cdot) = 0$ for $s+1 \leq j \leq p$. Three covariance matrices are considered in our simulations:

- Covariance matrix 1, $\sigma_{i,j}^{(1)}(u) = 0.5^{|i-j|}$, for $1 \leq i, j \leq p$;
- Covariance matrix 2. $\sigma_{i,j}^{(2)}(u) = u^{|i-j|}$, for $1 \leq i, j \leq p$;
- Covariance matrix 3. $\sigma_{i,j}^{(3)}(u) = u\mathbb{I}(i \neq j) + \mathbb{I}(i = j)$ for $1 \leq i, j \leq p$.

The combination of *Direction 1* and *Covariance matrix 1* is a classical static setting, where each entry of the mean vector and covariance matrix is a constant value. The other combinations are dynamic settings. We use the cubic spline in our simulation, and select the number of spline basis functions by 5-fold cross-validation. We compute the misclassification risk based on an independently generated test set with size 200.

5.1. Low-dimensional case

For low-dimensional case, the sample size of each class is fixed as $n = 100$ and the dimensionality p is varying from $\{5, 10, 20\}$. The proposed method in this paper (abbreviated as VCLDA) is deployed to the generated data. For comparison, we also conduct the following two classification rules:

1. Oracle: use the population Bayes' discriminant direction $\mathbf{\Sigma}(u)^{-1}(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))$ to conduct classification.
2. LDA: use the *static* estimators of mean vectors and covariance matrix, i.e., the sample means and sample covariance matrix, to compute discriminant direction.

We report the averaged misclassification risks computed from the test set in Table 1. The oracle classification rule is the most accurate among all procedures. In a static setting, we can see that LDA achieves nearly oracle performance. As we expected, the performance of LDA procedure degrades drastically in the dynamic case. Meanwhile, the misclassification risk of VCLDA is significantly lower than LDA, and very close to the oracle procedure in all dynamic settings.

5.2. High-dimensional case

In high-dimensional simulation, we fix the sample size of each class as $n = 100$ and consider the dimensionality $p = 100$ and $p = 200$. Moreover, the sparsity under each dimensionality varies in $\{5, 10, 20\}$. For comparison, we also conduct the oracle rule, the static LPD rule [4] and DLPD rule [21] in the test set. The misclassification risks and their standard errors under four discriminant direction functions are summarized in Table 2-5 respectively. Undoubtedly, the oracle

Table 1
Misclassification risk and its standard error (in parentheses) of each method in low-dimensional case.

p	$\boldsymbol{\Sigma}$	Orcale	VCLDA	LDA	Orcale	VCLDA	LDA
			$\boldsymbol{\beta}^{(1)}$			$\boldsymbol{\beta}^{(2)}$	
5	1	0.048	0.075(0.021)	0.050(0.016)	0.227	0.259(0.039)	0.255(0.032)
	2	0.055	0.078(0.021)	0.119(0.028)	0.221	0.249(0.034)	0.272(0.032)
	3	0.039	0.058(0.020)	0.093(0.023)	0.202	0.221(0.032)	0.245(0.030)
10	1	0.005	0.028(0.014)	0.006(0.006)	0.155	0.193(0.033)	0.197(0.029)
	2	0.014	0.038(0.015)	0.152(0.025)	0.163	0.198(0.029)	0.270(0.035)
	3	0.004	0.027(0.012)	0.104(0.023)	0.126	0.155(0.031)	0.212(0.026)
20	1	0.000	0.020(0.012)	0.000(0.001)	0.108	0.157(0.032)	0.166(0.028)
	2	0.002	0.041(0.019)	0.215(0.034)	0.125	0.182(0.034)	0.312(0.035)
	3	0.000	0.042(0.017)	0.117(0.025)	0.081	0.128(0.024)	0.222(0.029)
			$\boldsymbol{\beta}^{(3)}$			$\boldsymbol{\beta}^{(4)}$	
5	1	0.194	0.234(0.034)	0.317(0.034)	0.010	0.024(0.012)	0.029(0.013)
	2	0.192	0.234(0.033)	0.382(0.047)	0.025	0.040(0.015)	0.156(0.025)
	3	0.173	0.209(0.028)	0.351(0.047)	0.018	0.033(0.014)	0.137(0.023)
10	1	0.125	0.186(0.027)	0.281(0.038)	0.001	0.011(0.008)	0.016(0.009)
	2	0.117	0.183(0.030)	0.437(0.046)	0.007	0.027(0.014)	0.195(0.029)
	3	0.092	0.154(0.031)	0.353(0.051)	0.003	0.022(0.013)	0.157(0.023)
20	1	0.083	0.189(0.031)	0.286(0.039)	0.000	0.014(0.009)	0.011(0.008)
	2	0.077	0.200(0.037)	0.476(0.041)	0.001	0.041(0.023)	0.246(0.035)
	3	0.054	0.176(0.037)	0.391(0.059)	0.000	0.044(0.022)	0.172(0.029)

classification rule is the most accurate among all procedures. In a static setting, it can be seen that the DLPD rule almost achieves the same performance as the LPD rule in static settings (see Table 2). As we expected, the performance of the classical LDA procedure degrades drastically in the dynamic case, which performs like random guessing in a highly dynamic setting. Except for the static setting, we can see that the misclassification risk of our proposed VCLDA rule is significantly lower than the DLPD rule, especially for the setting with *Covariance matrix 2*. In addition, the results indicate that the performance of VCLDA is most close to the oracle procedure.

In fact, VCLDA fully uses the information that the discrimination direction varies with different values of U while the active set of the discrimination coefficient will not change in our simulation settings. The former leads to a lower misclassification risk than the static LPD rule, and the latter leads to better performance over the DLPD rule.

6. Real data analysis

Diffuse large B-cell lymphoma (DLBCL) is a heterogeneous disease with recognized variability in clinical outcome, genetic features, and cells of origin. It is of vital importance for precision medicine if we can predict DLBCL in advance.

Table 2
Misclassification risk and its standard error (in parenthesis) of each method under Direction 1 in high-dimensional case.

s	$\mathbf{\Sigma}$	Orcale	VCLDA	LPD	DLPD	Orcale	VCLDA	LPD	DLPD
		$p = 100$				$p = 200$			
5	1	0.048	0.076(0.019)	0.053(0.012)	0.053(0.015)	0.048	0.071(0.018)	0.057(0.016)	0.057(0.016)
	2	0.055	0.070(0.017)	0.195(0.154)	0.202(0.035)	0.056	0.067(0.017)	0.153(0.106)	0.134(0.040)
	3	0.039	0.060(0.017)	0.332(0.192)	0.085(0.018)	0.039	0.060(0.017)	0.150(0.058)	0.100(0.029)
10	1	0.005	0.015(0.009)	0.103(0.193)	0.101(0.187)	0.005	0.025(0.012)	0.007(0.006)	0.007(0.006)
	2	0.014	0.041(0.017)	0.152(0.028)	0.168(0.034)	0.014	0.043(0.019)	0.148(0.023)	0.160(0.030)
	3	0.004	0.012(0.009)	0.123(0.026)	0.040(0.019)	0.004	0.016(0.011)	0.128(0.021)	0.040(0.015)
20	1	0.000	0.009(0.006)	0.055(0.157)	0.055(0.157)	0.000	0.004(0.005)	0.000(0.001)	0.000(0.001)
	2	0.002	0.018(0.010)	0.208(0.067)	0.176(0.030)	0.002	0.014(0.009)	0.198(0.031)	0.164(0.026)
	3	0.000	0.009(0.007)	0.123(0.022)	0.026(0.013)	0.000	0.009(0.008)	0.125(0.021)	0.029(0.018)

Table 3
Misclassification risk and its standard error (in parentheses) of each method under Direction 2 in high-dimensional case.

s	$\mathbf{\Sigma}$	Orcale	VCLDA	LPD	DLPD	Orcale	VCLDA	LPD	DLPD
		$p = 100$				$p = 200$			
5	1	0.225	0.243(0.031)	0.371(0.109)	0.248(0.031)	0.227	0.244(0.032)	0.281(0.045)	0.300(0.042)
	2	0.217	0.252(0.028)	0.274(0.032)	0.338(0.046)	0.220	0.237(0.031)	0.280(0.042)	0.370(0.043)
	3	0.199	0.241(0.031)	0.268(0.030)	0.251(0.027)	0.204	0.232(0.031)	0.272(0.054)	0.239(0.032)
10	1	0.158	0.173(0.027)	0.204(0.029)	0.210(0.030)	0.160	0.189(0.028)	0.206(0.031)	0.210(0.031)
	2	0.164	0.208(0.031)	0.260(0.050)	0.347(0.033)	0.165	0.185(0.026)	0.258(0.026)	0.345(0.034)
	3	0.126	0.156(0.026)	0.212(0.025)	0.166(0.026)	0.127	0.140(0.024)	0.212(0.030)	0.174(0.028)
20	1	0.107	0.126(0.024)	0.182(0.032)	0.146(0.030)	0.108	0.132(0.025)	0.166(0.025)	0.146(0.027)
	2	0.126	0.184(0.028)	0.280(0.064)	0.336(0.028)	0.124	0.179(0.028)	0.270(0.027)	0.329(0.029)
	3	0.081	0.099(0.021)	0.207(0.056)	0.116(0.020)	0.081	0.093(0.020)	0.202(0.026)	0.124(0.023)

Using the data provided in [26], we establish the model to predict DLBCL according to the gene expression. It is mentioned in [26] that tumors had less frequent genetic abnormalities in younger patients. Thus, our proposed method VCLDA seems suitable for setting up the prediction model by setting *the age* as the exposure variable U.

The original data has 124 patients and 44972 gene expression levels. The binary response means whether a germinal center B-cell is normal or not, which is the significant signal of DLBCL. We screen out 150 gene expression levels to build a model according to the t test on the binary response. We conduct the following four procedures: LPD (exclude age as a covariate), LPD (include age as a covariate), DLPD (regard age as U), and VCLDA (regard age as U). We randomly choose ten patients as the test sample in each trial and regard the remaining samples as the training set to run the classification procedure. The average results of misclassification risks on the test sample over 100 trials are reported in Table 6. It shows that the contribution of U is negligible as a covariate in the static LPD rule. In contrast, it improves the classification accuracy tremendously as an exposure variable in the dynamic model.

Additionally, the active sets selected by the DLPD method under different ages are highly coincident. It means that the genes influencing DLBCL will not change significantly with age, which is also reasonable in the gene analysis. Two genes are excluded from the active set by the DLPD method during a

TABLE 4
Misclassification risk and its standard error (in parenthesis) of each method under Direction 3 in high-dimensional case.

s	$\mathbf{\Sigma}$	Orcale	VCLDA	LPD	DLPD	Orcale	VCLDA	LPD	DLPD
		$p = 100$				$p = 200$			
5	1	0.193	0.244(0.033)	0.337(0.055)	0.291(0.04)	0.194	0.221(0.030)	0.335(0.053)	0.286(0.031)
	2	0.193	0.211(0.029)	0.412(0.068)	0.302(0.032)	0.192	0.214(0.028)	0.395(0.068)	0.298(0.031)
	3	0.172	0.203(0.033)	0.334(0.048)	0.298(0.029)	0.172	0.209(0.028)	0.341(0.051)	0.291(0.032)
10	1	0.123	0.147(0.024)	0.280(0.038)	0.232(0.028)	0.125	0.152(0.025)	0.278(0.039)	0.237(0.027)
	2	0.119	0.148(0.026)	0.450(0.075)	0.293(0.05)	0.122	0.145(0.028)	0.442(0.075)	0.270(0.032)
	3	0.090	0.115(0.025)	0.282(0.046)	0.214(0.03)	0.091	0.126(0.025)	0.274(0.036)	0.226(0.035)
20	1	0.080	0.110(0.022)	0.249(0.036)	0.201(0.027)	0.084	0.109(0.021)	0.260(0.042)	0.201(0.031)
	2	0.080	0.114(0.023)	0.487(0.045)	0.234(0.027)	0.079	0.114(0.023)	0.498(0.020)	0.229(0.027)
	3	0.050	0.084(0.023)	0.218(0.037)	0.184(0.026)	0.052	0.071(0.019)	0.218(0.031)	0.221(0.033)

TABLE 5
Misclassification risk and its standard error (in parentheses) of each method under Direction 4 in high-dimensional case.

s	$\mathbf{\Sigma}$	Orcale	VCLDA	LPD	DLPD	Orcale	VCLDA	LPD	DLPD
		$p = 100$				$p = 200$			
5	1	0.010	0.018(0.009)	0.032(0.014)	0.015(0.010)	0.010	0.022(0.012)	0.033(0.012)	0.019(0.010)
	2	0.025	0.049(0.016)	0.174(0.087)	0.207(0.035)	0.025	0.035(0.013)	0.271(0.163)	0.177(0.029)
	3	0.018	0.037(0.014)	0.299(0.166)	0.056(0.021)	0.018	0.031(0.014)	0.175(0.021)	0.060(0.020)
10	1	0.001	0.010(0.007)	0.020(0.011)	0.005(0.009)	0.001	0.006(0.006)	0.020(0.011)	0.003(0.004)
	2	0.007	0.025(0.011)	0.408(0.141)	0.195(0.031)	0.006	0.020(0.011)	0.319(0.144)	0.174(0.028)
	3	0.003	0.022(0.012)	0.179(0.041)	0.033(0.014)	0.003	0.016(0.010)	0.186(0.060)	0.037(0.014)
20	1	0.000	0.007(0.006)	0.012(0.008)	0.000(0.001)	0.000	0.007(0.007)	0.011(0.008)	0.000(0.003)
	2	0.001	0.011(0.009)	0.426(0.116)	0.189(0.029)	0.001	0.015(0.009)	0.479(0.063)	0.179(0.027)
	3	0.000	0.008(0.005)	0.170(0.026)	0.027(0.016)	0.000	0.009(0.008)	0.178(0.022)	0.027(0.015)

very short age interval, which may be confusing and misleading to the relative researchers. Nearly all active genes selected by the DLPD method are also selected by the VCLDA method. Besides, as we find in coefficients estimated by VCLDA, most of the genes have a weak influence on DLBCL when U is small, which collaborates with the conclusion in [26] that tumors have less frequent genetic abnormalities in younger patients.

7. Discussion

This paper investigates the LDA model for dynamic data and proposes a new varying coefficient discriminant rule. The proposed classification procedure is more efficient than the dynamic linear programming rule [21]. We also establish the upper bounds for estimation error and uniform excess misclassification risk. The synthetic and real data experiments also demonstrate a better classification

TABLE 6
The average misclassification risk and its standard error of each method in DLBCL dataset.

Method	LPD (exclude age)	LPD (include age)	DLPD (U = age)	VCLDA (U = age)
Avg	0.432	0.432	0.192	0.171
SE	0.211	0.211	0.167	0.122

performance of our varying coefficient LDA method.

The Gaussian graphical model (GGM) is an essential formalism to infer dependence structures of contemporary data sets, whose structure is equivalent to the support of the precision matrix. Recently, [27] proposed the functional graphical model and assumed the covariate is a p-dimensional functional data. The authors proposed an estimator of the precision matrix function based on kernel smoothing and CLIME [5]. Therefore, studying the high-dimensional, varying coefficient GGM under a dynamic setting will be of great interest.

Appendix A: Preliminaries

A.1. Background of B-spline approximation

From now on, we will omit the argument in random vector $\boldsymbol{B}(U)$ and $\widetilde{\boldsymbol{B}}(U)$ and write $\boldsymbol{B}$ and $\widetilde{\boldsymbol{B}}$ respectively whenever the context is clear. We introduce the following facts about standard B-spline basis $\boldsymbol{B}^*(u) = (B_1^*(u), ..., B_{L_n}^*(u))^\top$ (see [11, 13]), which will be used in our proof:

1. For any $u \in [0,1]$, both $0 \leq \max_{1\leq k\leq L_n} B_k^*(u) \leq 1$ and $\sum_{k=1}^{L_n} B_k^*(u) = 1$ hold.
2. For any $\eta_k \in \mathbb{R}$, $k = 1, 2, ..., L_n$, we have

$$L_n^{-1}\sum_{k=1}^{L_n}\eta_k^2 \lesssim \int\left(\sum_{k=1}^{L_n}\eta_k B_k^*(w)\right)^2 dw \lesssim L_n^{-1}\sum_{k=1}^{L_n}\eta_k^2. \tag{A.1}$$

From the facts displayed above, for any $r \geq 1$, we also have

$$\mathbb{E}\left[|B_k^*(U)|^r\right] \asymp L_n^{-1},$$

and

$$\|\mathbb{E}[\boldsymbol{B}^*]\|_2 = \sup_{\|\boldsymbol{\nu}\|_2=1}|\mathbb{E}[\boldsymbol{\nu}^\top\boldsymbol{B}^*]| \leq \sup_{\|\boldsymbol{\nu}\|_2=1}\left(\mathbb{E}[\boldsymbol{\nu}^\top\boldsymbol{B}^*]^2\right)^{1/2} = O(L_n^{-1/2}).$$

Similarly, we can also obtain that

$$L_n^{-1} \lesssim \lambda_{\min}(\mathbb{E}[\boldsymbol{B}^*\boldsymbol{B}^{*\top}]) \leq \lambda_{\max}(\mathbb{E}[\boldsymbol{B}^*\boldsymbol{B}^{*\top}]) \lesssim L_n^{-1}.$$

Writing the scaled B-spline basis as $\boldsymbol{B}(u) = \sqrt{L_n}\boldsymbol{B}^*(u)$, we have

$$\|\mathbb{E}[\boldsymbol{B}]\|_2 = O(1),$$

and

$$\lambda_{\min}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]) = O(1), \quad \lambda_{\max}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]) = O(1).$$

A.2. Concentration inequality

The following concentration inequality will be used throughout the proof.

Lemma A.1 (Lemma 1, [4])**.** *Let $\xi_1, ..., \xi_n$ be independent random variables with mean 0. Suppose that there exists some $\phi > 0$ and s_n such that $\sum_{i=1}^n \mathbb{E}[\xi_i^2 e^{\phi|\xi_i|}] \leq s_n^2$. Then for $0 < x < s_n^2$,*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i \geq C_\phi s_n x\right) \leq \exp(-x^2)$$

where $C_\phi = \phi + \phi^{-1}$.

Next lemma gives the moment inequalities for normal random variable.

Lemma A.2. *Let $X \sim \mathcal{N}(0, \sigma^2)$, then for any $0 \leq \phi \leq \frac{1}{\sqrt{2}\sigma}$,*

$$\mathbb{E}[X^2 e^{\phi|X|}] \leq \frac{e}{\phi^2} \frac{1}{\sqrt{1 - 2\phi^2\sigma^2}};$$

and for any $\phi \geq 0$ and $k \geq 1$

$$\mathbb{E}[X^k e^{\phi|X|}] \leq e^{\frac{\phi^2\sigma^2}{2}} \left(\mathbb{E}[X^k_{-\phi\sigma^2}] + \mathbb{E}[X^k_{\phi\sigma^2}]\right) \quad \textit{holds for any } \phi \geq 0,$$

where $X_{-\phi\sigma^2} \sim \mathcal{N}(-\phi\sigma^2, \sigma^2)$ and $X_{\phi\sigma^2} \sim \mathcal{N}(\phi\sigma^2, \sigma^2)$.

Proof of Lemma A.2. Using the basic inequality $s^2 e^s \leq e^{2s}$ for any $s \geq 0$, we have

$$\begin{aligned}
\mathbb{E}[X^2 e^{\phi|X|}] &\leq \phi^{-2}\mathbb{E}[e^{2\phi|X|}] \\
&\leq \phi^{-2}\mathbb{E}\left[e^{1+\phi^2X^2}\right] \\
&= \frac{e}{\phi^2}\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\sigma^2}+\phi^2x^2} dx \\
&= \frac{e}{\phi^2}\frac{1}{\sqrt{1-2\phi^2\sigma^2}}.
\end{aligned}$$

In addition, we also have

$$\begin{aligned}
\mathbb{E}[X^k e^{\phi|X|}] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} x^k e^{-\frac{x^2}{2\sigma^2}+\phi|x|} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\left(\int_{-\infty}^{0} x^k e^{-\frac{x^2}{2\sigma^2}-\phi x} dx + \int_{0}^{+\infty} x^k e^{-\frac{x^2}{2\sigma^2}+\phi x} dx\right) \\
&\leq \frac{e^{\frac{\phi^2\sigma^2}{2}}}{\sqrt{2\pi}\sigma}\left(\int_{-\infty}^{+\infty} x^k e^{-\frac{(x+\phi\sigma^2)^2}{2\sigma^2}} dx + \int_{-\infty}^{+\infty} x^k e^{-\frac{(x-\phi\sigma^2)^2}{2\sigma^2}} dx\right).
\end{aligned}$$

□

A.3. Estimation error bound for mean functions

The estimation error bounds for mean function vectors in the following proposition contribute to establish the convergence rates of discriminant direction estimator and the excess misclassification risk.

Proposition A.1. *Denote the estimator of the mean functions by* $\widehat{\boldsymbol{\mu}}_1(u) = (\widehat{\boldsymbol{\alpha}}_{11}^\top \boldsymbol{B}(u), \cdots, \widehat{\boldsymbol{\alpha}}_{1p}^\top \boldsymbol{B}(u))^\top$ *and* $\widehat{\boldsymbol{\mu}}_2(u) = (\widehat{\boldsymbol{\alpha}}_{21}^\top \boldsymbol{B}(u), \cdots, \widehat{\boldsymbol{\alpha}}_{2p}^\top \boldsymbol{B}(u))^\top$*, then under condition* (**C**1)-(**C**4)*, for any* $\vartheta > 0$ *we have*

$$\sup_{u\in[0,1]} |\widehat{\boldsymbol{\mu}}_1(u) - \boldsymbol{\mu}_1(u)|_\infty \lesssim \sqrt{\frac{L_n \log n}{n}} + L_n^{-d},$$

and

$$\sup_{u\in[0,1]} |\widehat{\boldsymbol{\mu}}_2(u) - \boldsymbol{\mu}_2(u)|_\infty \lesssim \sqrt{\frac{L_n \log n}{n}} + L_n^{-d},$$

hold with probability at least $1 - 3pL_n n^{-\vartheta}$ *respectively.*

Lemma A.3. *For any* $\vartheta > 0$, *there exists some positive constant* C *such that*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\boldsymbol{B}_i^\top - \mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right\|_2 \geq CL_n\sqrt{\frac{\log n}{n}}\right) \leq n^{-\vartheta L_n}.$$

The proof of Lemma A.3 is deferred to Appendix C.1.

Remark A.1. *It is worthwhile noting that Lemma A.3 is more tight than the results Lemma A.7 in [13], where the authors established the following bound*

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i^*(\boldsymbol{B}_i^*)^\top - \mathbb{E}[\boldsymbol{B}^*(\boldsymbol{B}^*)^\top]\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{L_n \log n}{n}}\right).$$

By the fact that $\boldsymbol{B}_i = \sqrt{L_n}\boldsymbol{B}_i^*$ *and* $\boldsymbol{B} = \sqrt{L_n}\boldsymbol{B}^*$*, using the relation above, we can only obtain the following worse bound*

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i\boldsymbol{B}_i^\top - \mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right\|_2 = O_{\mathbb{P}}\left(L_n^{3/2}\sqrt{\frac{\log n}{n}}\right).$$

Proof of Proposition A.1. First we introduce the population form of the approximation coefficient,

$$\widetilde{\boldsymbol{\alpha}}_{1j} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{L_n}} \mathbb{E}\left(X_j - \boldsymbol{\alpha}^\top \boldsymbol{B}(U)\middle|Y=1\right)^2, \tag{A.2}$$

and denote $\widetilde{\boldsymbol{\mu}}_1(u) = (\widetilde{\boldsymbol{\alpha}}_{11}^\top \boldsymbol{B}, \cdots, \widetilde{\boldsymbol{\alpha}}_{1p}^\top \boldsymbol{B})^\top$. According to the splines' approximation property [11, 20], we are guaranteed that

$$\sup_{u\in[0,1]} |\widetilde{\boldsymbol{\mu}}_1(u) - \boldsymbol{\mu}_1(u)|_\infty \lesssim L_n^{-d}.$$

In addition, we have

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{1j}-\widetilde{\boldsymbol{\alpha}}_{1j}&=\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i(X_{ij}-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j})\right)\\&=\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[X_{ij}-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right).\end{aligned}\tag{A.3}$$

For any positive definite matrix $\mathbf{A}\in\mathbb{R}^{p\times p}$ and $\|\Delta\mathbf{A}\|_2=o(1)$,

$$\begin{aligned}\|(\mathbf{A}+\Delta\mathbf{A})^{-1}-\mathbf{A}^{-1}\|_2&\le\|\mathbf{A}^{-1}\|_2\|(\mathbf{I}+\mathbf{A}^{-1}\Delta\mathbf{A})^{-1}-\mathbf{I}\|_2\\&\le\|\mathbf{A}^{-1}\|_2\left(\|\mathbf{A}^{-1}\Delta\mathbf{A}\|_2+o(1)\right)\\&\le2\|\mathbf{A}^{-1}\|_2^2\|\Delta\mathbf{A}\|_2.\end{aligned}\tag{A.4}$$

Now let $\mathbf{A}=\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]$ and $\Delta\mathbf{A}=\sum_{i=1}^n\boldsymbol{B}_i\boldsymbol{B}_i^\top/n-\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]$. Since Lemma A.3 claims that $\|\Delta\mathbf{A}\|_2=o(1)$ almost surely, (A.4) results in

$$\begin{aligned}&\left\|\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}-\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]^{-1}\right\|_2\\&\qquad\le2\left\|\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]^{-1}\right\|_2^2\left\|\frac{1}{n}\sum_{i=1}^n\boldsymbol{B}_i\boldsymbol{B}_i^\top-\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right\|_2.\end{aligned}$$

In conjunction with Lemma A.3 and $\lambda_{\min}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top])\ge M_1$ we have

$$\left\|\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\right\|_2\le\frac{1}{M_1}+C\sqrt{\frac{L_n\log n}{n}}\tag{A.5}$$

holds with probability at least $1-L_nn^{-\vartheta}$. Substituting (A.5) into (A.3) yields

$$\|\widehat{\boldsymbol{\alpha}}_{1j}-\widetilde{\boldsymbol{\alpha}}_{1j}\|_2\lesssim\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[X_{ij}-\mu_{1j}(U_i)+\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right\|_2.\tag{A.6}$$

On the other hand, we note that given $Y_i=1$ and U_i, $X_{ij}\sim\mathcal{N}(\mu_{1j}(U_i),\boldsymbol{\Sigma}_{jj}(U_i))$. By choosing any $\eta>0$ and using Lemma A.2, we have

$$\begin{aligned}&\mathbb{E}\left[(B_k^*(U_i))^2\left(X_{ij}-\mu_{1j}(U_i)\right)^2\exp\left(\eta|B_k^*(U_i)||X_{ij}-\mu_{1j}(U_i)|\right)\right]\\\le&\mathbb{E}\left\{(B_k^*(U_i))^2\mathbb{E}\left[\left(X_{ij}-\mu_{1j}(U_i)\right)^2\exp\left(\eta|X_{ij}-\mu_{1j}(U_i)|\right)|U_i\right]\right\}\\\le&2\mathbb{E}\left[(B_k^*(U_i))^2e^{\frac{\eta^2\boldsymbol{\Sigma}_{jj}(U_i)}{2}}\left(\eta^2\boldsymbol{\Sigma}_{jj}^2(U_i)+\boldsymbol{\Sigma}_{jj}(U_i)\right)\right]\\\lesssim&\mathbb{E}[(B_k^*(U_i))^2]\le L_n^{-1},\end{aligned}$$

where the first inequality follows from $0 \leq B_k^*(U_i) \leq 1$ and $\Sigma_{jj}(U_i)$ is the conditional variance of X_{ij} given U_i. Applying Lemma A.1 and uniform bound, we can guarantee

$$\mathbb{P}\left(\max_{1\leq k\leq L_n}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}(X_{ij}-\mu_{1j}(U_i))B_k(U_i)\right| \lesssim \sqrt{\frac{\log n}{n}}\right) \leq L_n n^{-\vartheta}.$$

It yields that

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i(X_{ij}-\mu_{1j}(U_i))\right\|_2 \lesssim \sqrt{\frac{L_n\log n}{n}}, \tag{A.7}$$

holds with probability at least $1-2L_n n^{-\vartheta}$. In addition, we note that

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right\|_2 &\leq \left\|\mathbb{E}[\boldsymbol{B}[\mu_{1j}(U)-\boldsymbol{B}^\top\widetilde{\boldsymbol{\alpha}}_{1j}]]\right\|_2 \\ &+\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]-\mathbb{E}[\boldsymbol{B}_i[\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]]\right\|_2.\end{aligned}$$

Using Lemma A.1 again, we may show that

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]-\mathbb{E}[\boldsymbol{B}_i[\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]]\right\|_2 \lesssim L_n^{-d}\sqrt{\frac{L_n\log n}{n}},$$

holds with probability at least $1-L_n n^{-\vartheta}$. Combining (A.7) and the following fact

$$\|\mathbb{E}[\boldsymbol{B}[\mu_{1j}(U)-\boldsymbol{B}^\top\widetilde{\boldsymbol{\alpha}}_{1j}]]\|_2 \lesssim L_n^{-d}\|\mathbb{E}[\boldsymbol{B}]\|_2 \lesssim L_n^{-d},$$

we have with probability at least $1-3L_n n^{-\vartheta}$

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[X_{ij}-\mu_{1j}(U_i)+\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right\|_2 \lesssim \sqrt{\frac{L_n\log n}{n}}. \tag{A.8}$$

Together with (A.6), we have proved the first assertion.

For each fixed $u\in[0,1]$, we denote $\boldsymbol{\eta}(u)=\left(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right)^{-1}\boldsymbol{B}^*(u)$. It holds that

$$\begin{aligned}|\widehat{\boldsymbol{\mu}}_1(u)-\widetilde{\boldsymbol{\mu}}_1(u)|_\infty &= \max_j|\boldsymbol{B}(u)^\top(\hat{\boldsymbol{\alpha}}_{1j}-\widetilde{\boldsymbol{\alpha}}_{1j})| \\ &= \max_j\left|\boldsymbol{B}(u)^\top\left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i[X_{ij}-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right| \\ &\lesssim L_n\max_j\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\eta}(u)^\top\boldsymbol{B}_i^*[X_{ij}-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right|,\end{aligned}$$

where the last inequality comes from Lemma A.3 and (A.8). It follows from (A.1) that

$$\mathbb{E}\left[\left(\boldsymbol{\eta}(u)^\top \boldsymbol{B}_i^*\right)^2\right] = \mathbb{E}\left[\left(\sum_{k=1}^{L_n}\eta_k(u)B^*(U)\right)^2\right] \lesssim L_n^{-1}\sum_{k=1}^{L_n}\eta_k^2(u) \lesssim L_n^{-1}.$$

Further we have

$$\mathbb{E}\left[\left(\boldsymbol{\eta}(u)^\top \boldsymbol{B}_i^*\right)^2 (X_{ij}-\mu_{1j}(U_i))^2 e^{\eta|\boldsymbol{\eta}(u)^\top \boldsymbol{B}_i^*|X_{ij}-\mu_{1j}(U_i)||}\right] \lesssim L_n^{-1}.$$

Then applying Lemma A.1, we can show that

$$\max_j\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\eta}(u)^\top \boldsymbol{B}_i^*[X_{ij}-\mu_{1j}(U_i)+\mu_{1j}(U_i)-\boldsymbol{B}_i^\top\widetilde{\boldsymbol{\alpha}}_{1j}]\right| \lesssim \sqrt{\frac{\log n}{nL_n}}$$

holds with probability at least $1-pn^{-\vartheta}$. With the same probability, for any fixed u, it holds that

$$|\widehat{\boldsymbol{\mu}}_1(u)-\widetilde{\boldsymbol{\mu}}_1(u)|_\infty \lesssim \sqrt{\frac{L_n\log n}{n}}. \tag{A.9}$$

Next we use chaining technique to prove the uniform result. Notice that, we may divide the interval $[0,1]$ to n^M sub-intervals with end points $0=u_0\le u_1\le\cdots u_{n^M}=1$. Then for any $u\in[0,1]$, there exists some $0\le\ell\le n^M$ such that $|u-u_\ell|\le n^{-M}$. Thus we have

$$\begin{aligned}|\widehat{\boldsymbol{\mu}}_1(u)-\widetilde{\boldsymbol{\mu}}_1(u)|_\infty &\le |\widehat{\boldsymbol{\mu}}_1(u_\ell)-\widetilde{\boldsymbol{\mu}}_1(u_\ell)|_\infty\\ &\quad+\{|\widehat{\boldsymbol{\mu}}_1(u)-\widehat{\boldsymbol{\mu}}_1(u_\ell)|_\infty+|\widetilde{\boldsymbol{\mu}}_1(u)-\widetilde{\boldsymbol{\mu}}_1(u_\ell)|_\infty\}\\ &\lesssim |\widehat{\boldsymbol{\mu}}_1(u_\ell)-\widetilde{\boldsymbol{\mu}}_1(u_\ell)|_\infty+n^{-M},\end{aligned}$$

where we used both $\widehat{\boldsymbol{\mu}}_1$ and $\widetilde{\boldsymbol{\mu}}$ are Lipschitz continuous. It follows that

$$\sup_{u\in[0,1]}|\widehat{\boldsymbol{\mu}}_1(u)-\widetilde{\boldsymbol{\mu}}_1(u)|_\infty \lesssim \max_{0\le\ell\le n^M}|\widehat{\boldsymbol{\mu}}_1(u_\ell)-\widetilde{\boldsymbol{\mu}}_1(u_\ell)|_\infty+n^{-M}.$$

By choosing $M=1$ and large ϑ, the second assertion follows from (A.9) immediately. □

Appendix B: Proofs of main results

B.1. *Proof of Proposition 2.1*

The proof is adapted from [6], here we provide it for completeness.

Lemma B.1 (Lemma 7, [6])**.** *For two vectors $\boldsymbol{\theta}_n$ and $\hat{\boldsymbol{\theta}}_n$, if $\|\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n\|_2=o(1)$ as $n\to\infty$, and $\|\boldsymbol{\theta}\|_2\ge c$ for some constant c, then when $n\to\infty$,*

$$\|\boldsymbol{\theta}_n\|_2\|\hat{\boldsymbol{\theta}}_n\|_2-\boldsymbol{\theta}_n^\top\hat{\boldsymbol{\theta}}_n \asymp \|\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n\|_2^2.$$

Proof. Let $\boldsymbol{\delta}(u) = \boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u)$ and $\widehat{\Delta}(u) = c^*(u)(\widehat{\boldsymbol{\theta}}(u)^\top \boldsymbol{\Sigma}(u)\widehat{\boldsymbol{\theta}}(u))^{1/2}$ for $u \in [0,1]$. Recall the relation $\boldsymbol{\beta}^*(u) = \boldsymbol{\Sigma}^{-1}(u)\boldsymbol{\delta}(u) = c^*(u)\boldsymbol{\theta}^*(u)$ for $c^*(u) \in (0,1)$. To simplify notations, we will use f to denote $f(u)$ for any function of u. Following the proof technique in [6], we define a intermediate quantity

$$\widetilde{R}(u) = \frac{1}{2}\Phi\left(-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right) + \frac{1}{2}\bar{\Phi}\left(\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right).$$

Note that, $c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}} = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})$, then using Lemma B.1

$$\begin{aligned}\left|\Delta - \frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}}{\widehat{\Delta}}\right| &= \left|\|\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1/2}\|_2 - \frac{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})}{\left\|\boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})\right\|_2}\right| \\ &\lesssim \frac{\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta} - \boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})\right\|_2^2}{\Delta}. \end{aligned} \tag{B.1}$$

Using the fact that $c^*\boldsymbol{\theta}^* = \boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ and $\|\boldsymbol{\Sigma}\|_2$ is bounded,

$$\begin{aligned}\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta} - \boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})\right\|_2^2 &\le \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta} - \boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})\right\|_2^2 \\ &= \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta} - \boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}} - c^*\boldsymbol{\theta}^*) - \boldsymbol{\Sigma}^{1/2}c^*\boldsymbol{\theta}^*\right\|_2^2 \\ &\le (c^*)^2\|\boldsymbol{\Sigma}^{1/2}\|_2^2\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2. \end{aligned} \tag{B.2}$$

Then take Taylor expansion to the two terrms of $\widetilde{R}$ around $-\Delta/2$ and $\Delta/2$ respectively, we have

$$\begin{aligned}\widetilde{R} - R &= \frac{1}{2}\left(-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}} + \frac{\Delta}{2}\right)\Phi'\left(-\frac{\Delta}{2}\right) + \frac{1}{2}\left(\frac{\Delta}{2} - \frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)\Phi'\left(\frac{\Delta}{2}\right) \\ &\quad + \frac{1}{4}\left((\Phi''(b_{1n}) + \Phi''(b_{2n}))\left(\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}} - \frac{\Delta}{2}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\left(\Delta - \frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}}{\widehat{\Delta}}\right)\exp\left(-\frac{\Delta^2}{8}\right) \\ &\quad + \frac{1}{4}\left((\Phi''(b_{1n}) + \Phi''(b_{2n}))\left(\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}} - \frac{\Delta}{2}\right)^2\right),\end{aligned}$$

where b_{1n} is some point between $-\frac{\Delta}{2}$ and $-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$ and b_{2n} is some point between $\frac{\Delta}{2}$ and $\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$. Hence it holds that

$$\Phi''(b_{1n}) \asymp \Phi''(b_{2n}) \asymp \frac{\Delta}{2}\exp\left(-\frac{\Delta^2}{8}\right).$$

Together with (B.1) and (B.2), we can obtain

$$\begin{aligned}|\widetilde{R}-R| &\lesssim \left|\Delta-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}}{\widehat{\Delta}}\right|+\left|\Delta-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}}{\widehat{\Delta}}\right|^2\\ &\lesssim \left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\delta}-\boldsymbol{\Sigma}^{1/2}(c^*\widehat{\boldsymbol{\theta}})\right\|_2^2\\ &\lesssim \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2. \end{aligned} \tag{B.3}$$

Next we bound $|R_n-\widetilde{R}|$, note that

$$\begin{aligned}&\left|\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}-\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_2)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}\right|\\ &\quad=\frac{\left|(\boldsymbol{\delta}/2-\widehat{\boldsymbol{\mu}}+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}+(\boldsymbol{\delta}/2-\widehat{\boldsymbol{\mu}}+\boldsymbol{\mu}_2)^\top(c^*\widehat{\boldsymbol{\theta}}-c^*\boldsymbol{\theta}^*)\right|}{\widehat{\Delta}}\\ &\quad\lesssim\frac{1}{\Delta}\Big(|(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{\beta}^*|+\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2\|\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_2\Big)\\ &\quad\lesssim\frac{|(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{\beta}^*|}{\Delta},\end{aligned}$$

where the first inequality follows from that Δ is bounded. Take Taylor expansion on the two terms of $R_n(t)$ around $-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$ and $\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$ respectively, we have

$$\begin{aligned}R_n(t)-\widetilde{R}&=\frac{1}{2}\left(\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_1)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}+\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)\Phi'\left(-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)\\ &\quad-\frac{1}{2}\left(\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_2)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)\Phi'\left(\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)\\ &\quad+\frac{1}{4}\left(\Phi''(b_{3n})\left(\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_1)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}+\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)^2\right)\\ &\quad+\frac{1}{4}\left(\Phi''(b_{4n})\left(\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_2)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}\right)^2\right)\\ &=\frac{\Phi''(b_{3n})+\Phi''(b_{4n})}{4}\frac{(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}},\end{aligned} \tag{B.4}$$

where b_{3n} is some point between $\frac{\left(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_1\right)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}$ and $\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$ and b_{4n} is some point between $\frac{\left(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}_1\right)^\top(c^*\widehat{\boldsymbol{\theta}})}{\widehat{\Delta}}$ and $-\frac{c^*\boldsymbol{\delta}^\top\widehat{\boldsymbol{\theta}}/2}{\widehat{\Delta}}$, then

$$\Phi''(b_{3n})\asymp\Phi''(b_{4n})\asymp\frac{\Delta}{2}\exp\left(-\frac{\Delta^2}{8}\right).$$

In fact, we also used $\boldsymbol{\mu}-\boldsymbol{\mu}_1 = -\boldsymbol{\delta}/2$ and $\boldsymbol{\mu}-\boldsymbol{\mu}_2 = \delta/2$ to obtain (B.4). Then (B.4) implies

$$|R_n(t) - \widetilde{R}| \lesssim \left|(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top (c^* \boldsymbol{\theta}^*)\right|^2 + \left|(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right|^2 \\ \lesssim |\,(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{\beta}^*|^2. \tag{B.5}$$

Combining (B.3) and (B.5), for any $u \in [0,1]$, it holds that

$$|R_n(u) - R(u)| \leq |R_n(u) - \widetilde{R}(u)| + |\widetilde{R}(u) - R(u)| \\ \lesssim \|\widehat{\boldsymbol{\theta}}(u) - \boldsymbol{\theta}^*(u)\|_2^2 + \left|(\widehat{\boldsymbol{\mu}}(u) - \boldsymbol{\mu}(u))^\top \boldsymbol{\beta}^*(u)\right|^2. \qquad \square$$

B.2. Proof of Theorem 3.1 and 3.3

Here we only prove Theorem 3.1, and the proof of Theorem 3.1 can be easily obtained through the similar analysis. The following lemma provides the lower bound and upper bound for the eigenvalues of $\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]$, and the proof is deferred to Appendix C.2.

Lemma B.2. *Assume the assumptions hold, then there exist two positive constant M_1 and M_2 such that*

$$M_1\lambda_0 \leq \lambda_{\min}(\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]) \leq \lambda_{\max}(\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]) \leq M_2(\lambda_1 + \delta_p/4).$$

Proof of Theorem 3.1. There exists $\bar{\theta}_j(U) = \bar{\boldsymbol{\gamma}}_j^\top \boldsymbol{B}(U)$ for $j = 0, 1, ..., p$ such that

$$\sup_{u\in[0,1]} |\theta_j^*(u) - \bar{\theta}_j(u)| \leq M_0 L_n^{-d}. \tag{B.6}$$

Let $\bar{\boldsymbol{\gamma}} = (\bar{\boldsymbol{\gamma}}_1^\top, \cdots, \bar{\boldsymbol{\gamma}}_p^\top)^\top$, then note that

$$\widetilde{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}} = \left[\mathbb{E}\left(\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top\right)\right]^{-1} \mathbb{E}\left[\widetilde{\boldsymbol{B}}\left(Z - \widetilde{\boldsymbol{B}}^\top \bar{\boldsymbol{\gamma}}\right)\right], \tag{B.7}$$

The optimality condition of $\theta_j^*(U)$ implies that

$$\mathbb{E}\left[(X_j - \mu_j(U))\left(Z - \sum_{l=1}^p (X_j - \mu_j(U))\theta_j^*(U)\right)\Big| U\right] = 0,$$

which means

$$\mathbb{E}\left[(X_j - \mu_j(U))\boldsymbol{B}(U)\left(Z - \sum_{l=1}^p (X_j - \mu_j(U))\theta_j^*(U)\right)\right] = \mathbf{0}.$$

Recall $\widetilde{\boldsymbol{B}} = (\boldsymbol{X} - \boldsymbol{\mu}) \otimes \boldsymbol{B}$, then we can get

$$\mathbb{E}\left[\widetilde{\boldsymbol{B}}\left(Z - \widetilde{\boldsymbol{B}}^\top \bar{\gamma}\right)\right] = \mathbb{E}\left[\widetilde{\boldsymbol{B}}\left(\sum_{j=1}^p \theta_j^*(U)(X_j - \mu_j(U)) - \sum_{j=1}^p \bar{\theta}_j(U)(X_j - \mu_j(U))\right)\right]$$

$$= \mathbb{E}\left[\widetilde{\boldsymbol{B}}(\boldsymbol{X}-\boldsymbol{\mu})^\top(\boldsymbol{\theta}^*(U)-\bar{\boldsymbol{\theta}}(U))\right].$$

Let $\mathbf{C}(U) = \mathbb{E}\left[(\boldsymbol{X}-\boldsymbol{\mu}(U))(\boldsymbol{X}-\boldsymbol{\mu}(U))^\top|U\right]$, then simple calculation yields that

$$\mathbf{C}(U) = \boldsymbol{\Sigma}(U) + \frac{1}{4}(\boldsymbol{\mu}_1(U)-\boldsymbol{\mu}_2(U))(\boldsymbol{\mu}_1(U)-\boldsymbol{\mu}_2(U))^\top.$$

For $\boldsymbol{\nu} = (\boldsymbol{\nu}_{(1)}^\top, ..., \boldsymbol{\nu}_{(p)}^\top)^\top$, we denote $\widetilde{\boldsymbol{\nu}}(U) = (\boldsymbol{\nu}_{(1)}^\top\boldsymbol{B}(U), ..., \boldsymbol{\nu}_{(p)}^\top\boldsymbol{B}(U))^\top$. Then we have

$$\begin{aligned}
&\left\|\mathbb{E}[\widetilde{\boldsymbol{B}}(Z-\widetilde{\boldsymbol{B}}^\top\bar{\gamma})]\right\|_2 \\
&= \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\mathbb{E}[\boldsymbol{\nu}^\top\widetilde{\boldsymbol{B}}(Z-\widetilde{\boldsymbol{B}}^\top\bar{\gamma})]\right| \\
&= \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\mathbb{E}[\widetilde{\boldsymbol{\nu}}(U)^\top(\boldsymbol{X}-\boldsymbol{\mu}(U))(\boldsymbol{X}-\boldsymbol{\mu}(U))^\top(\boldsymbol{\theta}^*(U)-\bar{\boldsymbol{\theta}}(U))]\right| \\
&= \sup_{\|\boldsymbol{\nu}\|_2=1}\left|\mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}(U)\|_2\|\mathbf{C}(U)\|_2\|\boldsymbol{\theta}^*(U)-\bar{\boldsymbol{\theta}}(U)\|_2\right]\right| \\
&\lesssim \sqrt{p}L_n^{-d}\sup_{\|\boldsymbol{\nu}\|_2=1}\mathbb{E}[\|\widetilde{\boldsymbol{\nu}}(U)\|_2], \qquad \text{(B.8)}
\end{aligned}$$

where the last inequality follows from (B.6) and $\|\mathbf{C}(u)\|_2 \le \|\boldsymbol{\Sigma}(u)\|_2+\delta_p$. Using the inequality (A.1) and $\|\boldsymbol{\nu}\|_2 = 1$, we have

$$\mathbb{E}[\|\widetilde{\boldsymbol{\nu}}\|_2] \le \left(\mathbb{E}[\|\widetilde{\boldsymbol{\nu}}\|_2^2]\right)^{1/2} = L_n\left(\sum_{j=1}^p\mathbb{E}[(\boldsymbol{\nu}_{(j)}^\top\boldsymbol{B}^*)^2]\right)^{1/2} \lesssim \left(\sum_{j=1}^p\|\boldsymbol{\nu}_{(j)}\|_2^2\right)^{1/2}.$$

Combining (B.7) and (B.8), we are guaranteed that $\|\widetilde{\gamma}-\bar{\gamma}\|_2 \lesssim \sqrt{p}L_n^{-d}$. Recall that $\widetilde{\boldsymbol{\theta}}_j(u) = \widetilde{\gamma}_j^\top\boldsymbol{B}(u)$, together with (A.1), we can have

$$\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}\|_{\mathrm{L}_2}^2 = \int_0^1\left\|\widetilde{\boldsymbol{\theta}}(u)-\bar{\boldsymbol{\theta}}(u)\right\|_2^2du &= L_n\sum_{j=1}^p\int_0^1\left(\left(\widetilde{\gamma}_{(j)}-\bar{\boldsymbol{\theta}}_{(j)}\right)^\top\boldsymbol{B}^*(u)\right)^2du \\
&\lesssim \sum_{j=1}^p\|\widetilde{\gamma}_{(j)}-\bar{\boldsymbol{\theta}}_{(j)}\|_2^2 \\
&= \|\widetilde{\gamma}-\bar{\gamma}\|_2^2.
\end{aligned}$$

It yields that

$$\|\boldsymbol{\theta}^*-\widetilde{\boldsymbol{\theta}}\|_{\mathrm{L}_2} \le \|\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}}\|_{\mathrm{L}_2} + \|\widetilde{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}\|_{\mathrm{L}_2} \lesssim \sqrt{p}L_n^{-d}. \qquad \square$$

B.3. Proof of Theorem 3.2

Let $\mathbf{D}_n = \frac{1}{2n}\sum_{i=1}^{2n}\widetilde{\boldsymbol{B}}_i\widetilde{\boldsymbol{B}}_i^\top$ and $\boldsymbol{b}_n = \frac{1}{2n}\sum_{i=1}^{2n}\widetilde{\boldsymbol{B}}_iZ_i$. Correspondingly, we write $\mathbf{D} = \mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]$ and $\boldsymbol{b} = \mathbb{E}[\widetilde{\boldsymbol{B}}Z]$. The following two lemmas give the concentration bounds for two terms in estimation error $\|\widehat{\gamma}-\widetilde{\gamma}\|_2$. We defer the proofs in Section C.

Lemma B.3. *Under the conditions of Theorem 3.2, we have*

$$\|\mathbf{D}_n - \mathbf{D}\|_2 \lesssim L_n \sqrt{\frac{p \log n}{n}} + p L_n^{3/2} a_n \sqrt{\frac{\log n}{n}}, \tag{B.9}$$

holds with probability at least $1 - n^{-\vartheta L_n p} - pL_n n^{-\vartheta}$.

Lemma B.4. *Under the conditions of Theorem 3.2, we have*

$$\|\mathbf{D}_n \widetilde{\boldsymbol{\gamma}} - \mathbf{D}\widetilde{\boldsymbol{\gamma}}\|_2 \lesssim \sqrt{\frac{pL_n \log n}{n}} + pL_n a_n \sqrt{\frac{\log n}{n}}, \tag{B.10}$$

holds with probability at least $1 - n^{-\vartheta L_n p} - pL_n n^{-\vartheta}$.

Lemma B.5. *Under the conditions of Theorem 3.2, we have*

$$|\boldsymbol{b}_n - \boldsymbol{b}|_\infty \lesssim \sqrt{\frac{\log n}{n}} + L_n^{-\frac{1}{2}} a_n, \tag{B.11}$$

holds with probability $1 - pL_n n^{-\vartheta}$.

Proof of Theorem 3.2. From the definition of $\widehat{\boldsymbol{\gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, we have

$$\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}} = \mathbf{D}_n^{-1}\boldsymbol{b}_n - \widetilde{\boldsymbol{\gamma}} = \mathbf{D}_n^{-1}\left(\boldsymbol{b}_n - \mathbf{D}_n\widetilde{\boldsymbol{\gamma}}\right). \tag{B.12}$$

Now let us recall the optimal condition of $\widetilde{\boldsymbol{\gamma}}$,

$$\mathbf{0} = \mathbb{E}\left[\widetilde{\boldsymbol{B}}\left(Z - \widetilde{\boldsymbol{B}}^\top \widetilde{\boldsymbol{\gamma}}\right)\right] = \boldsymbol{b} - \mathbf{D}\widetilde{\boldsymbol{\gamma}}. \tag{B.13}$$

In addition, notice that

$$\begin{aligned}
\boldsymbol{b} &= \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}(U)) \otimes \boldsymbol{B} Z\right] \\
&= \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}(U)) \otimes \boldsymbol{B}\right] - \frac{1}{2}\mathbb{E}\left[(\boldsymbol{\mu}_2(U) - \boldsymbol{\mu}(U)) \otimes \boldsymbol{B}\right] \\
&= \mathbb{E}\left[(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U)) \otimes \boldsymbol{B}\right].
\end{aligned}$$

For $\boldsymbol{\nu} = (\boldsymbol{\nu}_{(1)}^\top, ..., \boldsymbol{\nu}_{(p)}^\top)^\top \in \mathbb{R}^{pL_n}$, we denote $\widetilde{\boldsymbol{\nu}}(U) = (\boldsymbol{\nu}_{(1)}^\top \boldsymbol{B}(U), ..., \boldsymbol{\nu}_{(p)}^\top \boldsymbol{B}(U))^\top$. By the definition of $\|\cdot\|_2$ and (A.1), we have

$$\begin{aligned}
\|\boldsymbol{b}\|_2 &= \sup_{\boldsymbol{\nu} \in \mathbb{S}^{pL_n - 1}} \left|\boldsymbol{\nu}^\top \boldsymbol{b}\right| \\
&= \sup_{\boldsymbol{\nu} \in \mathbb{S}^{pL_n - 1}} \left|\mathbb{E}\left[\boldsymbol{\nu}^\top \left((\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U)) \otimes \boldsymbol{B}\right)\right]\right| \\
&= \sup_{\boldsymbol{\nu} \in \mathbb{S}^{pL_n - 1}} \left|\mathbb{E}\left[\widetilde{\boldsymbol{\nu}}(U)^\top \left(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U)\right)\right]\right| \\
&\le \delta_p \sup_{\boldsymbol{\nu} \in \mathbb{S}^{pL_n - 1}} \mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}(U)\|_2\right] \\
&\le \delta_p \sup_{\boldsymbol{\nu} \in \mathbb{S}^{pL_n - 1}} \left(\mathbb{E}[\|\widetilde{\boldsymbol{\nu}}(U)\|_2^2]\right)^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&= \delta_p \sup_{\boldsymbol{\nu}\in\mathbb{S}^{pL_n-1}} \left(\sum_{j=1}^{p} \mathbb{E}\left[(\boldsymbol{\nu}_{(j)}^\top \boldsymbol{B}(U))^2\right]\right)^{1/2} \\
&\lesssim \delta_p \sup_{\boldsymbol{\nu}\in\mathbb{S}^{pL_n-1}} \left(\sum_{j=1}^{p} \|\boldsymbol{\nu}_{(j)}\|_2^2\right)^{1/2} = \delta_p, \qquad \text{(B.14)}
\end{aligned}
$$

where $\delta_p = \sup_{u\in[0,1]} \|\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u)\|_2$. According to Lemma B.3, we know that $\|\mathbf{D}_n - \mathbf{D}\|_2 = o_{\mathbb{P}}(1)$. Using the inequality (A.4), we get

$$
\begin{aligned}
\left\|\mathbf{D}_n^{-1} - \mathbf{D}^{-1}\right\|_2 &\le 2\left\|\mathbf{D}^{-1}\right\|_2 \|\mathbf{D}_n - \mathbf{D}\|_2 \\
&\le 2M_1^{-1}\lambda_0^{-1} \|\mathbf{D}_n - \mathbf{D}\|_2\,,
\end{aligned}
$$

where the second inequality follows from Lemma B.2. Hence we can guarantee that $\|\mathbf{D}_n^{-1}\|_2 = O(1)$ with high probability. By plugging the bounds in Lemma B.4 and B.5, together with (B.13), we have

$$
\begin{aligned}
\|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_2 &\le \|\mathbf{D}_n^{-1}\|_2 \|\boldsymbol{b}_n - \mathbf{D}_n\widetilde{\boldsymbol{\gamma}}\|_2 \\
&= \|\mathbf{D}_n^{-1}\|_2 \|\boldsymbol{b}_n - \mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \boldsymbol{b} + \mathbf{D}\widetilde{\boldsymbol{\gamma}}\|_2 \\
&\lesssim \|\boldsymbol{b}_n - \boldsymbol{b}\|_2 + \|\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \mathbf{D}\widetilde{\boldsymbol{\gamma}}\|_2 \\
&\lesssim \sqrt{\frac{pL_n \log n}{n}} + a_n p L_n \sqrt{\frac{\log n}{n}}. \qquad \text{(B.15)}
\end{aligned}
$$

Recall $\widehat{\boldsymbol{\theta}} = (\boldsymbol{B}(u)^\top \widehat{\boldsymbol{\gamma}}_{(1)}, ..., \boldsymbol{B}(u)^\top \widehat{\boldsymbol{\gamma}}_{(p)})^\top$ and $\widetilde{\boldsymbol{\theta}}(u) = (\boldsymbol{B}(u)^\top \widetilde{\boldsymbol{\gamma}}_{(1)}, ..., \boldsymbol{B}(u)^\top \widetilde{\boldsymbol{\gamma}}_{(p)})^\top$. Applying (A.1), we have

$$
\begin{aligned}
\int_0^1 \|\widehat{\boldsymbol{\theta}}(u) - \widetilde{\boldsymbol{\theta}}(u)\|_2^2 du &= L_n \sum_{j=1}^{p} \int_0^1 \left(\boldsymbol{B}^*(u)^\top (\widehat{\boldsymbol{\gamma}}_{(j)} - \widetilde{\boldsymbol{\gamma}}_{(j)})\right)^2 du \\
&\lesssim \sum_{j=1}^{p} \|\widehat{\boldsymbol{\gamma}}_{(j)} - \widetilde{\boldsymbol{\gamma}}_{(j)}\|_2^2 = \|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_2^2.
\end{aligned}
$$

Then we have finished the proof of Theorem 3.2 by plugging (B.15). □

B.4. Proof of Theorem 3.4

The following lemma provides the ℓ_2 error bound for general quadratic group lasso problem. We defer the proof of Lemma B.6 to Appendix C.6.

Lemma B.6. *For general quadratic group lasso problem*

$$
\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{pL_n}} \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{A}\boldsymbol{\gamma} - \boldsymbol{b}^\top\boldsymbol{\gamma} + \lambda \sum_{j=1}^{p} \|\boldsymbol{\gamma}_j\|_2,
$$

if the following two conditions hold

1. $\mathbf{A}$ *satisfies the restrictive eigenvalue condition with parameter* ζ*: for any* $\boldsymbol{\xi} \in \mathbb{R}^{pL_n}$ *such that* $\|\boldsymbol{\xi}\|_1 \leq 4\sqrt{sL_n}\|\boldsymbol{\xi}\|_2$*, it holds that*

$$\boldsymbol{\xi}^\top \mathbf{A}\boldsymbol{\xi} \geq \zeta\|\boldsymbol{\xi}\|_2.$$

2. *for any* $\check{\boldsymbol{\gamma}} \in \mathbb{R}^{pL_n}$ *such that* $\check{\boldsymbol{\gamma}}_{(j)} = \mathbf{0}$ *for* $j \in S^c$ *and*

$$\max_{1\leq j\leq p} \|(\mathbf{A}\check{\boldsymbol{\gamma}} - \boldsymbol{b})_{(j)}\|_2 \leq \frac{\lambda}{2}. \tag{B.16}$$

then we have

$$\|\widehat{\boldsymbol{\gamma}} - \check{\boldsymbol{\gamma}}\|_2 \leq \frac{12\sqrt{s}\lambda}{\zeta} \quad \textit{and} \quad \|\widehat{\boldsymbol{\gamma}} - \check{\boldsymbol{\gamma}}\|_1 \leq \frac{48s\sqrt{L_n}\lambda}{\zeta}.$$

Lemma B.7. *Under conditions* (**C**1)-(**C**5)*, let* $\boldsymbol{\nu} \in \mathbb{R}^{pL_n}$ *be a fixed vector with* $\boldsymbol{\nu}_{(S^c)} = \mathbf{0}$*, then for any* $\vartheta > 0$ *we have*

$$\max_{1\leq j\leq p} \|(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \mathbf{D}\widetilde{\boldsymbol{\gamma}})_{(j)}\|_2 \lesssim \|\boldsymbol{\nu}\|_2 \left(\sqrt{\frac{L_n \log p}{n}} + a_n L_n s\sqrt{\frac{\log p}{n}}\right),$$

holds with probability at least $1 - L_n p^{-\vartheta} - L_n s p^{-\vartheta} - s p^{-\vartheta L_n}$.

Proof of Theorem 3.4. According to Lemma B.6, it suffices to show the restrictive eigenvalue condition of $\mathbf{D}_n$ and the inequality (B.16). For any $\boldsymbol{\xi} \in \mathbb{R}^{pL_n}$ such that $\|\boldsymbol{\xi}\|_1 \leq 4\sqrt{sL_n}\|\boldsymbol{\xi}\|_2$, we have

$$\begin{aligned}
\boldsymbol{\xi}^\top \mathbf{D}_n\boldsymbol{\xi} &= \boldsymbol{\xi}^\top \mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]\boldsymbol{\xi} + \boldsymbol{\xi}^\top(\mathbf{D}_n - \mathbf{D})\boldsymbol{\xi} \\
&\geq M_1\lambda_0\|\boldsymbol{\xi}\|_2^2 - \|\boldsymbol{\xi}\|_1|(\mathbf{D}_n - \mathbf{D})\boldsymbol{\xi}|_\infty \\
&\geq M_1\lambda_0\|\boldsymbol{\xi}\|_2^2 - \|\boldsymbol{\xi}\|_1^2\left|\mathbf{D}_n - \mathbf{D}\right|_\infty \\
&\geq \left(M_1\lambda_0 - 4sL_n\left|\mathbf{D}_n - \mathbf{D}\right|_\infty\right)\|\boldsymbol{\xi}\|_2^2.
\end{aligned}$$

By tracing the proof of Lemma B.7, we can guarantee $sL_n|\mathbf{D}_n - \mathbf{D}|_\infty = o_{\mathbb{P}}(1)$. It implies that there exists some positive constant ζ such that,

$$\boldsymbol{\xi}^\top \mathbf{D}_n\boldsymbol{\xi} \geq \zeta\|\boldsymbol{\xi}\|_2^2.$$

Hence we have verified the restrictive eigenvalue condition. Recall the approximation coefficient $\widetilde{\boldsymbol{\gamma}}_{(S)} = \mathbf{D}_{(SS)}^{-1}\boldsymbol{b}_{(S)}$ and $\widetilde{\boldsymbol{\gamma}}_{(S^c)} = \mathbf{0}$, then we have

$$(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \boldsymbol{b}_n)_{(S)} = (\mathbf{D}_n)_{(SS)}\mathbf{D}_{(SS)}^{-1}\boldsymbol{b}_{(S)} - (\boldsymbol{b}_n)_{(S)}, \tag{B.17}$$

and

$$(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \boldsymbol{b}_n)_{(S^c)} = (\mathbf{D}_n)_{(S^cS)}\mathbf{D}_{(SS)}^{-1}\boldsymbol{b}_S - (\boldsymbol{b}_n)_{(S^c)}. \tag{B.18}$$

In addition, similar to (B.14), we can verify $\|\boldsymbol{b}_{(S)}\|_2 \lesssim \delta_s$. Together with Lemma B.2, we have

$$\|\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2 \leq \|\mathbf{D}_{(SS)}^{-1}\|_2\|\boldsymbol{b}_{(S)}\|_2 \lesssim \delta_s.$$

Then from (B.17), Lemma B.7 and B.5, for any $j \in S$

$$\begin{aligned}\|(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \boldsymbol{b}_n)_{(j)}\|_2 &\leq \max_{j\in S}\left\|\left[(\mathbf{D}_n)_{(j,S)} - \mathbf{D}_{(j,S)}\right]\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2 + \sqrt{L_n}\left|\boldsymbol{b}_{(S)} - (\boldsymbol{b}_n)_{(S)}\right|_\infty \\ &\lesssim \sqrt{\frac{L_n\log p}{n}} + a_nL_ns\sqrt{\frac{\log p}{n}} + \sqrt{\frac{L_n\log p}{n}} + a_n \\ &\lesssim \sqrt{\frac{L_n\log p}{n}} + a_nL_ns\sqrt{\frac{\log p}{n}} + a_n,\end{aligned}$$

holds with probability at least $1 - L_np^{-\vartheta} - L_nsp^{-\vartheta} - sp^{-\vartheta L_n}$. Next we will derive the bound for $j \in S^c$. From (B.18), we claim that for any $j \in S^c$

$$\begin{aligned}\|(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \boldsymbol{b}_n)_{(j)}\|_2 \leq{}& \left\|(\mathbf{D}_{(S^cS)}\widetilde{\boldsymbol{\gamma}}_{(S)})_{(j)} - \boldsymbol{b}_{(j)}\right\|_2 + \left\|\left[(\mathbf{D}_n)_{(j,S)} - \mathbf{D}_{(j,S)}\right]\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2 \\ &+ \sqrt{L_n}\left|\boldsymbol{b}_{(S^c)} - (\boldsymbol{b}_n)_{(S^c)}\right|_\infty.\end{aligned}$$

Using the optimality of $\boldsymbol{\theta}^*(U)$, we have

$$\mathbb{E}\left\{\widetilde{\boldsymbol{B}}_{(S^c)}\left[Z - \sum_{j\in S}(X_j - \mu_j(U))\theta_j^*(U)\right]\right\} = \mathbf{0}.$$

Together with $\widetilde{\theta}_j(U) = \boldsymbol{B}(U)^\top\widetilde{\boldsymbol{\gamma}}_{(j)}$, $\boldsymbol{b}_{(S^c)} = \mathbb{E}[\widetilde{\boldsymbol{B}}_{(S^c)}Z]$ and $\widetilde{\boldsymbol{B}}_{(j)} = (X_j - \mu_j(U))\boldsymbol{B}(U)$, we also have

$$\begin{aligned}\mathbf{D}_{(S^cS)}\widetilde{\boldsymbol{\gamma}}_{(S)} &= \mathbb{E}\left\{\widetilde{\boldsymbol{B}}_{(S^c)}\widetilde{\boldsymbol{B}}_{(S)}^\top\widetilde{\boldsymbol{\gamma}}_{(S)}\right\} \\ &= \mathbb{E}\left\{\widetilde{\boldsymbol{B}}_{(S^c)}\sum_{j\in S}(X_j - \mu_j(U))\boldsymbol{B}(U)^\top\widetilde{\boldsymbol{\gamma}}_{(j)}\right\} \\ &= \mathbb{E}\left\{\widetilde{\boldsymbol{B}}_{(S^c)}\left[\sum_{j\in S}(X_j - \mu_j(U))\widetilde{\theta}_j(U) - Z\right]\right\} + \boldsymbol{b}_{(S^c)} \\ &= \mathbb{E}\left\{\widetilde{\boldsymbol{B}}_{(S^c)}\left[\sum_{j\in S}(X_j - \mu_j(U))(\widetilde{\theta}_j(U) - \theta_j^*(U))\right]\right\} + \boldsymbol{b}_{(S^c)}.\end{aligned}$$

Let $\boldsymbol{c}_{j,S}(U) = \mathbb{E}[(X_j - \mu_j(U))(\boldsymbol{X} - \boldsymbol{\mu}(U))_S|U]$, then for $j \in S^c$, it holds that

$$(\mathbf{D}_{(S^cS)}\widetilde{\boldsymbol{\gamma}}_{(S)})_{(j)} - \boldsymbol{b}_{(j)} = \mathbb{E}\left[\widetilde{\boldsymbol{B}}_{(j)}(\boldsymbol{X} - \boldsymbol{\mu}(U))_S^\top(\boldsymbol{\theta}^*(U) - \widetilde{\boldsymbol{\theta}}(U))_S\right].$$

For any $\boldsymbol{\nu} \in \mathbb{S}^{L_n-1}$,

$$\begin{aligned}&\left|\mathbb{E}\left[\boldsymbol{\nu}^\top\widetilde{\boldsymbol{B}}_{(j)}(\boldsymbol{X} - \boldsymbol{\mu}(U))_S^\top(\boldsymbol{\theta}^*(U) - \widetilde{\boldsymbol{\theta}}(U))_S\right]\right| \\ &= \left|\mathbb{E}\left[\boldsymbol{\nu}^\top\boldsymbol{B}(X_j - \mu_j(U))(\boldsymbol{X} - \boldsymbol{\mu}(U))_S^\top(\boldsymbol{\theta}^*(U) - \widetilde{\boldsymbol{\theta}}(U))_S\right]\right|\end{aligned}$$

$$
\begin{aligned}
&= \left|\mathbb{E}\left[\boldsymbol{\nu}^\top \boldsymbol{B}\boldsymbol{c}_{j,S}(U)^\top (\boldsymbol{\theta}^*(U) - \widetilde{\boldsymbol{\theta}}(U))_S\right]\right| \\
&\le \sup_{u\in[0,1]} \left\{\|\boldsymbol{c}_{j,S}(u)\|_2 \|(\boldsymbol{\theta}^*(u) - \widetilde{\boldsymbol{\theta}}(u))_S\|_2\right\} \mathbb{E}\left[|\boldsymbol{\nu}^\top \boldsymbol{B}|\right] \\
&\lesssim \delta_s \sqrt{s} L_n^{-d} \left(\mathbb{E}\left[\left(\boldsymbol{\nu}^\top \boldsymbol{B}\right)^2\right]\right)^{1/2} \\
&\lesssim \sqrt{s} L_n^{-d},
\end{aligned}
$$

where we used Theorem 3.3 and

$$
\begin{aligned}
\sup_{u\in[0,1]} \|\boldsymbol{c}_{j,S}(U)\|_2 &= \sup_{u\in[0,1]} \|(\mu_{1j}(u) - \mu_j(u))(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}(u))_S\|_2 \\
&\lesssim \sup_{u\in[0,1]} \|(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))_S\|_2 \le \delta_s.
\end{aligned}
$$

Hence we have for any $j \in S^c$,

$$
\left\|(\mathbf{D}_{(S^c S)}\widetilde{\boldsymbol{\gamma}}_{(S)})_{(j)} - \boldsymbol{b}_{(j)}\right\|_2 \le \sup_{\boldsymbol{\nu}\in\mathbb{S}^{L_n-1}} \left|\boldsymbol{\nu}^\top \left((\mathbf{D}_{(S^c S)}\widetilde{\boldsymbol{\gamma}}_{(S)})_{(j)} - \boldsymbol{b}_{(j)}\right)\right| \lesssim \delta_s \sqrt{s} L_n^{-d}.
$$

Then applying Lemma B.7 and B.5, we claim that

$$
\max_{j\in S^c} \|(\mathbf{D}_n \widetilde{\boldsymbol{\gamma}} - \boldsymbol{b}_n)_{(j)}\|_2 \lesssim \sqrt{\frac{L_n \log p}{n}} + a_n L_n s \sqrt{\frac{\log p}{n}} + a_n + \sqrt{s} L_n^{-d}
$$

holds with probability at least $1 - 10 L_n p^{-\vartheta}$. □

Appendix C: Deferred proofs of Section A and B

C.1. Proof of Lemma A.3

Proof. Let $\mathbb{S}^{L_n-1}$ be the unit sphere in $\mathbb{R}^{L_n}$, we denote the $\frac{1}{8}$-covering of $\mathbb{S}^{L_n-1}$ by $\{\boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_K\}$ with $K \le 17^{L_n}$. Let $\mathbf{Q} = \sum_{i=1}^n \boldsymbol{B}_i \boldsymbol{B}_i^\top / n - \mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]$, then we have

$$
\|\mathbf{Q}\|_2 = \sup_{\boldsymbol{\nu}\in\mathbb{S}^{L_n-1}} |\boldsymbol{\nu}^\top \mathbf{Q}\boldsymbol{\nu}|.
$$

Based on the definition of covering set, for any $\boldsymbol{\nu} \in \mathbb{S}^{L_n-1}$, there exists some $1 \le k \le K$ such that $\|\boldsymbol{\nu} - \boldsymbol{\nu}_k\|_2 \le 1/8$. It follows that

$$
\begin{aligned}
|\boldsymbol{\nu}^\top \mathbf{Q}\boldsymbol{\nu}| &\le |\boldsymbol{\nu}_k^\top \mathbf{Q}\boldsymbol{\nu}_k| + 2|\boldsymbol{\nu}_k^\top \mathbf{Q}(\boldsymbol{\nu}_k - \boldsymbol{\nu})| + |(\boldsymbol{\nu}_k - \boldsymbol{\nu})^\top \mathbf{Q}(\boldsymbol{\nu}_k - \boldsymbol{\nu})| \\
&\le |\boldsymbol{\nu}_k^\top \mathbf{Q}\boldsymbol{\nu}_k| + \frac{1}{4}\|\mathbf{Q}\|_2 + \frac{1}{64}\|\mathbf{Q}\|_2 \\
&\le |\boldsymbol{\nu}_k^\top \mathbf{Q}\boldsymbol{\nu}_k| + \frac{1}{2}\|\mathbf{Q}\|_2.
\end{aligned}
$$

Thus we have

$$
\|\mathbf{Q}\|_2 \le 2 \max_{1\le k\le K} |\boldsymbol{\nu}_k^\top \mathbf{Q}\boldsymbol{\nu}_k| = 2 \max_{1\le k\le K} \left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{\nu}_k^\top \boldsymbol{B}_i)^2 - \mathbb{E}[(\boldsymbol{\nu}_k^\top \boldsymbol{B})^2]\right|. \tag{C.1}
$$

Since $\|\boldsymbol{B}_i^*\|_2^2 = \sum_{k=1}^{L_n}(B_k^*(U_i))^2 \leq \sum_{k=1}^{L_n} B_k^*(U_i) = 1$, together with (A.1), we have

$$\mathbb{E}\left[(\boldsymbol{\nu}_k^\top \boldsymbol{B}_i^*)^4 \exp\{\eta(\boldsymbol{\nu}_k^\top \boldsymbol{B}_i^*)^2\}\right] \leq e^\eta \mathbb{E}\left[(\boldsymbol{\nu}_k^\top \boldsymbol{B}_i^*)^2\right] \lesssim L_n^{-1}\|\boldsymbol{\nu}_k\|_2^2 = L_n^{-1},$$

together with Lemma A.1 we claim that

$$\mathbb{P}\left(\max_{1\leq k\leq K}\left|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\nu}_k^\top \boldsymbol{B}_i)^2 - \mathbb{E}[(\boldsymbol{\nu}_k^\top \boldsymbol{B})^2]\right| \geq CL_n\sqrt{\frac{L_n \log n}{nL_n}}\right) \leq Kn^{-\vartheta L_n^3} \leq n^{-\vartheta L_n}.$$

Then the conclusion follows immediately. □

C.2. Proof of Lemma B.2

Proof of Lemma B.2. Let $\mathbf{C}(u) = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu}(u))(\boldsymbol{X} - \boldsymbol{\mu}(u))^\top]$, and then it holds

$$\begin{aligned}\mathbf{C}(u) &= \frac{1}{2}\left(\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}(u))(\boldsymbol{X} - \boldsymbol{\mu}(u))^\top | Y = 1\right] + \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu}(u))(\boldsymbol{X} - \boldsymbol{\mu}(u))^\top | Y = 0\right]\right) \\ &= \boldsymbol{\Sigma}(u) + \frac{1}{4}(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))(\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))^\top.\end{aligned}$$

From conditions (**C**1) and (**C**4) we have

$$\lambda_0 \leq \lambda_{\min}(\mathbf{C}(u)) \leq \lambda_{\max}(\mathbf{C}(u)) \leq \lambda_1 + \frac{1}{4}\delta_p, \tag{C.2}$$

holds for any $u \in [0, 1]$. In addition, we notice that

$$\begin{aligned}\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top] &= \mathbb{E}\left[\left((\boldsymbol{X} - \boldsymbol{\mu}(U))(\boldsymbol{X} - \boldsymbol{\mu}(U))^\top\right) \otimes \left(\boldsymbol{B}\boldsymbol{B}^\top\right)\right] \\ &= \mathbb{E}\left[\mathbf{C}(U) \otimes \left(\boldsymbol{B}\boldsymbol{B}^\top\right)\right].\end{aligned}$$

Then for any $\boldsymbol{\eta} = (\boldsymbol{\eta}_{(1)}^\top, \cdots, \boldsymbol{\eta}_{(p)}^\top)^\top \in \mathbb{R}^{L_n p}$ with $\|\boldsymbol{\eta}\|_2 = 1$, we have

$$\begin{aligned}\boldsymbol{\eta}^\top \mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]\boldsymbol{\eta} &= \mathbb{E}\left(\sum_{j=1}^p (X_j - \mu_j(U))\boldsymbol{B}^\top \boldsymbol{\eta}_j\right)^2 \\ &= \mathbb{E}\left[\left(\boldsymbol{B}^\top \boldsymbol{\eta}_{(1)}, \cdots, \boldsymbol{B}^\top \boldsymbol{\eta}_{(p)}\right) \mathbf{C}(U) \left(\boldsymbol{B}^\top \boldsymbol{\eta}_{(1)}, \cdots, \boldsymbol{B}^\top \boldsymbol{\eta}_{(p)}\right)^\top\right] \\ &\geq \inf_{u\in[0,1]} \lambda_{\min}(\mathbf{C}(u))\, \mathbb{E}\left[\sum_{j=1}^p \boldsymbol{\eta}_{(j)}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{\eta}_{(j)}\right] \\ &\geq \lambda_0 \lambda_{\min}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]).\end{aligned}$$

Similarly, we have

$$\boldsymbol{\eta}^\top \mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]\boldsymbol{\eta} \leq \left(\lambda_1 + \frac{\delta_p}{4}\right) \lambda_{\max}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]).$$

Then the result follows from $\lambda_{\min}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]) = O(1)$ and $\lambda_{\max}(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]) = O(1)$ (see Section A.1). □

C.3. Proof of Lemma B.3

To prove Lemma B.3, we impose the following five lemmas on the concentration inequalities of random matrices. The proofs can be found in Appendix D.1 - D.4.

Lemma C.1. *Let $\boldsymbol{Z}_i = \boldsymbol{Z}(U_i) = (\boldsymbol{X}_i - \boldsymbol{\mu}_k(U_i)) \otimes \boldsymbol{B}_i$, then under condition* (**C**1)-(**C**4)*, we have for any $\vartheta > 0$ and $k = 1, 2$*

$$\left\| \frac{1}{n} \sum_{i \in \mathcal{I}_k} \left\{ \boldsymbol{Z}_i \boldsymbol{Z}_i^\top - \mathbb{E}\left[\boldsymbol{Z}_i \boldsymbol{Z}_i^\top | Y_i = k \right] \right\} \right\|_2 \lesssim L_n \sqrt{\frac{p \log n}{n}},$$

and

$$\left\| \frac{1}{n} \sum_{i \in \mathcal{I}_k} \left\{ \boldsymbol{Z}_i \boldsymbol{Z}_i^\top - \mathbb{E}\left[\boldsymbol{Z}_i \boldsymbol{Z}_i^\top | Y_i = k \right] \right\} \widetilde{\boldsymbol{\gamma}} \right\|_2 \lesssim \sqrt{\frac{p L_n \log n}{n}},$$

hold with probability at least $1 - n^{-\vartheta p L_n}$.

Lemma C.2. *Under conditions* (**C**1)-(**C**4)*, for $k = 1, 2$ and any $\vartheta > 0$, we have*

$$\left\| \frac{1}{n} \sum_{i \in \mathcal{I}_k} [(\boldsymbol{X}_i - \boldsymbol{\mu}_k(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes \left(\boldsymbol{B}_i \boldsymbol{B}_i^\top \right) \right\|_2 \lesssim L_n \sqrt{\frac{p \log n}{n}},$$

$$\left\| \frac{1}{n} \sum_{i \in \mathcal{I}_k} \left\{ [(\boldsymbol{X}_i - \boldsymbol{\mu}_k(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes \left(\boldsymbol{B}_i \boldsymbol{B}_i^\top \right) \right\} \widetilde{\boldsymbol{\gamma}} \right\|_2 \lesssim \sqrt{\frac{p L_n \log n}{n}},$$

hold with probability at least $1 - n^{-\vartheta L_n p}$.

Lemma C.3. *Under conditions* (**C**1)-(**C**4)*, let $\mathbf{A}_i = [(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top)$, then for any $\vartheta > 0$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{A}_i - \mathbb{E}[\mathbf{A}_i] \right\|_2 \lesssim L_n \sqrt{\frac{p \log n}{n}},$$

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]) \widetilde{\boldsymbol{\gamma}} \right\|_2 \lesssim \sqrt{\frac{p L_n \log n}{n}}$$

hold with probability at least $1 - n^{-\vartheta L_n p}$.

Lemma C.4. *Under conditions* (**C**1)-(**C**4)*, then for $k = 1, 2$ and any $\vartheta > 0$,*

$$\left\| \frac{1}{n} \sum_{i \in \mathcal{I}_k} [(\boldsymbol{X}_i - \boldsymbol{\mu}_k(U_i))(\widehat{\boldsymbol{\mu}}(U_i) - \boldsymbol{\mu}(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\|_2$$
$$\lesssim p L_n^{3/2} \sqrt{\frac{\log n}{n}} \left(\sqrt{\frac{L_n \log n}{n}} + L_n^{-d} \right),$$

$$\begin{aligned}&\left\| \frac{1}{n}\sum_{i\in\mathcal{I}_k}\left\{[(\boldsymbol{X}_i-\boldsymbol{\mu}_k(U_i))(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))^\top]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}\widetilde{\gamma}\right\|_2\\ &\qquad\lesssim pL_n\sqrt{\frac{\log n}{n}}\left(\sqrt{\frac{L_n\log n}{n}}+L_n^{-d}\right),\end{aligned}$$

hold with probability at least $1-n^{-\vartheta pL_n}-pL_nn^{-\vartheta}$.

Lemma C.5. *Under conditions* (**C**1)-(**C**4), *let* $\mathbf{G}(U_i)=[(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))^\top]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)$, *we have for any* $\vartheta>0$

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\mathbf{G}(U_i)-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\mathbf{G}(U_i)\right\|_2\lesssim pL_n^{3/2}\sqrt{\frac{\log n}{n}}\left(\sqrt{\frac{L_n\log n}{n}}+L_n^{-d}\right),$$

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\mathbf{G}(U_i)\widetilde{\gamma}-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\mathbf{G}(U_i)\widetilde{\gamma}\right\|_2\lesssim pL_n\sqrt{\frac{\log n}{n}}\left(\sqrt{\frac{L_n\log n}{n}}+L_n^{-d}\right),$$

hold with probability at least $1-n^{-\vartheta pL_n}-pL_nn^{-\vartheta}$.

Proof of Lemma B.3. Note that we can rewrite $\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]$ and $\sum_{i=1}^{2n}\widetilde{\boldsymbol{B}}_i\widetilde{\boldsymbol{B}}_i^\top/2n$ as

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top\right]&=\underbrace{\frac{1}{2}\mathbb{E}\left\{\left[(\boldsymbol{X}-\boldsymbol{\mu}_1(U))\,(\boldsymbol{X}-\boldsymbol{\mu}_1(U))^\top\right]\otimes(\boldsymbol{B}\boldsymbol{B}^\top)\middle|Y=1\right\}}_{\mathbf{I}_1^*}\\&+\underbrace{\frac{1}{2}\mathbb{E}\left\{\left[(\boldsymbol{X}-\boldsymbol{\mu}_2(U))\,(\boldsymbol{X}-\boldsymbol{\mu}_2(U))^\top\right]\otimes(\boldsymbol{B}\boldsymbol{B}^\top)\middle|Y=0\right\}}_{\mathbf{I}_2^*}\\&+\underbrace{\frac{1}{4}\mathbb{E}\left\{[(\boldsymbol{\mu}_1(U)-\boldsymbol{\mu}_2(U))(\boldsymbol{\mu}_1(U)-\boldsymbol{\mu}_2(U))^\top]\otimes(\boldsymbol{B}\boldsymbol{B}^\top)\right\}}_{\mathbf{I}_3^*}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{2n}\widetilde{\boldsymbol{B}}_i\widetilde{\boldsymbol{B}}_i^\top&=\underbrace{\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left\{[(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}(U_i))^\top]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}^1}\\&+\underbrace{\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left\{[(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\widehat{\boldsymbol{\mu}}(U_i))^\top]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}^2}.\end{aligned}$$

To upper bound $\|\sum_{i=1}^{2n}\widetilde{\boldsymbol{B}}_i\widetilde{\boldsymbol{B}}_i^\top/2n-\mathbb{E}[\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^\top]\|_2$, we begin with the following decompositions for $\mathbf{I}_1$ and $\mathbf{I}_2$,

$$
\begin{aligned}
\mathbf{I}^1 = & \underbrace{\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_1^1} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left\{\left[(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}_2^1} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left\{\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}_3^1} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_4^1}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{I}^2 = & \underbrace{\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_1^2} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left\{\left[(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}_2^2} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left\{\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}}_{\mathbf{I}_3^2} \\
& \underbrace{+\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_4^2}
\end{aligned}
$$

Step 1.1. upper bounding $\|\mathbf{I}_1^1-\mathbf{I}_1^*+\mathbf{I}_1^2-\mathbf{I}_2^*-\mathbf{I}_3^*\|_2$ First, we decompose $\mathbf{I}_1^1$ as

$$
\begin{aligned}
\mathbf{I}_1^1 = & \underbrace{\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_{11}^1} \\
& \underbrace{+\frac{1}{4n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_{12}^1}
\end{aligned}
$$

$$+ \underbrace{\frac{1}{4n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_{13}^1}$$

$$+ \underbrace{\frac{1}{8n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)}_{\mathbf{I}_{14}^1}$$

By Lemma C.1, we have

$$\mathbb{P}\left(\left\|\mathbf{I}_{11}^1-\mathbf{I}_1^*\right\|_2 \geq CL_n\sqrt{\frac{p\log n}{n}}\right) \leq n^{-\vartheta L_n p}.$$

By Lemma C.2, we have

$$\mathbb{P}\left(\left\|\mathbf{I}_{12}^1\right\|_2 \lesssim L_n\sqrt{\frac{p\log n}{n}}\right) \geq 1-n^{-\vartheta L_n p},$$

and

$$\mathbb{P}\left(\left\|\mathbf{I}_{13}^1\right\|_2 \lesssim CL_n\sqrt{\frac{p\log n}{n}}\right) \geq 1-n^{-\vartheta L_n p}.$$

By Lemma C.3, we have

$$\mathbb{P}\left(\|\mathbf{I}_{14}^1-\frac{1}{2}\mathbf{I}_3^*\|_2 \lesssim L_n\sqrt{\frac{p\log n}{n}}\right) \geq 1-pL_n n^{-\vartheta}.$$

Combining the results displayed above, it follows that

$$\mathbb{P}\left(\|\mathbf{I}_1^1-\mathbf{I}_1^*-\frac{1}{2}\mathbf{I}_3^*\|_2 \lesssim L_n\sqrt{\frac{p\log n}{n}}\right) \geq 1-3n^{-\vartheta L_n p}-pL_n n^{-\vartheta}. \tag{C.3}$$

Similarly, we also have

$$\mathbb{P}\left(\|\mathbf{I}_1^2-\mathbf{I}_2^*-\frac{1}{2}\mathbf{I}_3^*\|_2 \lesssim L_n\sqrt{\frac{p\log n}{n}}\right) \geq 1-3n^{-\vartheta L_n p}-pL_n n^{-\vartheta}. \tag{C.4}$$

Step 1.2. upper bounding $\|\mathbf{I}_2^1+\mathbf{I}_2^2\|_2$, $\|\mathbf{I}_3^1+\mathbf{I}_3^2\|_2$ and $\|\mathbf{I}_4^1\|_2+\|\mathbf{I}_4^2\|_2$ Note that

$$\begin{aligned}\|\mathbf{I}_2^1+\mathbf{I}_2^2\|_2 \leq &\left\|\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\|_2\\ &+\left\|\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left[(\boldsymbol{X}_i-\boldsymbol{\mu}_2(U_i))(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\|_2\\ &+\left\|\frac{1}{4n}\sum_{i\in\mathcal{I}_1}\mathbf{G}(U_i)-\frac{1}{4n}\sum_{i\in\mathcal{I}_2}\mathbf{G}(U_i)\right\|_2,\end{aligned}$$

where $\mathbf{G}(U_i) = [(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))(\widehat{\boldsymbol{\mu}}(U_i) - \boldsymbol{\mu}(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top)$. By invoking Lemma C.5 and C.3, we have

$$\mathbb{P}\left(\|\mathbf{I}_2^1 + \mathbf{I}_2^2\|_2 \lesssim pL_n^{3/2} a_n \sqrt{\frac{\log n}{n}} \right) \geq 1 - n^{-\vartheta L_n p}, \tag{C.5}$$

where $a_n = \sqrt{L_n \log n/n} + L_n^{-d}$. Similarly, we can obtain

$$\mathbb{P}\left(\|\mathbf{I}_3^1 + \mathbf{I}_3^2\|_2 \lesssim pL_n^{3/2} a_n \sqrt{\frac{\log n}{n}} \right) \geq 1 - n^{-\vartheta L_n p}. \tag{C.6}$$

In addition, by Proposition A.1 we have

$$\begin{aligned} \|\mathbf{I}_4^1\|_2 + \|\mathbf{I}_4^2\|_2 &\leq 2 \max_{1 \leq i \leq n} \|\widehat{\boldsymbol{\mu}}(U_i) - \boldsymbol{\mu}(U_i)\|_2^2 \|\boldsymbol{B}_i\|_2^2 \\ &\leq 2pL_n a_n^2 \end{aligned} \tag{C.7}$$

Combining (C.3)-(C.7), we have

$$\left\| \frac{1}{2n} \sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i \widetilde{\boldsymbol{B}}_i^\top - \mathbb{E}[\widetilde{\boldsymbol{B}} \widetilde{\boldsymbol{B}}^\top] \right\|_2 \lesssim L_n \sqrt{\frac{p \log n}{n}} + pL_n^{3/2} a_n \sqrt{\frac{\log n}{n}} \tag{C.8}$$

holds with probability at least $1 - n^{-\vartheta L_n p} - pL_n n^{-\vartheta}$. □

C.4. Proof of Lemma B.4

Proof. By replacing the bounds for the operator norm of matrices with those for ℓ_2-norm of matrix-vector-products in Section C.3, we can finish the proof Due to the fact that $\widetilde{\boldsymbol{B}}_i^\top \widetilde{\gamma}_{(j)} = \widetilde{\theta}_j(U_i)$ is bounded, we can drop a $\sqrt{L_n}$ factor for matrix-vector-product bounds in Lemma C.1- C.3. □

C.5. Proof of Lemma B.5

Proof. Note that for any $j = 1, 2, ..., p$ and $k = 1, 2, ..., L_n$, we find that

$$\begin{aligned} &\frac{1}{2n} \sum_{i=1}^{2n} B_k(U_i)(X_{ij} - \widehat{\mu}_j(U_i)) Z_i \\ =&\frac{1}{4n} \sum_{i \in \mathcal{I}_1} B_k(U_i)(X_{ij} - \widehat{\mu}_j(U_i)) - \frac{1}{4n} \sum_{i \in \mathcal{I}_2} B_k(U_i)(X_{ij} - \widehat{\mu}_j(U_i)) \\ =&\frac{1}{4} \underbrace{\frac{1}{2n} \sum_{i=1}^{2n} B_k(U_i) \widehat{\mu}_{1j}(U_i)}_{I_1} - \frac{1}{4} \underbrace{\frac{1}{2n} \sum_{i=1}^{2n} B_k(U_i) \widehat{\mu}_{2j}(U_i)}_{I_2}, \end{aligned}$$

where we used $\widehat{\mu}_{1j} = \boldsymbol{B}_i^\top \widehat{\boldsymbol{\alpha}}_{1j}$, $\widehat{\mu}_{2j} = \boldsymbol{B}_i^\top \widehat{\boldsymbol{\alpha}}_{2j}$ and the optimal condition for $\widehat{\boldsymbol{\alpha}}_{1j}$ and $\widehat{\boldsymbol{\alpha}}_{2j}$ such that

$$\sum_{i \in \mathcal{I}_1} \boldsymbol{B}_i(X_{ij} - \boldsymbol{B}_i^\top \widehat{\boldsymbol{\alpha}}_{1j}) = \mathbf{0}, \quad \sum_{i \in \mathcal{I}_2} \boldsymbol{B}_i(X_{ij} - \boldsymbol{B}_i^\top \widehat{\boldsymbol{\alpha}}_{2j}) = \mathbf{0}.$$

Moreover, note that

$$\begin{aligned}
&\mathbb{E}[B_k(U)(X_j - \mu_j(U))Z] \\
=&\frac{1}{4}\mathbb{E}[B_k(U)(X_j - \mu_j(U))|Y=1] - \frac{1}{4}\mathbb{E}[B_k(U)(X_j - \mu_j(U))|Y=0] \\
=&\frac{1}{4}\underbrace{\mathbb{E}[B_k(U)\mu_{1j}(U)]}_{I_1^*} - \frac{1}{4}\underbrace{\mathbb{E}[B_k(U)\mu_{2j}(U)]}_{I_2^*}
\end{aligned}$$

The remaining detail is to upper bound $|I_1 - I_1^*|$. According to Proposition A.1 and $B_k = \sqrt{L_n} B_k^*$, we have

$$\begin{aligned}
|I_1 - I_1^*| \le \sqrt{L_n}&\left|\frac{1}{2n}\sum_{i=1}^{2n} B_k^*(U_i)[\widehat{\mu}_{1j}(U_i) - \mu_{1j}(U_i)]\right| \\
&+ \sqrt{L_n}\left|\frac{1}{2n}\sum_{i=1}^{2n} B_k^*(U_i)\mu_{1j}(U_i) - \mathbb{E}[B_k^*(U)\mu_{1j}(U)]\right| \\
\le a_n\sqrt{L_n}&\left|\frac{1}{2n}\sum_{i=1}^{2n} |B_k^*(U_i)| - \mathbb{E}[|B_k^*(U_i)|]\right| + a_n\sqrt{L_n}\mathbb{E}[|B_k^*(U)|] \\
&+ \sqrt{L_n}\left|\frac{1}{2n}\sum_{i=1}^{2n} B_k^*(U_i)\mu_{1j}(U_i) - \mathbb{E}[B_k^*(U)\mu_{1j}(U)]\right|.
\end{aligned}$$

Using Lemma A.1, we can verify

$$\mathbb{P}\left(\left|\frac{1}{2n}\sum_{i=1}^{2n} B_k^*(U_i)\mu_{1j}(U_i) - \mathbb{E}[B_k^*(U)\mu_{1j}(U)]\right| \lesssim \sqrt{\frac{\log n}{L_n n}}\right) \ge 1 - n^{-\vartheta}$$

and

$$\mathbb{P}\left(\left|\frac{1}{2n}\sum_{i=1}^{2n} |B_k^*(U_i)| - \mathbb{E}[|B_k^*(U_i)|]\right| \lesssim \sqrt{\frac{\log n}{L_n n}}\right) \ge 1 - n^{-\vartheta}.$$

In addition, recall the fact that $\mathbb{E}|B_k^*(U)| \le M_2 L_n^{-1}$, which yields that

$$\mathbb{P}\left(|I_1 - I_1^*| \lesssim \sqrt{\frac{\log n}{n}} + L_n^{-1/2} a_n\right) \ge 1 - n^{-\vartheta}.$$

Thus we are guaranteed that

$$\mathbb{P}\left(\left|\frac{1}{2n}\sum_{i=1}^{2n} \widetilde{\boldsymbol{B}}_i Z_i - \mathbb{E}[\widetilde{\boldsymbol{B}} Z]\right| \lesssim \sqrt{\frac{\log n}{n}} + L_n^{-1/2} a_n\right) \ge 1 - 4pL_n n^{-\vartheta}. \quad \text{(C.9)}$$

□

C.6. Proof of Lemma B.6

Proof. By the optimality of $\widehat{\boldsymbol{\gamma}}$, we have

$$\frac{1}{2}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)^{\top}\mathbf{A}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)+\lambda_n\sum_{j=1}^{p}\|\widehat{\boldsymbol{\gamma}}_{(j)}\|_2\leq\left(\mathbf{A}\check{\boldsymbol{\gamma}}-\boldsymbol{b}\right)^{\top}\left(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}\right)+\lambda_n\sum_{j=1}^{p}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2. \tag{C.10}$$

Using the condition (B.16) and dropping the first non-negative term in the left hand side of (C.10), we are guaranteed that

$$\begin{aligned}\lambda_n\sum_{j=1}^{p}\|\widehat{\boldsymbol{\gamma}}_{(j)}\|_2&\leq\sum_{j=1}^{p}\|(\mathbf{A}\check{\boldsymbol{\gamma}}-\boldsymbol{b})_{(j)}\|_2\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\lambda_n\sum_{j=1}^{p}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2\\&\leq\frac{\lambda_n}{2}\sum_{j=1}^{p}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\lambda_n\sum_{j=1}^{p}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2\\&=\frac{\lambda_n}{2}\sum_{j\in S}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\frac{\lambda_n}{2}\sum_{j\in S^c}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\lambda_n\sum_{j\in S}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2.\end{aligned}$$

It follows from the assumption $\check{\boldsymbol{\gamma}}_{(j)}=\mathbf{0}$ for $j\in S^c$ that

$$\begin{aligned}\frac{1}{2}\sum_{j\in S^c}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2&\leq\frac{1}{2}\sum_{j\in S}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\sum_{j\in S}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2-\sum_{j\in S}\|\widehat{\boldsymbol{\gamma}}_{(j)}\|_2\\&\leq\frac{1}{2}\sum_{j\in S}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\sum_{j\in S}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2\\&\leq\frac{3}{2}\sum_{j\in S}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2,\end{aligned}\tag{C.11}$$

where we used the fact $\|\check{\boldsymbol{\gamma}}_{(j)}\|_2-\|\widehat{\boldsymbol{\gamma}}_{(j)}\|_2\leq\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2$. From (C.10), we can also obtain that

$$\begin{aligned}\frac{1}{2}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)^{\top}\mathbf{A}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)&\leq\frac{\lambda_n}{2}\sum_{j=1}^{p}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2+\lambda_n\sum_{j=1}^{p}\|\check{\boldsymbol{\gamma}}_{(j)}\|_2-\lambda_n\sum_{j=1}^{p}\|\widehat{\boldsymbol{\gamma}}_{(j)}\|_2\\&\leq\frac{3\lambda_n}{2}\sum_{j=1}^{p}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2.\end{aligned}\tag{C.12}$$

By the restrictive eigenvalue condition of $\mathbf{A}$, we know

$$\frac{1}{2}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)^{\top}\mathbf{A}\left(\widehat{\boldsymbol{\gamma}}-\check{\boldsymbol{\gamma}}\right)\geq\frac{\zeta}{2}\|\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}\|_2^2.$$

Together with (C.12) and (C.11), we further have

$$\zeta\|\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}\|_2^2\leq3\lambda_n\sum_{j=1}^{p}\|(\check{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}})_{(j)}\|_2$$

$$
\begin{aligned}
&= 3\lambda_n \left(\sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 + \sum_{j\in S^c} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 \right) \\
&\leq 12\lambda_n \sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 \\
&\leq 12\lambda_n \sqrt{s} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(S)}\|_2 \\
&\leq 12\lambda_n \sqrt{s} \|\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}\|_2,
\end{aligned}
$$

which yields the first conclusion in Lemma B.6. In fact, we also used the following relation

$$
\left(\sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 \right)^2 \leq s \sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2^2 = s\|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(S)}\|_2^2.
$$

And the second conclusion holds since

$$
\begin{aligned}
\|\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}\|_1 &= \sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_1 + \sum_{j\in S^c} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_1 \\
&\leq \sqrt{L_n} \sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 + \sqrt{L_n} \sum_{j\in S^c} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 \\
&\leq 4\sqrt{L_n} \sum_{j\in S} \|(\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}})_{(j)}\|_2 \\
&\leq 4\sqrt{sL_n} \|\check{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}\|_2.
\end{aligned}
$$

□

C.7. Proof of Lemma B.7

Proof of Lemma B.7. Recall that $\widetilde{\boldsymbol{\gamma}}_{(S^c)} = \mathbf{0}$, thus

$$
\max_{1\leq j\leq p} \|(\mathbf{D}_n\widetilde{\boldsymbol{\gamma}} - \mathbf{D}\widetilde{\boldsymbol{\gamma}})_{(j)}\|_2 = \max_{1\leq j\leq p} \|(\mathbf{D}_n - \mathbf{D})_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2.
$$

Then it suffices to show that

$$
\max_{1\leq j\leq p} \|(\mathbf{D}_n - \mathbf{D})_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2 \lesssim \sqrt{\frac{\log p}{n}} + \frac{L_n \log p}{n} + L_n^{-2d}, \tag{C.13}
$$

with high probability. We use the same decomposition for $\mathbf{D}_n - \mathbf{D}$ in the Section C.3 and only prove the counterpart to $\{i : Y_i = 1\}$. Correspondingly, the expectation $\mathbb{E}[\cdot]$ means conditional expectation $\mathbb{E}[\cdot|Y_i = 1]$. We split the proof into the following two main steps.

Step 1. upper bounding $\max_{1\leq j\leq p} \|(\mathbf{I}_1^1 - \mathbf{I}_1^* + \mathbf{I}_1^2 - \mathbf{I}_2^* - \mathbf{I}_3^*)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2$ Recall

$$
\left\| (\mathbf{I}_1^1 - \mathbf{I}_1^* - \frac{1}{2}\mathbf{I}_3^*)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \leq \left\| (\mathbf{I}_{11}^1 - \mathbf{I}_1^*)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2
$$

$$+ \left\| (\mathbf{I}_{12}^1)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 + \left\| (\mathbf{I}_{13}^1)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 + \left\| (\mathbf{I}_{14}^1 - \frac{1}{2}\mathbf{I}_3^*)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 .$$

Notice that

$$\left\| (\mathbf{I}_{11}^1)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 = \sqrt{L_n} \left\| \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mu_{1j}(U_i)) \widetilde{\boldsymbol{\theta}}_S(U_i)^\top \left(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i) \right)_S \boldsymbol{B}_i^* \right\|_2 .$$

Given $Y_i = 1$ and U_i, $\widetilde{\boldsymbol{\theta}}_S(U_i)^\top \left(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i) \right)_S$ is a normal random variable with mean-zero. Due to our assumption $\sup_{u\in[0,1]} \|\boldsymbol{\theta}^*(u)\|_2 \le \delta_s$, together with Theorem 3.3, we have

$$\begin{aligned}
\mathbb{E}_1 \left[\left(\widetilde{\boldsymbol{\theta}}_S(U_i)^\top \left(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i) \right)_S \right)^2 |U_i \right] &= \widetilde{\boldsymbol{\theta}}_S(U_i)^\top \boldsymbol{\Sigma}_{SS}(U_i) \widetilde{\boldsymbol{\theta}}_S(U_i) \\
&\le \lambda_1 \|\widetilde{\boldsymbol{\theta}}_S(U_i)\|_2^2 \\
&\le 2\lambda_1 \left(\|\widetilde{\boldsymbol{\theta}}_S(U_i)\|_2^2 + \|\widetilde{\boldsymbol{\theta}}_S(U_i) - \boldsymbol{\theta}_S^*(U_i)\|_2^2 \right) \\
&\lesssim 2\lambda_1 \left(\delta_s + sL_n^{-d} \right) = O(1).
\end{aligned}$$

Let $T_{i,jk} = B_k^*(U_i) \widetilde{\boldsymbol{\theta}}_S(U_i)^\top \left(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i) \right)_S (X_{ij} - \mu_{1j}(U_i))$ for $1 \le j \le p$ and $1 \le k \le L_n$, then we have

$$\left\| (\mathbf{I}_{11}^1 - \mathbf{I}_1^*)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \le L_n \max_{1\le k\le L_n} \left| \frac{1}{n} \sum_{i\in\mathcal{I}_1} T_{i,jk} - \mathbb{E}_1[T_{i,jk}] \right| . \tag{C.14}$$

In addition,

$$\begin{aligned}
&\mathbb{E}_1 \left\{ (T_{i,jk} - \mathbb{E}[T_{i,jk}])^2 \exp\left[\eta \left| T_{i,jk} - \mathbb{E}_1[T_{i,jk}] \right| \right] \right\} \\
\le& \mathbb{E}_1 \left\{ T_{i,jk}^2 \exp\left[\eta \left| T_{i,jk} - \mathbb{E}_1[T_{i,jk}] \right| \right] \right\} + [\mathbb{E}_1[T_{i,jk}]]^2 \mathbb{E}_1 \left\{ \exp\left[\eta \left| T_{i,jk} - \mathbb{E}_1[T_{i,jk}] \right| \right] \right\} \\
\le& \exp(\eta |\mathbb{E}_1[T_{i,jk}]|) \left(\mathbb{E}_1 \left\{ T_{i,jk}^2 \exp\left[\eta |T_{i,jk}| \right] \right\} + [\mathbb{E}_1[T_{i,jk}]]^2 \mathbb{E}_1 \left\{ \exp\left[\eta \left| T_{i,jk} \right| \right] \right\} \right) .
\end{aligned}$$

Recall that $\mathbb{E}_1[(X_{ij} - \mu_{1j}(U_i))(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i))|U_i] = \boldsymbol{\Sigma}_{j,S}(U_i)$, we have

$$\begin{aligned}
\mathbb{E}_1[T_{i,jk}] \le& \mathbb{E}_1 \left[B_k^*(U_i) \widetilde{\boldsymbol{\theta}}_S(U_i)^\top \left(\boldsymbol{X} - \boldsymbol{\mu}_1(U_i) \right)_S (X_{ij} - \mu_{1j}(U_i)) \right] \\
\le& \mathbb{E} \left[B_k^*(U_i) \left| \widetilde{\boldsymbol{\theta}}_S(U_i)^\top \boldsymbol{\Sigma}_{j,S}(U_i) \right| \right] \\
\le& \lambda_1 \sup_{u\in[0,1]} \|\widetilde{\boldsymbol{\theta}}_S(u)\|_2 \mathbb{E}\left[B_k^*(U_i) \right] \\
\lesssim& L_n^{-1}.
\end{aligned}$$

Denote $H_i = \widetilde{\boldsymbol{\theta}}_S(U_i)^\top (\boldsymbol{X} - \boldsymbol{\mu}_1(U_i))_S$. Applying the second assertion of Lemma A.2, for any $\eta >$ for sufficiently large $C > 0$, we have

$$\begin{aligned}
&\mathbb{E}_1 \left\{T_{i,jk}^2 \exp\left[\eta|T_{i,jk}|\right]\right\}\\
&\le \mathbb{E}_1 \left\{(B_k^*(U_i))^2(X_{ij} - \mu_{1j}(U_i))^2 H_i^2 \exp\left[\eta|(X_{ij} - \mu_{1j}(U_i))H_i|\right]\right\}\\
&\le \mathbb{E}_1 \left[(B_k^*(U_i))^2 \left(\mathbb{E}\left[(X_{ij} - \mu_{1j}(U_i))^4 e^{2\eta|X_{ij}-\mu_{1j}(U_i)|}\middle|U_i\right] \mathbb{E}\left[H_i^4 e^{2\eta|H_i|}\middle|U_i\right]\right)^{1/2}\right]\\
&\lesssim \mathbb{E}\left[(B_k^*(U_i))^2\right] \lesssim L_n^{-1}.
\end{aligned}$$

In fact, we also used the fact $\mathbb{E}[H^2(U_i)|U_i]$ and $\mathbb{E}[(X_{ij} - \mu_{1j}(U_i))^2]$ are both bounded. Moreover, for $\eta = 1/(2C\lambda_1)$ for sufficiently large $C > 0$, it holds

$$\begin{aligned}
\mathbb{E}_1\left[e^{\eta|T_{i,jk}|}\right] &\le \mathbb{E}_1\left[e^{\eta|(X_{ij}-\mu_{1j}(U_i))H_i|}\right]\\
&\le \mathbb{E}_1\left[e^{\eta(|X_{ij}-\mu_{1j}(U_i)|^2+H_i^2|)}\right]\\
&\le \mathbb{E}_1\left[\left(\mathbb{E}\left[e^{2\eta|X_{ij}-\mu_{1j}(U_i)|^2}\middle|U_i\right]\mathbb{E}_1\left[e^{2\eta|H_i|^2}\middle|U_i\right]\right)^{1/2}\right]\\
&= O(1).
\end{aligned}$$

Combing the results above, we conclude that

$$\mathbb{E}_1\left\{(T_{i,jk} - \mathbb{E}[T_{i,jk}])^2 \exp\left[\eta\left|T_{i,jk} - \mathbb{E}[T_{i,jk}]\right|\right]\right\} \lesssim L_n^{-1}.$$

According to Lemma A.1, we are guaranteed that

$$\mathbb{P}\left(\max_{j,k}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1} T_{i,jk} - \mathbb{E}_1[T_{i,jk}]\right| \lesssim \sqrt{\frac{\log p}{nL_n}}\right) \ge 1 - L_n p^{-\vartheta}.$$

In conjunction with (C.14), it follows that

$$\mathbb{P}\left(\max_{1\le j\le p}\left\|(\mathbf{I}_{11}^1 - \mathbf{I}_1^*)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2 \lesssim \sqrt{\frac{L_n \log p}{n}}\right) \ge 1 - L_n p^{-\vartheta}. \tag{C.15}$$

For $\mathbf{I}_{12}^1$ and $\mathbf{I}_{13}^1$, we have

$$\left\|(\mathbf{I}_{12}^1)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2 \le L_n \max_{1\le k\le L_n}\left|\frac{1}{n}\sum_{i=1}^n E_{i,jk}\right|, \tag{C.16}$$

and

$$\left\|(\mathbf{I}_{13}^1)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2 \le L_n \max_{1\le k\le L_n}\left|\frac{1}{n}\sum_{i=1}^n F_{i,jk}\right|, \tag{C.17}$$

where

$$E_{i,jk} = B_k^*(U_i)(X_{ij} - \mu_{1j}(U_i))\widetilde{\boldsymbol{\theta}}_S(U_i)^\top (\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))_S,$$

and

$$F_{i,jk} = B_k^*(U_i)(X_{ij} - \mu_{1j}(U_i))\widetilde{\boldsymbol{\theta}}_S(U_i)^\top (\boldsymbol{X} - \boldsymbol{\mu}_1(U_i))_S .$$

Due to the fact that $\sum_{l=1}^s (\widetilde{\boldsymbol{\nu}}_{(l)}^\top \boldsymbol{B}_i^*)^2 \le 1$, we can verify that

$$\max \left\{ \mathbb{E}_1 \left(E_{i,jk}^2 e^{\eta |E_{i,jk}|} \right)^2, \mathbb{E}_1 \left(F_{i,jk}^2 e^{\eta |F_{i,jk}|} \right)^2 \right\} \lesssim L_n^{-1}.$$

Combining with Lemma A.1, (C.16) and (C.17), we are guaranteed that

$$\mathbb{P}\left(\max_{1\le j\le p} \left\{ \left\| (\mathbf{I}_{12}^1)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 + \left\| (\mathbf{I}_{13}^1)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \right\} \le \sqrt{\frac{L_n \log p}{n}} \right) \ge 1 - L_n p^{-\vartheta}. \tag{C.18}$$

For

$$\begin{aligned} \mathbf{I}_{14}^1 - \mathbf{I}_3^*/2 = {} & \frac{1}{8n} \sum_{i\in\mathcal{I}_1} \left[(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top \right] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \\ & - \frac{1}{8}\mathbb{E} \left\{ \left[(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))(\boldsymbol{\mu}_1(U) - \boldsymbol{\mu}_2(U))^\top \right] \otimes (\boldsymbol{B}\boldsymbol{B}^\top) \right\}, \end{aligned}$$

it is easy to obtain that

$$\mathbb{P}\left(\max_{1\le j\le p} \left\| (\mathbf{I}_{14}^1 - \frac{1}{2}\mathbf{I}_3^*)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \lesssim \sqrt{\frac{L_n \log p}{n}} \right) \ge 1 - L_n p^{-\vartheta}. \tag{C.19}$$

Combining (C.16), (C.18) and (C.19), we have

$$\mathbb{P}\left(\max_{1\le j\le p} \left\| (\mathbf{I}_1^1 - \mathbf{I}_1^* - \frac{1}{2}\mathbf{I}_3^*)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \lesssim \sqrt{\frac{L_n \log p}{n}} \right) \ge 1 - L_n p^{-\vartheta}. \tag{C.20}$$

Similarly, we also have

$$\mathbb{P}\left(\max_{1\le j\le p} \left\| (\mathbf{I}_1^2 - \mathbf{I}_2^* - \frac{1}{2}\mathbf{I}_3^*)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)} \right\|_2 \lesssim \sqrt{\frac{L_n \log p}{n}} \right) \ge 1 - L_n p^{-\vartheta}. \tag{C.21}$$

Step 2. upper bounding $\max_{1\le j\le p} \|(\mathbf{I}_2^1 + \mathbf{I}_2^2 + \mathbf{I}_3^1 + \mathbf{I}_3^2 + \mathbf{I}_4^1 + \mathbf{I}_4^2)_{(j,S)} \widetilde{\boldsymbol{\gamma}}_{(S)}\|_2$ Recall that

$$\begin{aligned} \mathbf{I}_2^1 + \mathbf{I}_2^2 = {} & \frac{1}{2n} \sum_{i\in\mathcal{I}_1} \left\{ \left[(\boldsymbol{X}_i - \boldsymbol{\mu}(U_i))(\boldsymbol{\mu}(U_i) - \widehat{\boldsymbol{\mu}}(U_i))^\top \right] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\} \\ & + \frac{1}{2n} \sum_{i\in\mathcal{I}_2} \left\{ \left[(\boldsymbol{X}_i - \boldsymbol{\mu}(U_i))(\boldsymbol{\mu}(U_i) - \widehat{\boldsymbol{\mu}}(U_i))^\top \right] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\}, \end{aligned}$$

then we have

$$
\begin{aligned}
&\|(\mathbf{I}_2^1+\mathbf{I}_2^2)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\|_2\\
&\leq L_n \max_{1\leq k\leq L_n}\left|\frac{1}{2n}\sum_{i\in\mathcal{I}_1}(X_{ij}-\mu_{1j}(U_i))B_k^*(U_i)\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S\right|\\
&+L_n \max_{1\leq k\leq L_n}\left|\frac{1}{2n}\sum_{i\in\mathcal{I}_2}(X_{ij}-\mu_{2j}(U_i))B_k^*(U_i)\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S\right|\\
&+L_n \max_{1\leq k\leq L_n}\left|\frac{1}{4n}\sum_{i\in\mathcal{I}_1}G_{i,jk}(U_i)-\frac{1}{4n}\sum_{i\in\mathcal{I}_2}G_{i,jk}(U_i)\right|,
\end{aligned}
$$

where $G_{i,jk}(U_i)=(\mu_{1j}(U_i)-\mu_{2j}(U_i))B_k^*(U_i)\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S$. Notice that

$$
\begin{aligned}
\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S &= \sum_{l=1}^{s}(\widehat{\boldsymbol{\alpha}}_l-\widetilde{\boldsymbol{\alpha}}_l)^\top\boldsymbol{B}_i\widetilde{\theta}_l(U_i)\\
&+\sum_{l=1}^{s}(\widetilde{\boldsymbol{\alpha}}_l^\top\boldsymbol{B}_i-\mu_l(U_i))\widetilde{\theta}_l(U_i)\\
&=\widetilde{\boldsymbol{\theta}}(U_i)_S^\top\left(\widehat{\mathbf{M}}_{.S}-\widetilde{\mathbf{M}}_{.S}\right)^\top\boldsymbol{B}_i+\widetilde{\boldsymbol{\theta}}(U_i)_S^\top\left(\widetilde{\mathbf{M}}_{.S}^\top\boldsymbol{B}_i\right),
\end{aligned}
$$

where $\widehat{\mathbf{M}}_{.S}=(\widehat{\boldsymbol{\alpha}}_1,...,\widehat{\boldsymbol{\alpha}}_s)$ and $\widetilde{\mathbf{M}}_{.S}=(\widetilde{\boldsymbol{\alpha}}_1,...,\widetilde{\boldsymbol{\alpha}}_s)$. Then we have

$$
\begin{aligned}
&\left|\frac{1}{2n}\sum_{i\in\mathcal{I}_1}(X_{ij}-\mu_{1j}(U_i))B_k^*(U_i)\sum_{l=1}^{s}[\hat{\mu}_l(U_i)-\mu_l(U_i)]\widetilde{\theta}_l(U_i)\right|\\
&\leq\left|\frac{1}{2n}\sum_{i\in\mathcal{I}_1}(X_{ij}-\mu_{1j}(U_i))B_k^*(U_i)\widetilde{\boldsymbol{\theta}}(U_i)_S^\top\left(\widehat{\mathbf{M}}_{.S}-\widetilde{\mathbf{M}}_{.S}\right)^\top\boldsymbol{B}_i\right|\\
&+\left|\frac{1}{2n}\sum_{i\in\mathcal{I}_1}(X_{ij}-\mu_{1j}(U_i))B_k^*(U_i)\widetilde{\boldsymbol{\theta}}(U_i)_S^\top\left(\widetilde{\mathbf{M}}_{.S}^\top\boldsymbol{B}_i\right)\right|.
\end{aligned}
$$

From Proposition A.1, we know that $\|\widehat{\mathbf{M}}_{.S}-\widetilde{\mathbf{M}}_{.S}\|_F\lesssim\sqrt{s}a_n$. By utilizing the same chaining technique in Section D.3 to $\widehat{\mathbf{M}}_{.S}-\widetilde{\mathbf{M}}_{.S}$, we can show that

$$
\mathbb{P}\left(\max_{1\leq j\leq p}\left\|(\mathbf{I}_2^1+\mathbf{I}_2^2)_{(j,S)}\widetilde{\boldsymbol{\gamma}}_{(S)}\right\|_2\lesssim a_nL_ns\sqrt{\frac{\log p}{n}}\right)\geq 1-p^{-\vartheta sL_n}. \tag{C.22}
$$

Recall that

$$
\begin{aligned}
\mathbf{I}_3^1+\mathbf{I}_3^2=&\frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left\{\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}\\
&+\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left\{\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{X}_i-\boldsymbol{\mu}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\right\}.
\end{aligned}
$$

Similarly, we can verify

$$\mathbb{P}\left(\max_{1\le j\le p}\left\|(\mathbf{I}_3^1+\mathbf{I}_3^2)_{(j,S)}\widetilde{\gamma}_{(S)}\right\|_2 \lesssim a_nL_n\sqrt{\frac{\log p}{n}}\right) \ge 1-p^{-\vartheta L_n}. \tag{C.23}$$

For

$$\begin{aligned}\mathbf{I}_4^1+\mathbf{I}_4^2 &= \frac{1}{2n}\sum_{i\in\mathcal{I}_1}\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top)\\ &\quad+\frac{1}{2n}\sum_{i\in\mathcal{I}_2}\left[(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))(\boldsymbol{\mu}(U_i)-\widehat{\boldsymbol{\mu}}(U_i))^\top\right]\otimes(\boldsymbol{B}_i\boldsymbol{B}_i^\top),\end{aligned}$$

it follows from Proposition A.1 that

$$\begin{aligned}&\left\|(\mathbf{I}_4^1+\mathbf{I}_4^2)_{(j,S)}\widetilde{\gamma}_{(S)}\right\|_2\\ &\le L_n\max_{1\le k\le L_n,1\le i\le 2n}\left|B_k^*(U_i)[\hat{\mu}_j(U_i)-\mu_j(U_i)]\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S\right|\\ &\le L_na_n\max_{1\le i\le 2n}\left|\widetilde{\boldsymbol{\theta}}(U_i)_S^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S\right|\\ &\lesssim L_na_n\max_{1\le i\le 2n}\|(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))_S\|_2\,\|\widetilde{\boldsymbol{\theta}}(U_i)_S\|_2\\ &\le \sqrt{s}L_na_n^2,\end{aligned} \tag{C.24}$$

holds with probability at least $1-sL_np^{-\vartheta}$. Combining (C.20)-(C.24), we have

$$\max_{1\le j\le p}\|(\mathbf{D}_n-\mathbf{D})_{(j,S)}\widetilde{\gamma}_{(S)}\|_2 \lesssim \sqrt{\frac{L_n\log p}{n}}+a_nL_ns\sqrt{\frac{\log p}{n}},$$

holds with probability at least $1-sL_np^{-\vartheta}-L_np^{-\vartheta sL_n}-L_np^{-\vartheta}$. □

Appendix D: Proofs of auxiliary lemmas in Section C

D.1. Proof of Lemma C.1

Proof of Lemma C.1. We only prove the bound for $\mathcal{I}_1$, and the case in $\mathcal{I}_2$ is similar. Here we use $\mathbb{E}_1[\cdot]$ to denote the conditional expectation $\mathbb{E}[\cdot|Y=1]$.

Let $\mathbb{S}^{pL_n-1}$ be the unit sphere in $\mathbb{R}^{pL_n}$, we denote the 1/8-covering set of $\mathbb{S}^{pL_n-1}$ by $\{\boldsymbol{\nu}_1,\dots,\boldsymbol{\nu}_N\}$ with $N\le 17^{(pL_n)}$. It follows that for any $\boldsymbol{\nu}\in\mathbb{S}^{pL_n-1}$, there exist some $\boldsymbol{\nu}_l$ such that $\|\boldsymbol{\nu}-\boldsymbol{\nu}_l\|_2\le 1/8$. Let $\mathbf{Q}_i=\boldsymbol{Z}_i\boldsymbol{Z}_i^\top-\mathbb{E}_1[\boldsymbol{Z}_i\boldsymbol{Z}_i^\top]$, then we have

$$\left\|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\mathbf{Q}_i\right\|_2 \le 2\max_{1\le l\le N}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}_l^\top\mathbf{Q}_i\boldsymbol{\nu}_l\right|.$$

Note that

$$\begin{aligned}L_n^{-1}\boldsymbol{\nu}_l^\top\mathbf{Q}_i\boldsymbol{\nu}_l &= \left(\boldsymbol{\nu}_l^\top[(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))\otimes\boldsymbol{B}_i^*)]\right)^2-\mathbb{E}_1\left(\boldsymbol{\nu}_l^\top[(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))\otimes\boldsymbol{B}_i^*)]\right)^2\\ &= [\widetilde{\boldsymbol{\nu}}_{li}^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))]^2-\mathbb{E}_1[\widetilde{\boldsymbol{\nu}}_{li}^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))]^2\end{aligned}$$

where $\widetilde{\boldsymbol{\nu}}_l(U_i) = ((\boldsymbol{\nu}_l)_{(1)}^\top \boldsymbol{B}_i^*, ..., (\boldsymbol{\nu}_l)_{(p)}^\top \boldsymbol{B}_i^*)^\top$ for $\boldsymbol{\nu}_l = ((\boldsymbol{\nu}_l)_{(1)}^\top, ..., (\boldsymbol{\nu}_l)_{(p)}^\top)^\top$. Let $R_{li} = \widetilde{\boldsymbol{\nu}}_l(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))$, then $R_{li} \sim \mathcal{N}(0, \widetilde{\boldsymbol{\nu}}_l(U_i)^\top \boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}_l(U_i))$ given U_i and $Y_i = 1$. Let $\sigma_{li}^2 = \widetilde{\boldsymbol{\nu}}_l(U_i)^\top \boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}_l(U_i)$, then we have $\sigma_{li}^2 \le \lambda_2$

$$\begin{aligned}\mathbb{E}_1[\sigma_{li}^2] &\le \mathbb{E}_1\left[\|\widetilde{\boldsymbol{\nu}}_l(U_i)\|_2^2 \|\boldsymbol{\Sigma}(U_i)\|_2\right] \le \lambda_1 \mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}_l(U_i)\|_2^2\right] \\ &\le \lambda_1 \sum_{j=1}^p \mathbb{E}_1\left[\left(\boldsymbol{\nu}_{(j)}^\top \boldsymbol{B}_i^*\right)^2\right] \lesssim L_n^{-1} \sum_{j=1}^p \|\boldsymbol{\nu}_{(j)}\|_2^2 = L_n^{-1}.\end{aligned}$$

For $\eta = 1/(8\lambda_2 M_2)$, simple calculation gives $\mathbb{E}_1[e^{\eta R_{li}^2}] = O(1)$. Using Hölder's inequality, we also have

$$\begin{aligned}&\mathbb{E}_1\left[\left(R_{li}^2 - \mathbb{E}_1[R_{li}^2]\right)^2 \exp\left(\eta |R_{li}^2 - \mathbb{E}_1[R_{li}^2]|\right)\right] \\ \le& 2\exp\left(\eta \mathbb{E}_1[R_{li}^2]\right)\left(\mathbb{E}_1\left[R_{li}^4 e^{\eta R_{li}^2}\right] + \left(\mathbb{E}_1\left[R_{li}^2\right]\right)^2 \mathbb{E}_1\left[e^{\eta R_{li}^2}\right]\right) \\ \lesssim& \mathbb{E}_1\left[\sigma_{li}^4 \mathbb{E}_1\left[\frac{R_{li}^4}{\sigma_{li}^4} e^{\eta R_{li}^2} | U_i\right]\right] + \left(\mathbb{E}_1\left[\sigma_{li}^2\right]\right)^2 \\ \le& \mathbb{E}_1\left[\sigma_{li}^2 \left(\mathbb{E}_1\left[(R_{li}/\sigma_{li})^8 | U_i\right] \mathbb{E}_1\left[e^{2\eta R_{li}^2} | U_i\right]\right)^{1/2}\right] + L_n^{-2} \\ \lesssim& \mathbb{E}_1\left[\sigma_{li}^2\right] + L_n^{-2} \lesssim L_n^{-1}.\end{aligned}$$

Invoking Lemma A.1, we can obtain that

$$\mathbb{P}\left(\max_{1\le j\le N}\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\nu}_l^\top \mathbf{Q}_i \boldsymbol{\nu}_l\right| \lesssim L_n\sqrt{\frac{p\log n}{n}}\right) \ge 1 - n^{-\vartheta p L_n}.$$

Next we prove the conclusion for matrix-vector-product. It suffices to show that for any fixed $\boldsymbol{\nu} \in \mathbb{R}^{pL_n - 1}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1} \boldsymbol{\nu}^\top\left(\boldsymbol{Z}_i\boldsymbol{Z}_i^\top - \mathbb{E}_1\left[\boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right]\right)\widetilde{\boldsymbol{\gamma}}\right| \lesssim \sqrt{\frac{L_n p\log n}{n}}\right) \le n^{-\vartheta p L_n}. \tag{D.1}$$

Let $\widetilde{\boldsymbol{\nu}}(U_i) = (\boldsymbol{\nu}_{(1)}^\top \boldsymbol{B}_i^*, ..., \boldsymbol{\nu}_{(p)}^\top \boldsymbol{B}_i^*)^\top$, then notice that

$$\begin{aligned}\boldsymbol{\nu}^\top \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \widetilde{\boldsymbol{\gamma}} &= \boldsymbol{\nu}^\top\left((\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)) \otimes \boldsymbol{B}_i\right)\widetilde{\boldsymbol{\gamma}}^\top\left((\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)) \otimes \boldsymbol{B}_i\right) \\ &= \sqrt{L_n}\left(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)\right)^\top \widetilde{\boldsymbol{\nu}}(U_i)\left(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)\right)^\top \widetilde{\boldsymbol{\theta}}(U_i).\end{aligned}$$

Denote $R_i = \left(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)\right)^\top \widetilde{\boldsymbol{\nu}}(U_i)$ and $S_i = \left(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)\right)^\top \widetilde{\boldsymbol{\theta}}(U_i)$, then

$$R_i | U_i \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}(U_i)\right), \quad S_i | U_i \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{\theta}}(U_i)^\top \boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\theta}}(U_i)\right).$$

Also, we know that $\mathbb{E}_1[R_i^2|U_i] \le \lambda_1$ and $\mathbb{E}_1[S_i^2|U_i] \lesssim \lambda_1$ since $\|\widetilde{\boldsymbol{\nu}}(u)\|_2 \le 1$ and

$$\sup_{u\in[0,1]} \|\widetilde{\boldsymbol{\theta}}(u)\|_2 \lesssim \|\boldsymbol{\theta}^*(u)\|_2 + \sqrt{p}L_n^{-d} = O(1),$$

which is true due to the assumptions $\sup_{u\in[0,1]}\|\boldsymbol{\theta}^*(u)\|_2 = O(1)$ and $\sqrt{p}L_n^{-d} = o(1)$. Using Hölder's inequality, we have

$$\begin{aligned}
&\mathbb{E}_1\left[(R_iS_i - \mathbb{E}_1[R_iS_i])^2 e^{\eta|R_iS_i - \mathbb{E}_1[R_iS_i]|}\right] \\
\leq& 2\exp\left(\eta|\mathbb{E}_1[R_iS_i]|\right)\left(\mathbb{E}_1\left[R_i^2S_i^2e^{\eta|R_iS_i|}\right] + \left(\mathbb{E}_1\left[R_iS_i\right]\right)^2\mathbb{E}_1\left[e^{\eta|R_iS_i|}\right]\right) \\
\lesssim& \mathbb{E}_1\left[R_i^2S_i^2e^{\eta|R_iS_i|}\right] + \mathbb{E}_1[R_i^2]\mathbb{E}_1[S_i^2]\mathbb{E}_1\left[e^{\eta|R_iS_i|}\right]. \qquad \text{(D.2)}
\end{aligned}$$

Denote $\sigma_i^2 = \widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}(U_i)$. Then applying Lemma A.2 and moment formula of normal distribution, it holds that

$$\begin{aligned}
\mathbb{E}_1\left[R_i^4e^{2\eta|R_i|}\middle|U_i\right] &\leq 32e^{2\eta^2\sigma_i^2}\left((\eta\sigma_i^2)^4 + 6(\eta\sigma_i^2)^2\sigma_i^2 + 3\sigma_i^4\right) \\
&\lesssim \sigma_i^4 = \left(\widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}(U_i)\right)^2,
\end{aligned}$$

where the second inequality holds since σ_i is bounded. Similarly, we can also verify that

$$\mathbb{E}_1\left[S_i^4e^{2\eta|S_i|}\middle|U_i\right] \lesssim \left(\widetilde{\boldsymbol{\theta}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\theta}}(U_i)\right)^2.$$

It follows from $\widetilde{\boldsymbol{\theta}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\theta}}(U_i)$ is bounded that

$$\begin{aligned}
\mathbb{E}_1\left[R_i^2S_i^2e^{\eta|R_iS_i|}\right] &= \mathbb{E}_1\left[\mathbb{E}_1\left[R_i^2S_i^2e^{\eta|R_iS_i|}\middle|U_i\right]\right] \\
&\leq \mathbb{E}_1\left[\left(\mathbb{E}_1\left[R_i^4e^{2\eta|R_i|}\middle|U_i\right]\mathbb{E}_1\left[S_i^4e^{2\eta|S_i|}\middle|U_i\right]\right)^{1/2}\right] \\
&\lesssim \mathbb{E}_1\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}(U_i)\right)\left(\widetilde{\boldsymbol{\theta}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\theta}}(U_i)\right)\right] \\
&\lesssim \mathbb{E}_1\left[\widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\nu}}(U_i)\right] \\
&\lesssim \mathbb{E}_1\left[\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2^2\right] \lesssim L_n^{-1}. \qquad \text{(D.3)}
\end{aligned}$$

Then we take $\eta = 1/(C\lambda_1)$ for sufficiently large $C > 0$, it holds that

$$\mathbb{E}_1\left[e^{\eta|R_iS_i|}\right] \leq \left(\mathbb{E}_1\left[e^{\eta(R_i^2+S_i^2)}\right]\right) \leq \left(\mathbb{E}_1\left[e^{2\eta R_i^2}\right]\mathbb{E}_1\left[e^{2\eta S_i^2}\right]\right)^{1/2} = O(1). \qquad \text{(D.4)}$$

Substituting (D.3) and (D.4) into (D.2), we have

$$\mathbb{E}_1\left[(R_iS_i - \mathbb{E}_1[R_iS_i])^2e^{\eta|R_iS_i - \mathbb{E}_1[R_iS_i]|}\right] \lesssim L_n^{-1}.$$

Applying Lemma A.1 again, we can prove (D.1) immediately. □

D.2. Proof of Lemma C.2 and C.3

Proof. The proofs of Lemma C.2 and C.3 are similar, here we only prove Lemma C.2.

For the first assertion for operator norm of matrix, it suffices to show for any fixed $\boldsymbol{\nu} \in \mathbb{S}^{pL_n-1}$ and $\boldsymbol{\xi} \in \mathbb{S}^{pL_n-1}$ such that

$$\left| \frac{1}{n} \sum_{i \in \mathcal{I}_1} \boldsymbol{\nu}^\top \left\{ [(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\} \boldsymbol{\xi} \right| \lesssim L_n \sqrt{\frac{p \log n}{n}},$$

holds with probability at least $1 - n^{-\vartheta p L_n}$. For $\boldsymbol{\nu} = (\boldsymbol{\nu}_{(1)}^\top, ..., \boldsymbol{\nu}_{(p)}^\top)$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_{(1)}^\top, ..., \boldsymbol{\xi}_{(p)}^\top)$, we write

$$\widetilde{\boldsymbol{\nu}}(U_i) = (\boldsymbol{\nu}_{(1)}^\top \boldsymbol{B}_i^*, ..., \boldsymbol{\nu}_{(p)}^\top \boldsymbol{B}_i^*)^\top, \quad \widetilde{\boldsymbol{\xi}}(U_i) = (\boldsymbol{\xi}_{(1)}^\top \boldsymbol{B}_i^*, ..., \boldsymbol{\xi}_{(p)}^\top \boldsymbol{B}_i^*)^\top.$$

Then we have

$$\begin{aligned} &\boldsymbol{\nu}^\top \left\{ [(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\} \boldsymbol{\xi} \\ &\quad = L_n (\widetilde{\boldsymbol{\xi}}(U_i)^\top (\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i)))[\widetilde{\boldsymbol{\nu}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))]. \end{aligned}$$

Let

$$T_i = (\widetilde{\boldsymbol{\xi}}(U_i)^\top (\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i)))[\widetilde{\boldsymbol{\nu}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))],$$

and $\sigma_i^2 = \widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\Sigma}(U_i) \widetilde{\boldsymbol{\nu}}(U_i)$, then for $\eta > 0$ it holds that

$$\begin{aligned} \mathbb{E}_1[T_i^2 e^{\eta |T_i|}] &\lesssim \mathbb{E}_1 \left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)) \right)^2 e^{\eta |\widetilde{\boldsymbol{\nu}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))|} \right] \\ &\le 2\mathbb{E} \left[e^{\frac{\eta^2 \sigma_i^2}{2}} \left(\sigma_i^2 + \phi^2 \sigma_i^4 \right) \right] \\ &\lesssim \mathbb{E} \left[\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2^2 \right] \lesssim L_n^{-1}, \end{aligned}$$

where the first inequality follows from $|\widetilde{\boldsymbol{\xi}}(u)^\top (\boldsymbol{\mu}_1(u) - \boldsymbol{\mu}_2(u))|$ is uniformly bounded. By Lemma A.1, we are guaranteed that

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i \in \mathcal{I}_1} T_i \right| \lesssim \sqrt{\frac{p \log n}{n}} \right) \ge 1 - n^{-\vartheta p L_n}.$$

For the assertion for matrix-vector-product, it suffices to show

$$\left| \frac{1}{n} \sum_{i \in \mathcal{I}_1} \boldsymbol{\nu}^\top \left\{ [(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\} \widetilde{\gamma} \right| \lesssim \sqrt{\frac{p L_n \log n}{n}}$$

holds with probability at least $1 - n^{-\vartheta p L_n}$. Notice that

$$\begin{aligned} &\boldsymbol{\nu}^\top \left\{ [(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))(\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i))^\top] \otimes (\boldsymbol{B}_i \boldsymbol{B}_i^\top) \right\} \widetilde{\gamma} \\ = &\sqrt{L_n} \left(\widetilde{\boldsymbol{\nu}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i)) \right) \left(\widetilde{\boldsymbol{\theta}}(U_i)^\top (\boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i)) \right), \end{aligned}$$

and

$$\left|\widetilde{\boldsymbol{\theta}}(U_i)^\top(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))\right| \le \delta_p \sup_{u\in[0,1]} \|\widetilde{\boldsymbol{\theta}}(u)\|_2 = O(1).$$

Denote $R_i = \left(\widetilde{\boldsymbol{\nu}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))\right)\left(\widetilde{\boldsymbol{\theta}}(U_i)^\top(\boldsymbol{\mu}_1(U_i)-\boldsymbol{\mu}_2(U_i))\right)$. Then we can get

$$\begin{aligned}\mathbb{E}_1\left[R_i^2 e^{\eta|R_i|}\right] &\lesssim \mathbb{E}_1\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))\right)^2 e^{\eta|\widetilde{\boldsymbol{\nu}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))|}\right]\\ &\lesssim L_n^{-1}.\end{aligned}$$

Applying Lemma A.1, we can prove the desired result. □

D.3. Proof of Lemma C.4

Proof. Let $\widetilde{\mathbf{M}} = (\widetilde{\mathbf{M}}_1, ..., \widetilde{\mathbf{M}}_p) \in \mathbb{R}^{L_n\times p}$ where the j-th column $\mathbf{M}_j = \frac{1}{2}(\widetilde{\boldsymbol{\alpha}}_{1j} + \widetilde{\boldsymbol{\alpha}}_{2j})$ and

$$\widetilde{\boldsymbol{\alpha}}_{1j} = \left(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right)^{-1}\mathbb{E}[\boldsymbol{B}X_j|Y=1], \quad \widetilde{\boldsymbol{\alpha}}_{1j} = \left(\mathbb{E}[\boldsymbol{B}\boldsymbol{B}^\top]\right)^{-1}\mathbb{E}[\boldsymbol{B}X_j|Y=0].$$

Similarly, we denote $\widehat{\mathbf{M}} = (\widehat{\mathbf{M}}_1, ..., \widehat{\mathbf{M}}_p) \in \mathbb{R}^{L_n\times p}$, where the j-th column $\widehat{\mathbf{M}}_j = \frac{1}{2}(\widehat{\boldsymbol{\alpha}}_{1j} + \widehat{\boldsymbol{\alpha}}_{2j})$ and

$$\widehat{\boldsymbol{\alpha}}_{1j} = \left(\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{B}_iX_{ij}, \ \widehat{\boldsymbol{\alpha}}_{2j} = \left(\frac{1}{n}\sum_{i\in\mathcal{I}_2}\boldsymbol{B}_i\boldsymbol{B}_i^\top\right)^{-1}\frac{1}{n}\sum_{i\in\mathcal{I}_2}\boldsymbol{B}_iX_{ij}.$$

Recalling the approximation error bound in the proof of Proposition A.1:

$$\sup_{u\in[0,1]}|\widetilde{\boldsymbol{\alpha}}_{1j}^\top\boldsymbol{B}(u) - \mu_{1j}(u)| \lesssim L_n^{-d}, \quad \sup_{u\in[0,1]}|\widetilde{\boldsymbol{\alpha}}_{2j}^\top\boldsymbol{B}(u) - \mu_{2j}(u)| \lesssim L_n^{-d}.$$

It means that $\|\widetilde{\mathbf{M}}^\top\boldsymbol{B}_i - \boldsymbol{\mu}(U_i)\|_2 \lesssim \sqrt{p}L_n^{-d}$. In addition, we define the good event:

$$\mathcal{A} := \left\{\|\widehat{\mathbf{M}}_j - \widetilde{\mathbf{M}}_j\|_2 \lesssim a_n : \text{ for } 1\le j\le p\right\},$$

where $a_n = \sqrt{\frac{L_n\log n}{n}} + L_n^{-d}$. By Proposition A.1, we know that $\mathbb{P}(\mathcal{A}^c) \le L_n n^{-\vartheta}$. Next we prove Lemma C.3 under the good event $\mathcal{A}$.

We first prove the bound for the operator norm of the matrix. Let $\mathbf{A}_i = [(\widehat{\boldsymbol{\mu}}(U_i) - \boldsymbol{\mu}(U_i))(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))^\top] \otimes (\boldsymbol{B}_i\boldsymbol{B}_i^\top)$. According to the proof of Lemma C.1, it suffices to show that for any fixed $\boldsymbol{\nu}, \boldsymbol{\xi} \in \mathbb{S}^{pL_n-1}$ such that

$$\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}^\top\mathbf{A}_i\boldsymbol{\xi}\right| \lesssim a_n L_n^{3/2} p\sqrt{\frac{\log n}{n}}$$

holds with probability at least $1 - n^{-\vartheta p L_n}$. Notice that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}^\top \mathbf{A}_i\boldsymbol{\xi}\right| &= L_n \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\nu}}_i^\top(\widehat{\boldsymbol{\mu}}(U_i)-\boldsymbol{\mu}(U_i))][\widetilde{\boldsymbol{\xi}}_i^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))]\right| \\ &\le L_n \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\widetilde{\mathbf{M}}^\top\boldsymbol{B}_i-\boldsymbol{\mu}(U_i))]\right| \\ &\quad + L_n \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\widehat{\mathbf{M}}-\widetilde{\mathbf{M}})^\top\boldsymbol{B}_i]\right|,\end{aligned} \tag{D.5}$$

where $\widetilde{\boldsymbol{\nu}}(U_i) = (\boldsymbol{\nu}_{(1)}^\top\boldsymbol{B}_i^*, ..., \boldsymbol{\nu}_{(p)}^\top\boldsymbol{B}_i^*)^\top$ and $\widetilde{\boldsymbol{\xi}}(U_i) = (\boldsymbol{\xi}_{(1)}^\top\boldsymbol{B}_i^*, ..., \boldsymbol{\xi}_{(p)}^\top\boldsymbol{B}_i^*)^\top$ for $\boldsymbol{\nu} = (\boldsymbol{\nu}_{(1)}^\top, ..., \boldsymbol{\nu}_{(p)}^\top)^\top$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_{(1)}^\top, ..., \boldsymbol{\xi}_{(p)}^\top)^\top$. Also, we know that

$$\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))|U_i \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)\right).$$

Denote $T_i = [\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\widetilde{\mathbf{M}}^\top\boldsymbol{B}_i-\boldsymbol{\mu}(U_i))]$. Apply the second assertion in Lemma A.2 with any constant $\eta > 0$, we get

$$\begin{aligned}&\mathbb{E}_1\left[T_i^2 e^{\eta|T_i|}\right] \le \mathbb{E}_1\left[T_i^2 e^{\eta|\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))|}\right] \\ &\lesssim pL_n^{-2d}\mathbb{E}_1\left[\left(\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))\right)^2 e^{\eta|\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))|}\right] \\ &\le 2pL_n^{-2d}\mathbb{E}\left[e^{\frac{\eta^2\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)}{2}}\left(\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)+\eta^2\left(\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)\right)^2\right)\right] \\ &\lesssim pL_n^{-2d}\mathbb{E}\left[\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)\right] \lesssim pL_n^{-2d}L_n^{-1},\end{aligned}$$

where we also used the fact

$$\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i) \le \lambda_1\|\widetilde{\boldsymbol{\xi}}(U_i)\|_2^2 \le \lambda_1\|\boldsymbol{B}_i^*\|_2^2\sum_{j=1}^p\|\boldsymbol{\nu}_{(j)}\|_2^2 \le \lambda_1.$$

Applying Lemma A.1, we can verify that

$$\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\widetilde{\mathbf{M}}^\top\boldsymbol{B}_i-\boldsymbol{\mu}(U_i))]\right| \lesssim pL_n^{-d}\sqrt{\frac{L_n\log n}{n}}, \tag{D.6}$$

holds with probability at least $1-n^{-\vartheta p L_n}$. Next we proceed to bound the second term in (D.5). We define a matrix set as

$$\boldsymbol{\Xi} = \left\{\mathbf{M} \in \mathbb{R}^{L_n\times p}:\ \|\mathbf{M}_j\|_2 \le a_n \text{ for } j = 1, 2, ..., p\right\},$$

where $a_n = O(\sqrt{L_n \log n/n} + L_n^{-d})$ and $\mathbf{M}_j$ is the j-th column of $\mathbf{M}$. For each $1 \le j \le p$, we may find a set $\{\boldsymbol{\zeta}^\ell \in \mathbb{R}^{L_n}, 1 \le \ell \le n^{ML_n}:\ \|\boldsymbol{\zeta}^\ell\|_2 \le a_n\}$ such that there exists some $1 \leq \ell \leq n^{ML_n}$ satisfying that $\|\mathbf{M}_j - \boldsymbol{\zeta}^\ell\|_2 \leq n^{-M}\sqrt{L_n}a_n$[1]. It means that we can find a subset $\boldsymbol{\Xi}' = \{\mathbf{M}^\ell : 1 \leq \ell \leq n^{MpL_n}\} \subseteq \boldsymbol{\Xi}$. And for any $\mathbf{M} \in \boldsymbol{\Xi}$, there exists some $1 \leq \ell \leq n^{MpL_n}$ such that $\|\mathbf{M}_j^\ell - \mathbf{M}_j\|_2 \leq a_n\sqrt{L_n}n^{-M}$. It follows that for any $\mathbf{M} \in \boldsymbol{\Xi}$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top\mathbf{M}^\top\boldsymbol{B}_i]\right| \\
&\quad \leq \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\mathbf{M}^\ell)^\top\boldsymbol{B}_i]\right| \\
&\qquad + \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\mathbf{M}^\ell - \mathbf{M})^\top\boldsymbol{B}_i]\right| \\
&\quad \leq \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\mathbf{M}^\ell)^\top\boldsymbol{B}_i]\right| \\
&\qquad + a_n\sqrt{p}L_n n^{-M}\max_{i\in\mathcal{I}_1}\left|\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))\right|,
\end{aligned}$$

where the last inequality is true since

$$\left|\widetilde{\boldsymbol{\nu}}(U_i)^\top(\mathbf{M}^\ell)^\top\boldsymbol{B}_i\right| \leq \|\widetilde{\boldsymbol{\nu}}(U_i)\|_2\left(\sum_{j=1}^p((\mathbf{M}_j^\ell - \mathbf{M}_j)^\top\boldsymbol{B}_i)^2\right)^{1/2} \leq \sqrt{p}L_n a_n n^{-M}.$$

Let $V_{\ell,i} = [\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\mathbf{M}^\ell)^\top\boldsymbol{B}_i^*]$. Then we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top(\widehat{\mathbf{M}} - \widetilde{\mathbf{M}})^\top\boldsymbol{B}_i]\right| \\
&\leq \sup_{\mathbf{M}\in\boldsymbol{\Xi}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}[\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top\mathbf{M}^\top\boldsymbol{B}_i]\right| \qquad \text{(D.7)} \\
&\leq \max_{1\leq\ell\leq n^{MpL_n}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}V_{\ell,i}\right| + a_n\sqrt{p}L_n n^{-M}\sqrt{\log n},
\end{aligned}$$

where the last inequality follows from the bound for the maximal of n independent Gaussian random variables and the variance of $\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))$ is bounded almost surely. Notice that

$$\widetilde{\boldsymbol{\xi}}(U_i)^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))|U_i \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)\right),$$

where $\mathbb{E}[\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)] \lesssim L_n^{-1}$ and $\widetilde{\boldsymbol{\xi}}(U_i)^\top\boldsymbol{\Sigma}(U_i)\widetilde{\boldsymbol{\xi}}(U_i)$ is almost surely bounded. Using the second assertion of Lemma A.2 with any $\eta > 0$, together

[1] For each coordinate of $\mathbf{M}_j$, we can divide the interval $[-a_n, a_n]$ into n^M small intervals with equal length $2a_n/N^M$.

with $\|\mathbf{M}^\ell \widetilde{\boldsymbol{\nu}}(U_i)\|_2 \leq \|\mathbf{M}^\ell\|_F \|\widetilde{\boldsymbol{\nu}}(U_i)\|_2 \leq \sqrt{p} a_n$, we can verify that

$$\mathbb{E}\left[V_i^2 e^{\eta|V_i|}\right] \lesssim p a_n^2 L_n^{-1},$$

Applying Lemma A.1, it is easy to show

$$\max_{1\leq \ell \leq n^{MpL_n}} \left| \frac{1}{n} \sum_{i\in\mathcal{I}_1} [\widetilde{\boldsymbol{\xi}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top (\mathbf{M}^\ell)^\top \boldsymbol{B}_i] \right| \lesssim p a_n \sqrt{\frac{L_n \log n}{n}} \tag{D.8}$$

with probability at least $1 - n^{-\vartheta L_n p}$. By choosing sufficiently large M and substituting (D.6)-(D.8) into (D.5), we can finish the proof.

For the case in the matrix-vector-product, we notice that

$$\begin{aligned}\boldsymbol{\nu}^\top \mathbf{A}_i \widetilde{\boldsymbol{\gamma}} &= \sqrt{L_n}[\widetilde{\boldsymbol{\theta}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top (\widetilde{\mathbf{M}}^\top \boldsymbol{B}_i - \boldsymbol{\mu}(U_i))] \\ &\quad + \sqrt{L_n}[\widetilde{\boldsymbol{\theta}}(U_i)^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_1(U_i))][\widetilde{\boldsymbol{\nu}}(U_i)^\top (\widehat{\mathbf{M}} - \widetilde{\mathbf{M}})^\top \boldsymbol{B}_i].\end{aligned}$$

By utilizing the same chaining technique and applying Lemma A.1, we can prove

$$\left| \frac{1}{n} \sum_{i\in\mathcal{I}_1} \boldsymbol{\nu}^\top \mathbf{A}_i \widetilde{\boldsymbol{\gamma}} \right| \lesssim a_n L_n p \sqrt{\frac{\log n}{n}},$$

holds with probability at least $1 - n^{-\vartheta p L_n}$. Then we can finish the proof. □

D.4. Proof of Lemma C.5

Proof. We prove Lemma C.5 under the good event $\mathcal{A}$ defined in Section D.4. We prove the bound for the operator norm of the matrix, and ℓ_2 norm bound for the matrix-vector-product is similar. It suffices to show that for any fixed $\boldsymbol{\nu}, \boldsymbol{\xi} \in \mathbb{S}^{pL_n - 1}$ such that

$$\left| \frac{1}{n} \sum_{i\in\mathcal{I}_1} \boldsymbol{\nu}^\top \mathbf{G}(U_i) \boldsymbol{\xi} - \frac{1}{n} \sum_{i\in\mathcal{I}_2} \boldsymbol{\nu}^\top \mathbf{G}(U_i) \boldsymbol{\xi} \right| \leq C L_n^{3/2} p \sqrt{\frac{\log n}{n}} \left(\sqrt{\frac{L_n \log n}{n}} + L_n^{-d} \right),$$

holds with probability at least $1 - n^{-\vartheta p L_n}$. To simplify notations, we denote $\boldsymbol{\delta}(U_i) = \boldsymbol{\mu}_1(U_i) - \boldsymbol{\mu}_2(U_i)$ and write $\widehat{\boldsymbol{\mu}}(u) = \widehat{\mathbf{M}}^\top \boldsymbol{B}(u)$, where the j-th column of $\widehat{\mathbf{M}}$ equals to $(\widehat{\boldsymbol{\alpha}}_{1j} + \widehat{\boldsymbol{\alpha}}_{2j})/2$. Recall the definition of $\mathbf{G}(U_i)$, we have

$$\begin{aligned}\mathbf{G}(U_i) &= \left(\boldsymbol{\delta}(U_i) \left(\widehat{\boldsymbol{\mu}}(U_i) - \boldsymbol{\mu}(U_i)\right)^\top\right) \otimes \left(\boldsymbol{B}_i \boldsymbol{B}_i^\top\right) \\ &= \underbrace{\left(\boldsymbol{\delta}(U_i) \left(\widehat{\mathbf{M}}^\top \boldsymbol{B}(U_i) - \widetilde{\mathbf{M}}^\top \boldsymbol{B}(U_i)\right)^\top\right) \otimes \left(\boldsymbol{B}_i \boldsymbol{B}_i^\top\right)}_{\mathbf{G}_1(U_i)} \\ &\quad + \underbrace{\left(\boldsymbol{\delta}(U_i) \left(\widetilde{\mathbf{M}}^\top \boldsymbol{B}(U_i) - \boldsymbol{\mu}(U_i)\right)^\top\right) \otimes \left(\boldsymbol{B}_i \boldsymbol{B}_i^\top\right)}_{\mathbf{G}_2(U_i)}.\end{aligned}$$

It yields that

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}^\top\mathbf{G}(U_i)\boldsymbol{\xi}-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\boldsymbol{\nu}^\top\mathbf{G}(U_i)\boldsymbol{\xi}\right|\\
&\leq\underbrace{\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}^\top\mathbf{G}_1(U_i)\boldsymbol{\xi}-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\boldsymbol{\nu}^\top\mathbf{G}_1(U_i)\boldsymbol{\xi}\right|}_{\Pi_1}\\
&+\underbrace{\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\boldsymbol{\nu}^\top\mathbf{G}_2(U_i)\boldsymbol{\xi}-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\boldsymbol{\nu}^\top\mathbf{G}_1(U_i)\boldsymbol{\xi}\right|}_{\Pi_2}.
\end{aligned}
\tag{D.9}
$$

Denote $\widetilde{\boldsymbol{\nu}}(U_i)=(\boldsymbol{\nu}_{(1)}^\top\boldsymbol{B}_i^*,...,\boldsymbol{\nu}_{(p)}^\top\boldsymbol{B}_i^*)^\top$ and $\widetilde{\boldsymbol{\xi}}(U_i)=(\boldsymbol{\xi}_{(1)}^\top\boldsymbol{B}_i^*,...,\boldsymbol{\xi}_{(p)}^\top\boldsymbol{B}_i^*)^\top$. Then we have

$$
\boldsymbol{\nu}^\top\mathbf{G}_1(U_i)\boldsymbol{\xi}=\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\delta}(U_i)\right)\left(\widetilde{\boldsymbol{\xi}}(U_i)^\top\left(\widehat{\mathbf{M}}-\widetilde{\mathbf{M}}\right)^\top\boldsymbol{B}_i\right)=:T_i\left(\widehat{\mathbf{M}}-\widetilde{\mathbf{M}}\right).
$$

Here we use the same notation $\boldsymbol{\Xi}$ in Section D.3. From Proposition A.1, we know that

$$
\begin{aligned}
\Pi_1&=\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}T_i\left(\widehat{\mathbf{M}}-\widetilde{\mathbf{M}}\right)-\frac{1}{n}\sum_{i\in\mathcal{I}_2}T_i\left(\widehat{\mathbf{M}}-\widetilde{\mathbf{M}}\right)\right|\\
&\leq\sup_{\mathbf{M}\in\boldsymbol{\Xi}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}T_i\left(\mathbf{M}\right)-\frac{1}{n}\sum_{i\in\mathcal{I}_2}T_i\left(\mathbf{M}\right)\right|\\
&=\sup_{\mathbf{M}\in\boldsymbol{\Xi}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}\left(T_i\left(\mathbf{M}\right)-\mathbb{E}\left[T_i\left(\mathbf{M}\right)\right]\right)-\frac{1}{n}\sum_{i\in\mathcal{I}_2}\left(T_i\left(\mathbf{M}\right)-\mathbb{E}\left[T_i\left(\mathbf{M}\right)\right]\right)\right|\\
&\leq\underbrace{\sup_{\mathbf{M}\in\boldsymbol{\Xi}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}T_i(\mathbf{M})-\mathbb{E}\left[T_i(\mathbf{M})\right]\right|}_{\Pi_{11}}+\underbrace{\sup_{\mathbf{M}\in\boldsymbol{\Xi}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_2}T_i(\mathbf{M})-\mathbb{E}\left[T_i(\mathbf{M})\right]\right|}.
\end{aligned}
\tag{D.10}
$$

In fact, the second equality holds since the randomness of $T_i(\mathbf{M})$ is from U_i and U_i is independent of Y_i. Using the facts $\|\boldsymbol{B}_i\|_2\leq\sqrt{L_n}$, $\|\boldsymbol{\delta}(U_i)\|_2\leq\delta_p$, $\|\widetilde{\boldsymbol{\xi}}(U_i)\|_2\leq 1$ and $\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2\leq 1$, and following the chaining arguments in Section D.3, we have

$$
\Pi_{11}\leq\max_{1\leq\ell\lesssim n^{MpL_n}}\left|\frac{1}{n}\sum_{i\in\mathcal{I}_1}T_i\left(\mathbf{M}^\ell\right)-\mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]\right|+a_n\sqrt{p}L_nn^{-M},
\tag{D.11}
$$

where $a_n=O(\sqrt{L_n\log n/n}+L_n^{-d})$. In addition, notice that

$$
\left(\mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]\right)^2\leq\mathbb{E}[(\widetilde{\boldsymbol{\nu}}(U_i)^\top\boldsymbol{\delta}(U_i))^2]\mathbb{E}[(\widetilde{\boldsymbol{\xi}}(U_i)^\top(\mathbf{M}^\ell)^\top\boldsymbol{B}_i^*)^2]
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2^2 \|\boldsymbol{\delta}(U_i)\|_2^2\right] \mathbb{E}\left[\|\mathbf{M}^\ell \widetilde{\boldsymbol{\xi}}(U_i)\|_2^2 \|\boldsymbol{B}_i^*\|_2^2\right] \\
&\leq p\delta_p^2 a_n^2 \mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2^2\right] \\
&\leq p\delta_p^2 a_n^2 \mathbb{E}\left[\sum_{j=1}^{p} \|\boldsymbol{\nu}_j\|_2^2 \|\boldsymbol{B}_i^*\|_2^2\right] \\
&\lesssim p\delta_p^2 a_n^2 L_n^{-1},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[T_i^2\left(\mathbf{M}^\ell\right)\right] &= \mathbb{E}\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\delta}(U_i)\right)^2 \left(\widetilde{\boldsymbol{\xi}}(U_i)^\top (\mathbf{M}^\ell)^\top \boldsymbol{B}_i^*\right)^2\right] \\
&\leq p a_n^2 \mathbb{E}\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\delta}(U_i)\right)^2\right] \\
&\lesssim p\delta_p^2 a_n^2 L_n^{-1}.
\end{aligned}
$$

In fact, we used $\|\mathbf{M}^\ell \widetilde{\boldsymbol{\xi}}(U_i)\|_2^2 = \|\mathbf{M}^\ell\|_F^2 \|\widetilde{\boldsymbol{\xi}}(U_i)\|_2^2 \leq p a_n$ and $\|\boldsymbol{B}_i\|_2^2 \leq L_n$ in the relations above. Thus we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(T_i\left(\mathbf{M}^\ell\right) - \mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]\right)^2 e^{\eta|T_i\left(\mathbf{M}^\ell\right) - \mathbb{E}[T_i\left(\mathbf{M}^\ell\right)]|}\right] \\
\leq& 2e^{\eta|\mathbb{E}[T_i\left(\mathbf{M}^\ell\right)]|}\left\{\mathbb{E}\left[T_i^2\left(\mathbf{M}^\ell\right) e^{\eta|T_i\left(\mathbf{M}^\ell\right)|}\right] + \left(\mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]\right)^2 \mathbb{E}[e^{\eta|\mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]|}]\right\} \\
\lesssim& \mathbb{E}\left[T_i^2\left(\mathbf{M}^\ell\right)\right] + \left(\mathbb{E}\left[T_i\left(\mathbf{M}^\ell\right)\right]\right)^2 \\
\leq& 2\mathbb{E}\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\delta}(U_i)\right)^2 \left(\widetilde{\boldsymbol{\xi}}(U_i)^\top \left(\mathbf{M}^\ell\right)^\top \boldsymbol{B}_i\right)^2\right] \\
\lesssim& pL_n a_n^2 \mathbb{E}\left[\left(\widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\delta}(U_i)\right)^2\right] \leq pL_n \delta_p^2 a_n^2 \mathbb{E}\left[\|\widetilde{\boldsymbol{\nu}}(U_i)\|_2^2\right] \\
\leq& pL_n \delta_p^2 a_n^2 \sum_{j=1}^{p} \mathbb{E}\left[\left(\boldsymbol{\nu}_{(j)}^\top \boldsymbol{B}_i^*\right)^2\right] \lesssim p\delta_p^2 a_n^2 \sum_{j=1}^{p} \|\boldsymbol{\nu}_j\|_2^2 = p\delta_p^2 a_n^2.
\end{aligned}
$$

where the second inequality holds since $T_i\left(\mathbf{M}^\ell\right)$ is bounded. In accordance with Lemma A.1, we obtain that

$$
\mathbb{P}\left(\Pi_{11} \lesssim p a_n L_n^{3/2} \sqrt{\frac{\log n}{n}} + 2a_n \sqrt{p} L_n n^{-M}\right) \geq 1 - n^{-\vartheta p L_n}. \tag{D.12}
$$

Similar bound also holds for Π_{12}, in conjunction with (D.10)-(D.12) and proper choice for M, we are guaranteed that

$$
\mathbb{P}\left(\Pi_1 \lesssim p a_n L_n^{3/2} \sqrt{\frac{\log n}{n}}\right) \geq 1 - n^{-\vartheta p L_n}. \tag{D.13}
$$

Now let $W_i = \left(\widetilde{\boldsymbol{\nu}}(U_i)^\top \boldsymbol{\delta}(U_i)\right)\left(\widetilde{\boldsymbol{\xi}}(U_i)^\top[\widetilde{\mathbf{M}}^\top \boldsymbol{B}_i - \boldsymbol{\mu}(U_i)]\right)$, then it follows from the dependence between U_i and Y_i that

$$\Pi_2 \le \left|\frac{1}{n}\sum_{i\in\mathcal{I}_1} W_i - \mathbb{E}[W_i]\right| + \left|\frac{1}{n}\sum_{i\in\mathcal{I}_2} W_i - \mathbb{E}[W_i]\right|.$$

Since W_i is bounded and $\sup_u |\widetilde{\mathbf{M}}_j^\top \boldsymbol{B}(u) - \mu_j(u)| \lesssim L_n^{-d} \le a_n$, we have

$$\begin{aligned}\mathbb{E}\left[(W_i - \mathbb{E}[W_i])^2 e^{\eta|W_i - \mathbb{E}[W_i]|}\right] &\lesssim \mathbb{E}[W_i^2] + (\mathbb{E}[W_i])^2 \\ &\lesssim \delta_p^2 \mathbb{E}[\|\widetilde{\mathbf{M}}^\top \boldsymbol{B}_i - \boldsymbol{\mu}(U_i)\|_2^2 \|\widetilde{\boldsymbol{\xi}}(U_i)\|_2^2] \\ &\le \delta_p^2 \mathbb{E}[\sup_u \|\widetilde{\mathbf{M}}^\top \boldsymbol{B}(u) - \boldsymbol{\mu}(u)\|_2^2 \|\widetilde{\boldsymbol{\xi}}(U_i)\|_2^2] \\ &\lesssim \delta_p^2 p a_n^2 L_n^{-1}.\end{aligned}$$

Applying Lemma A.1, we get

$$\mathbb{P}\left(\Pi_2 \lesssim p a_n \sqrt{\frac{\log n}{n}}\right) \ge 1 - n^{-\vartheta p L_n}. \tag{D.14}$$

Plugging (D.14) and (D.13) into (D.9), we finish the proof Lemma C.3. □

Appendix E: Iterative shrinkage thresholding algorithm

Next we take ISTA as an example to illustrate the optimization procedure to solve (2.9). Denote $g(\boldsymbol{\gamma}) = \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{D}_n \boldsymbol{\gamma} - \boldsymbol{b}_n^\top \boldsymbol{\gamma}$. Given a point $\boldsymbol{\gamma} \in \mathbb{R}^{pL_n}$, ISTA approach updates the solution through solving the following subproblem

$$\begin{aligned}\boldsymbol{\gamma}_+ &= Q_\eta(\boldsymbol{\gamma}) \\ &= \arg\min_{\boldsymbol{z}\in\mathbb{R}^{pL_n}} \left\{ g(\boldsymbol{\gamma}) + (\boldsymbol{z} - \boldsymbol{\gamma})^\top \nabla g(\boldsymbol{\gamma}) + \frac{1}{2\eta}\|\boldsymbol{z} - \boldsymbol{\gamma}\|_2^2 + \lambda_n \sum_{j=1}^p \|\boldsymbol{z}_{(j)}\|_2 \right\} \\ &= \arg\min_{\boldsymbol{z}\in\mathbb{R}^{pL_n}} \left\{ \frac{1}{2\eta}\sum_{j=1}^p \|\boldsymbol{z}_{(j)} - (\boldsymbol{\gamma} - \eta\nabla g(\boldsymbol{\gamma}))_{(j)}\|_2^2 + \lambda_n \|\boldsymbol{z}_{(j)}\|_2 \right\},\end{aligned}$$

where η is the step size. The solution of the subproblem is given by the soft-thresholding operator, that is

$$(\boldsymbol{\gamma}_+)_{(j)} = \frac{(\boldsymbol{\gamma} - \eta\nabla g(\boldsymbol{\gamma}))_{(j)}}{\|(\boldsymbol{\gamma} - \eta\nabla g(\boldsymbol{\gamma}))_{(j)}\|_2} \max\left\{0, \|(\boldsymbol{\gamma} - \eta\nabla g(\boldsymbol{\gamma}))_{(j)}\|_2 - \eta\lambda_n\right\}.$$

The step size η is determined by a backtracking line-search [2] such that

$$g(\boldsymbol{\gamma}_+) \le g(\boldsymbol{\gamma}) + (\boldsymbol{\gamma}_+ - \boldsymbol{\gamma})^\top \nabla g(\boldsymbol{\gamma}) + \frac{1}{2\eta}\|\boldsymbol{\gamma}_+ - \boldsymbol{\gamma}\|_2^2.$$

Another simple choice for η is $1/\|\mathbf{D}_n\|_2$, whereas it usually leads a very small step sizes and slow convergence [28]. For the ease of reference, we provide the detailed procedure in Algorithm 1.

Algorithm 1 ISTA with backtracking line-search

```
Input: Initial point $\gamma_0 \in \mathbb{R}^{pL_n}$, number of iterations $T$, shrinking rate $\rho \in (0,1)$, initial step size $\eta_0 \in (0,1)$.
for $t = 0, 1, ..., T-1$ do
    Compute the gradient: $\nabla g(\boldsymbol{\gamma}^t) = \mathbf{D}_n \boldsymbol{\gamma}^t - \boldsymbol{b}_n$.
    Find the smallest nonnegative integer $i_t$ such that with $\eta = \rho^{i_t} \eta_{t-1}$
        $$g(Q_\eta(\boldsymbol{\gamma}^t)) \leq g(\boldsymbol{\gamma}^t) + (Q_\eta(\boldsymbol{\gamma}^t) - \boldsymbol{\gamma}^t)^\top \nabla g(\boldsymbol{\gamma}^t) + \frac{1}{2\eta} \|Q_\eta(\boldsymbol{\gamma}^t) - \boldsymbol{\gamma}^t\|_2^2.$$
    Set $\eta_t = \rho^{i_t} \eta_{t-1}$ and update $\boldsymbol{\gamma}^{t+1} = Q_{\eta_t}(\boldsymbol{\gamma}^t)$.
end for
Output: The final solution $\boldsymbol{\gamma}^\top$.
```

References

[1] Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. Wiley, New York. MR0091588

[2] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** 183–202. MR2486527

[3] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

[4] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* **106** 1566–1577. MR2896857

[5] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

[6] Cai, T. and Zhang, L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 675–705. MR3997097

[7] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* **95** 888–902. MR1804446

[8] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351. MR2382651

[9] Chen, J., Li, D. and Linton, O. (2019). A new semiparametric estimation approach for large dynamic covariance matrices with multiple conditioning variables. *Journal of Econometrics* **212** 155–176. MR3994012

[10] Chen, Z. and Leng, C. (2016). Dynamic covariance models. *Journal of the American Statistical Association* **111** 1196–1207. MR3561942

[11] De Boor, C. (1978). *A practical guide to splines* **27**. Springer-Verlag New

York. MR507062

[12] Fan, J., Han, F. and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review* **1** 293–314.

[13] Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association* **109** 1270–1284. MR3265696

[14] Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface* **1** 179. MR2425354

[15] Fan, J., Zhang, W. et al. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics* **27** 1491–1518. MR1742497 MR1742497

[16] Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8** 86–100.

[17] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)* **55** 757–779. MR1229881

[18] Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference* **121** 113–125. MR2027718

[19] Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

[20] Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* **31** 1600–1635. MR2012827

[21] Jiang, B., Chen, Z. and Leng, C. (2020). Dynamic linear discriminant analysis in high dimensional space. *Bernoulli* **26** 1234–1268. MR4058366

[22] Lu, X., Dong, F., Liu, X. and Chang, X. (2018). Varying coefficient support vector machines. *Statistics & Probability Letters* **132** 107–115. MR3718095

[23] Mai, Q., Yang, Y. and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica* **29** 97–111. MR3889359

[24] Mai, Q. and Zou, H. (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis* **135** 175–188. MR3306434

[25] Mai, Q., Zou, H. and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42. MR2899661

[26] Monti and S. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105** 1851–1861.

[27] Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika* **107** 415–431. MR4108937

[28] Qin, Z., Scheinberg, K. and Goldfarb, D. (2013). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation* **5** 143–169. MR3069877

[29] Schumaker, L. (2007). *Spline functions: basic theory.* Cambridge University Press. MR2348176

[30] Shao, J., Wang, Y., Deng, X., Wang, S. et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics* **39** 1241–1265. MR2816353

[31] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10** 1040–1053. MR0673642

[32] Wang, H., Peng, B., Li, D. and Leng, C. (2021). Nonparametric estimation of large covariance matrices with conditional sparsity. *Journal of Econometrics* **223** 53–72. MR4252147

[33] Wang, H. J., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics* **37** 3841–3866. MR2572445

[34] Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 615–636. MR2749910

[35] Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C. and Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics* **25** 1145–1151.

[36] Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22** 1112–1137. MR2328530

[37] Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research* **13** 1973–1998. MR2956349

[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574