Statistical Analysis of High-Dimensional Models

新久，章

(Degree)
博士（経済学）
(Date of Degree)
2023-03-25
(Date of Publication)
2025-03-25
(Resource Type)
doctoral thesis
(Report Number)
甲第8570号
(URL)
https://hdl.handle.net/20.500.14094/0100482318

博　士　論　文

令和 4 年 12 月

神戸大学大学院経済学研究科
経済学専攻
指導教員　末石　直也
新久　章
(Akira Shinkyu)

博　士　論　文

Statistical Analysis of High-Dimensional Models

(高次元モデルの統計解析)

令和 4 年 12 月

神戸大学大学院経済学研究科
経済学専攻
指導教員　末石　直也
新久　章
(Akira Shinkyu)

Acknowledgments

This dissertation contains results of my three years research conducted in the Graduate School of Economics, Kobe University. This thesis is composed of three articles. The first article investigates model selection in ultra-high dimensional varying coefficient models, which is based on Shinkyu (2021), "Forward selection for feature screening and structure identification in varying coefficient models," Sankhya A, to appear. The second article considers statistical inference for high-dimensional linear regression models. The third article studies prediction under misspecified high-dimensional models.

I would like to express the deepest gratitude to my supervisor, Prof. Naoya Sueishi. Without his helpful support and patient guidance throughout three years, I would never have completed this dissertation. I would also like to deeply thank Prof. Toshio Honda, Prof. Yoshimasa Uematsu, Prof. Mototsugu Shintani, Prof. Wenjie Wang, Prof. Akio Namba, Prof. Kaiji Motegi, and participants at my presentations for their valuable comments, which give significant improvement to this dissertation.

I am deeply grateful to my parents, who always encourage and support me. Special thanks go to the Japan Science and Technology Agency for financial support.

December, 2022

Akira Shinkyu

Contents

Chapter 1

Forward Selection for Feature Screening and Structure Identification in Varying Coefficient Models

1.1 Introduction

Nowadays, high dimensional data are frequently collected in diverse scientific fields such as genomics, finance, economics, biology, and so on. Their main characteristic is that the total number of covariates is much larger than the sample size. However, it is believed that the number of relevant covariates is relatively smaller than the total number of covariates. Sparse regression models draw much attention to the analysis of high dimensional data, and variable selection is an important task. To carry out variable selection, penalized least squares procedures have been proposed. They include the Lasso by Tibshirani (1996), the SCAD by Fan and Li (2001), and the group Lasso by Yuan and Lin (2006), to name just a few. However, when there are too many covariates, these procedures may not work well due to computational complexities and algorithmic instabilities.

To cope with this issue, Fan and Lv (2008) proposed the sure independence screening (SIS). The purpose of SIS is to remove irrelevant covariates sufficiently, so that penalized least squares methods can work well at the subsequent step. SIS selects covariates which are highly correlated with the response, and the computational cost is small. SIS can select all relevant covariates with probability tending to one under some conditions. After their work, feature screening procedures such as model-based methods which specify marginal models between the response and each covariate (Li et al. 2012a; Fan et al. 2011; Fan et al. 2014), or model-free methods which employ relevance measures between the response and each covariate (Zhu et al. 2011; Li et al. 2012b; Mai and Zou 2015) have been proposed. Liu et al. 2015 gave reviews for these methods.

However, these feature screening procedures tend to select many irrelevant covariates when they are highly correlated with relevant covariates. They also tend to miss relevant covariates which are marginally uncorrelated with the response. Iterative methods have been proposed to address these matters, but theoretical justification is not provided. An additional approach to carry out feature screening is forward selection. Wang (2009) proved that a forward selection procedure can also select all relevant covariates with probability tending to one in linear models. His simulation studies also showed that the forward selection procedure selects much fewer irrelevant covariates than SIS, although both are competitive in selecting relevant covariates. After his work, Cheng et al. (2018) proposed more efficient forward selection methods for linear models. Cheng et al. (2016), Honda and Lin (2021), and Zhong et al. (2020) studied forward selection procedures for nonparametric models.

In real practice, complex relationships between covariates and the response often arise, and varying coefficient models are useful in capturing them. Variable selection procedures have been proposed for varying coefficient models, as in Wang et al. (2008), Wang and Xia (2009), and so on. Some authors also studied feature screening for varying coefficient models, as in Song et al. (2014), Liu et al. (2014), Fan et al. (2014), and Cheng et al. (2016). In particular, Cheng et al. (2016) demonstrated desirable properties of forward selection procedures.

It is possible that there are covariates having constant coefficients in varying coefficient models, and such models are called as partially linear varying coefficient models. Identifying these covariates is also crucial to reduce model complexities. Some authors have proposed simultaneous variable selection and structure identification procedures based on penalized least squares. For example, see Cheng et al. (2014), Wang and Lin (2016), Honda and Yabe (2017), and Honda et al. (2019). On the other hand, as far as we know, no author has developed simultaneous feature screening and structure identification procedures. Existing feature screening methods for varying coefficient models do not include structure identification. Hence, they do not explicitly select partially linear varying coefficient models. We aim to address this issue.

In this paper, motivated by desirable properties of forward selection procedures proposed by Cheng et al. (2016) and the orthonormal basis proposed by Honda and Yabe (2017), we propose a forward selection procedure for simultaneous feature screening and structure identification in ultra-high dimensional varying coefficient models. Unlike existing feature screening procedures, our procedure classifies all covariates into three groups: covariates with constant coefficients, covariates with non-constant coefficients, and covariates with zero coefficients. Hence, our procedure can explicitly select partially linear varying coefficient models. Under some conditions, we establish the screening consistency for our procedure with EBIC stopping rules. Numerical studies show that our procedure performs well in finite sample.

The rest of this paper is organized as follows. In Section 1.2, we introduce notations and our forward selection procedure. In Section 1.3, we give assumptions and main theoretical results for our procedure. In Section 1.4, results of numerical studies are presented. Conclusions are stated in Section 1.5, and proofs of theoretical results are given in Section 1.6.

1.2 Forward Selection Procedure

1.2.1 Model and Notations

Before we describe our forward selection procedure, we introduce notations used throughout this paper. $|A|$ represents the number of elements in a set A. $\{x\}_+$ is the maximum value of x and 0 for a scalar x. $\|\boldsymbol{a}\|$ and $\boldsymbol{a}^T$ stand for the Euclidean norm and the transpose for a vector $\boldsymbol{a}$. $\|f\|_2$ and $\|f\|_\infty$ denote the L_2 and sup norms for a function f on the unit interval. $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ denote the maximum and minimum eigenvalues of a matrix $\boldsymbol{A}$. The spectral norm and the maximum element in absolute value of a matrix $\boldsymbol{A}$ are denoted as $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}\|_\infty$. $\xrightarrow{P}$ stands for convergence in probability.

Suppose that we have n i.i.d. observations $\{(\boldsymbol{X}_i, Y_i, Z_i)\}_{i=1}^n$ from the varying coefficient model

$$Y_i = \sum_{j=1}^{p} g_j(Z_i) X_{ij} + \epsilon_i, \tag{1.1}$$

where Y_i is a response variable, $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^T$ is a p dimensional covariate vector, $X_{i1} = 1$, Z_i is an index variable on the unit interval, $g_j(z)$ is an unknown smooth function on the unit interval for $j = 1, \ldots, p$, and ϵ is an error term with mean 0 and variance $\sigma^2 > 0$. We denote $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$. To identify constant coefficients and non-constant coefficients, we decompose $g_j(z)$ into its constant part and its non-constant part. This means

$$
\begin{aligned}
g_j(z) &= g_{cj} + g_{vj}(z), \\
g_{cj} &= \int_0^1 g_j(z)dz, \\
g_{vj}(z) &= g_j(z) - \int_0^1 g_j(z)dz.
\end{aligned}
$$

g_{cj} and $g_{vj}(z)$ are orthogonal since $\int_0^1 g_{cj} g_{vj}(z)dz = 0$. Let us denote the index sets by

$$
\mathcal{S}_c = \{j : g_{cj} \neq 0\}, \quad \mathcal{S}_v = \{j : g_{vj}(z) \not\equiv 0\}.
$$

Notice that $j \in \mathcal{S}_c \cap \mathcal{S}_v^c$ implies that $g_j(z)$ is a constant coefficient, and $j \in \mathcal{S}_v$ implies that $g_j(z)$ is a non-constant coefficient. If $j \notin \mathcal{S}_c \cup \mathcal{S}_v$, then $g_j(z)$ is a zero coefficient. Let $|\mathcal{S}_c \cup \mathcal{S}_v| = p_0$. Throughout this paper, we assume that p_0 is a fixed positive constant. We also assume that $\{1\} \subset \mathcal{S}_c$.

We employ the orthonormal basis $\boldsymbol{B}(z) = (B_1(z), \ldots, B_L(z))^T$ introduced by Honda and Yabe (2017) to estimate coefficients. We can construct $\boldsymbol{B}(z)$ from the equispaced B-spline basis $\boldsymbol{B}_0(z) = (B_{01}(z), \ldots, B_{0L}(z))^T$ whose order is larger than or equal to two. We can write $\boldsymbol{B}(z)$ as $\boldsymbol{B}(z) = \boldsymbol{A}_0 \boldsymbol{B}_0(z)$ where $\boldsymbol{A}_0$ is calculated numerically. Note that $B_1(z) = 1/\sqrt{L}, B_2(z) = 12/\sqrt{L}(z - 1/2)$, and

$$
\int_0^1 \boldsymbol{B}(z)\boldsymbol{B}^T(z)dz = L^{-1}\boldsymbol{I}_L,
$$

where $\boldsymbol{I}_L$ is the $L \times L$ identity. Here we take $L = C_L n^{1/5}$ for some positive constant C_L to simplify our presentation. By choosing a suitable L dimensional vector $\boldsymbol{\gamma}_j = (\gamma_{1j}, \boldsymbol{\gamma}_{-1j}^T)^T$, we can represent g_{cj} and $g_{vj}(z)$ as

$$
\begin{aligned}
g_{cj} &= B_1(z)\gamma_{1j}, \\
g_{vj}(z) &= \boldsymbol{B}_{-1}^T(z)\boldsymbol{\gamma}_{-1j} + r_j(z),
\end{aligned}
$$

where $\boldsymbol{B}_{-1}(z) = (B_2(z), \ldots, B_L(z))^T$ and $r_j(z)$ is an approximation error. Hence, $B_1(z)\gamma_{1j}$ is the constant part of $g_j(z)$, and $\boldsymbol{B}_{-1}^T(z)\boldsymbol{\gamma}_{-1j}$ gives the approximation for the non-constant part of $g_j(z)$. Note that we take $\gamma_{1j} = 0$ if $g_{cj} = 0$ and $\boldsymbol{\gamma}_{-1j} = 0$ if $g_{vj}(z) \equiv 0$. Then the model (1.1) can be expressed as

$$Y_i = \sum_{j=1}^{p} \{B_1(Z_i)\gamma_{1j} + \boldsymbol{B}_{-1}^T(Z_i)\boldsymbol{\gamma}_{-1j}\} X_{ij} + \epsilon_i^*, \tag{1.2}$$

where $\epsilon_i^* = \sum_{j=1}^{p} r_j(Z_i)X_{ij} + \epsilon_i$. Let us define

$$\begin{aligned}
\boldsymbol{W}_{1j} &= (X_{1j}B_1(Z_1), \ldots, X_{nj}B_1(Z_n))^T, \\
\boldsymbol{W}_{-1j} &= (X_{1j}\boldsymbol{B}_{-1}(Z_1), \ldots, X_{nj}\boldsymbol{B}_{-1}(Z_n))^T, \\
\boldsymbol{\gamma} &= (\gamma_{11}, \boldsymbol{\gamma}_{-11}^T, \ldots, \gamma_{1p}, \boldsymbol{\gamma}_{-1p}^T)^T, \\
\boldsymbol{W} &= (\boldsymbol{W}_{11}, \boldsymbol{W}_{-11}, \ldots, \boldsymbol{W}_{1p}, \boldsymbol{W}_{-1p}).
\end{aligned}$$

$\boldsymbol{W}_{1j}$ is an n dimensional covariate vector for the jth constant part, and $\boldsymbol{W}_{-1j}$ is an $n \times (L-1)$ covariate matrix for the jth non-constant part. Notice that $\boldsymbol{\gamma}$ is a pL dimensional vector, and $\boldsymbol{W}$ is an $n \times pL$ matrix. Then the matrix form of (1.2) is given by

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\epsilon}^*, \tag{1.3}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \ldots, \epsilon_n^*)^T$. We will apply our forward selection procedure introduced below to (1.3) for feature screening and structure identification in (1.1).

Let $\mathcal{F}_c = \{11, \ldots, 1p\}$, $\mathcal{F}_v = \{-11, \ldots, -1p\}$, and $\mathcal{F} = \mathcal{F}_c \cup \mathcal{F}_v$. Note that $\mathcal{F}_c$ includes all indices of covariates for constant parts in (1.3), and $\mathcal{F}_v$ includes all indices of covariates for non-constant parts in (1.3). Let

$$\mathcal{S}_0 = \{1j : j \in \mathcal{S}_c\} \cup \{-1j : j \in \mathcal{S}_v\}. \tag{1.4}$$

$\mathcal{S}_0$ includes indices of all relevant covariates in (1.3). Let $\boldsymbol{W}_A$ denote an $n \times \{|A \cap \mathcal{F}_c| + (L-1)|A \cap \mathcal{F}_v|\}$ matrix consisting of columns of $\boldsymbol{W}$ indexed by $A \subset \mathcal{F}$. Let $\boldsymbol{\gamma}_A$ denote a vector whose dimension is $\{|A \cap \mathcal{F}_c| + (L-1)|A \cap \mathcal{F}_v|\}$ and it consists of components of $\boldsymbol{\gamma}$ indexed by $A \subset \mathcal{F}$. For example, if $A = \{11, 12, -11\}$, then $\boldsymbol{W}_A = (\boldsymbol{W}_{11}, \boldsymbol{W}_{12}, \boldsymbol{W}_{-11})$, and $\boldsymbol{\gamma}_A = (\gamma_{11}, \gamma_{12}, \boldsymbol{\gamma}_{-11}^T)^T$. We also define $\widehat{\boldsymbol{\gamma}}_A = (\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Y}$ and $\hat{\sigma}_A^2 = \|\boldsymbol{Y} - \boldsymbol{W}_A\widehat{\boldsymbol{\gamma}}_A\|^2/n$. For the Gram matrix $\boldsymbol{W}^T\boldsymbol{W}/n$, let us define the sub Gram matrix and its expectation as $\widehat{\boldsymbol{\Sigma}}_S = \boldsymbol{W}_S^T\boldsymbol{W}_S/n$ and $\boldsymbol{\Sigma}_S = \mathrm{E}[\widehat{\boldsymbol{\Sigma}}_S]$ for any $S \subset \mathcal{F}$, respectively. We use the extended BIC (EBIC) proposed by Chen and Chen (2008) as the stopping criterion of our forward selection procedure. EBIC for any $A \subset \mathcal{F}$ is defined as

$$\mathrm{EBIC}(A) = n \cdot \log(\hat{\sigma}_A^2) + \{|A \cap \mathcal{F}_c| + (L-1)|A \cap \mathcal{F}_v|\}(\log n + 2\eta \log p),$$

where $0 \leq \eta \leq 1$. Note that EBIC is equal to the BIC when $\eta = 0$. Let us denote E_k and F_k as sets of candidate indices for $\mathcal{S}_c$ and $\mathcal{S}_v$ which our forward selection procedure has selected from $\{1, \ldots, p\}$ at the completion of the kth step, respectively. We also define $S_k = \{1j : j \in E_k\} \cup \{-1j : j \in F_k\}$.

1.2.2 Forward selection procedure

We introduce our forward selection procedure.

(1) Set $E_1 = \{1\},\ S_1 = \{11\}$, and $F_1 = \emptyset$. Compute $\mathrm{EBIC}(S_1)$.

(2) In the $k+1$th step for $k \geq 1$, S_k, F_k, and E_k are given. Find j^* such that

$$j^* = \operatorname*{argmin}_{j \in \mathcal{F} \cap S_k^c} \hat{\sigma}^2_{S_k \cup \{j\}}.$$

(3) Suppose that $j^* = -1j$ for some $j \in \{1, \ldots, p\}$. Suppose further that we get $\mathrm{EBIC}(S_k \cup \{j^*\}) \leq \mathrm{EBIC}(S_k)$. If $j \notin E_k$, then set $S_{k+1} = S_k \cup \{1j, -1j\}$, $F_{k+1} = F_k \cup \{j\}$, $E_{k+1} = E_k \cup \{j\}$, and go back to (2) with changing k to $k+1$. Otherwise, set $S_{k+1} = S_k \cup \{-1j\}$, $F_{k+1} = F_k \cup \{j\}, E_{k+1} = E_k$, and go back to (2) with changing k to $k+1$. If $\mathrm{EBIC}(S_k \cup \{j^*\}) > \mathrm{EBIC}(S_k)$, then end (3).

Suppose that $j^* = 1j$ for some $j \in \{1, \ldots, p\}$. If we get $\mathrm{EBIC}(S_k \cup \{j^*\}) \leq \mathrm{EBIC}(S_k)$, then set $S_{k+1} = S_k \cup \{1j\}$, $F_{k+1} = F_k$, $E_{k+1} = E_k \cup \{j\}$, and go back to (2) with changing k to $k+1$. Otherwise, end (3).

After the end of (3), we get E_k and F_k. If $j \in E_k \cap F_k^c$, we declare that X_{ij} is a relevant covariate with a constant coefficient. If $j \in F_k$, we declare that X_{ij} is a relevant covariate with a non-constant coefficient. If $j \notin E_k \cup F_k$, we declare that X_{ij} is a covariate with a zero coefficient. Note that when our forward selection procedure adds an index to F_k, it also adds the same index to E_k, if the index does not belong to it yet. This is because $g_{vj}(z) \not\equiv 0$ implies $g_{cj} \neq 0$ in many cases. We may incidentally have some coefficient $g_j(z)$ such that $g_{cj} = 0$ and $g_{vj}(z) \not\equiv 0$. For example, if $g_j(z) = z - 1/2$, then $g_{cj} = 0$ and $g_{vj}(z) \not\equiv 0$. However, having such coefficients rarely happens. Hence, F_k should be the subset of E_k.

Remark 1. Recently, spatio-temporal varying coefficient models draw attention. We can modify our procedure for feature screening and structure identification of these models. For example, consider a spatio-temporal varying coefficient model proposed by Serban (2011):

$$Y_{ij} = \sum_{k=1}^{p} g_k(t_i, s_j) X_{k,ij} + \epsilon_{ij}, \tag{1.5}$$

where t_i is the time point for $i = 1, \dots, T$, $s_j = (s_{j1}, s_{j2})$ is the location point for $j = 1, \dots, S$. With appropriate scaling, we can assume that $t_i \in [0, 1]$ for $i = 1, \dots, T$ and $s_j \in [0, 1] \times [0, 1]$ for $j = 1, \dots, S$. For simplicity, suppose that we can decompose $g_k(t_i, s_j)$ into a time dependent function and a location dependent function:

$$g_k(t_i, s_j) = h_k(t_i) + f_k(s_j).$$

Apply the orthogonal decomposition to h_k and f_k:

$$h_k(t) = h_{ck} + h_{vk}(t),\ h_{ck} = \int_0^1 h_k(t)dt,\ h_{vk}(t) = h_k(t) - h_{ck},$$

$$f_k(x, y) = f_{ck} + f_{vk}(x, y), \quad f_{ck} = \int_0^1 \int_0^1 f_k(x, y)dxdy, \quad f_{vk}(x, y) = f(x, y) - f_{ck}.$$

Then, there are some γ^*_{1k} and $\boldsymbol{\gamma}^*_{-1k}$ such that

$$h_{ck} = B_1(t_i)\gamma^*_{1k}, \quad h_{vk}(t_i) = \boldsymbol{B}^T_{-1}(t_i)\boldsymbol{\gamma}^*_{-1k} + r_{t,k}(t_i),$$

where $r_{t,k}(t_i)$ is an approximation error, and $\boldsymbol{B}(z) = (B_1(z), \dots, B_L(z))^T = (B_1(z), \boldsymbol{B}_{-1}(z)^T)^T$ is the orthonormal basis given in this section. Note that if $h_{ck} = 0$ and $h_{vk}(t) \equiv 0$, then we set $\gamma^*_{1k} = 0$ and $\boldsymbol{\gamma}^*_{-1k} = 0$, respectively. We can also write

$$f_{ck} = B_1(s_{j1})B_1(s_{j2})a^*_{1,1,k}, \quad f_{vk}(s_j) = \boldsymbol{B}^T_{-1,-1}(s_j)\boldsymbol{a}^*_{-1,-1,k} + r_{s,k}(s_j),$$

where $r_{s,k}(s_j)$ is an approximation error, and

$$\begin{aligned} \boldsymbol{B}_{-1,-1}(s_j) = (&B_1(s_{j1})B_2(s_{j2}), \dots, B_1(s_{j1})B_L(s_{j2}), \\ &B_2(s_{j1})B_1(s_{j2}), \dots, B_L(s_{j1})B_L(s_{j2}))^T \in \mathbb{R}^{L^2-1}, \\ \boldsymbol{a}^*_{-1,-1,k} = (&a^*_{1,2,k}, \dots, a^*_{1,L,k}, a^*_{2,1,k}, \dots, a^*_{L,L,k})^T \in \mathbb{R}^{L^2-1}, \end{aligned}$$

for some $(a_{l,s,k})_{l,s=1}^{L}$. Note that if $f_{ck} = 0$ and $f_{vk} \equiv 0$, then we take $a^*_{1,1,k} = 0$ and $\boldsymbol{a}^*_{-1,-1,k} = 0$, respectively. Let

$$\mathcal{S}_{t,c} = \{k : \gamma^*_{1k} \neq 0\},\ \mathcal{S}_{t,v} = \{k : \boldsymbol{\gamma}^*_{-1k} \neq 0\},$$
$$\mathcal{S}_{s,c} = \{k : a^*_{1,1,k} \neq 0\},\ \mathcal{S}_{s,v} = \{k :\ \boldsymbol{a}^*_{-1,-1,k} \neq 0\}.$$

These four sets above give the true structure of the model (1.5). We can also define covariates for constant components and non-constant components for time dependent functions and location dependent functions as in Section 1.2 of this paper. In this case, we will employ the group Lasso in the forward selection of covariates with separating penalties as in Luo and Chen (2014) and Honda and Yabe (2017), since non-constant components for location dependent functions are relatively high dimensional. Theoretical analysis and simulation studies for the method describe in this remark may be interesting topics for future work.

1.3 Theoretical Results

1.3.1 Assumptions

To get theoretical insights of our procedure, we introduce following assumptions. First, we state an assumption on the number of covariates.

Assumption 1. There exists a positive constant $1/5 < \xi_0 < 4/5$ such that

$$\log p = O(n^{\xi_0}).$$

This assumption allows us that the number of covariates is exponentially larger than the sample size. In our theoretical analysis, the constant ξ_0 should be larger than $1/5$ to stop our procedure after selecting all relevant covariates. ξ_0 also should be smaller than $4/5$ to get the penalty term of EBIC not too large so that our procedure does not stop when some relevant covariates are not selected yet.

The next assumption is about eigenvalues for submatrices of the expected Gram matrix.

Assumption 2. For any given integer M larger than p_0, there exist positive constants

$\tau_{\min}$ and $\tau_{\max}$ such that

$$\frac{\tau_{\min}}{L} \le \lambda_{\min}(\boldsymbol{\Sigma}_S) \le \lambda_{\max}(\boldsymbol{\Sigma}_S) \le \frac{\tau_{\max}}{L},$$

for any $S \subset \mathcal{F}$ satisfying $|S| \le M$.

We can see this type of assumption in other papers such as Cheng et al. (2016), Wang (2009), Lee et al. (2014), and Honda and Lin (2021). We use this assumption to bound eigenvalues for submatrices of the Gram matrix. In this paper, we have to assume that M diverges slowly, such that $M = O(n^{\xi})$ for some $0 < \xi < 1/10$. This assumption looks restrictive. However, if we admit that the order of M is larger than $O(n^{\xi})$, then it is difficult to bound eigenvalues for submatrices of the Gram matrix via Bernstein's inequality. Hence, we decide to impose the small order on M to simplify presentation of our paper.

Next, we introduce an assumption which is about conditional tail probabilities of covariates.

Assumption 3. There exist positive constants $K_1, \ldots, K_p$ and $r_1, \ldots, r_p \ge 2$ such that

$$P(|X_j| > t|Z) \le \exp\left(1 - \left(\frac{t}{K_j}\right)^{r_j}\right),\ j = 1, \ldots, p,$$

for any $t \ge 0$, uniformly on the compact support of Z.

We can see this assumption in Fan et al. (2014). This exponential tail assumption allows us to bound the higher order expectations of covariates in a simple manner. Under this assumption, we can use Bernstein's inequality in evaluation of eigenvalues for submatrices of the Gram matrix.

Next, we state an assumption on each coefficient of our varying coefficient model.

Assumption 4. For $j = 1, \ldots, p$, $g_j(z)$ is twice continuously differentiable and there exists a positive constant C_g such that

$$\sum_{j=1}^{p} \|g_j\|_\infty \le C_g, \quad \sum_{j=1}^{p} \|g_j'\|_\infty \le C_g, \quad \sum_{j=1}^{p} \|g_j''\|_\infty \le C_g.$$

Furthermore, there exists a positive constant $C_{\min}$ such that

$$\min\left\{\max_{j \in \mathcal{S}_c} |g_{cj}|^2, \max_{j \in \mathcal{S}_v} \|g_{vj}\|_2^2\right\} > C_{\min}.$$

We can see the first part of this assumption in Honda and Yabe (2017). Under the first part of this assumption, a linear combination of the orthonormal basis can approximate nonzero coefficients up to the order of L^{-2}. See Appendix A of Honda and Yabe (2017) for properties of the orthonormal basis. Note that since we consider sparse models, only the bounded number of coefficients are related to the first part of this assumption. Under the second part of this assumption, at least a function among $g_j(z)$ is bounded away from zero. The second part of this assumption may be restrictive compared with other assumptions such as the beta min condition. However, under this assumption, we can simply evaluate $\|\boldsymbol{\gamma}\|$ so that we can avoid unnecessary technical discussion.

Next, we state a sub-Gaussianity assumption on error terms conditionally on all the covariates and the index variable.

Assumption 5. There exists a positive constant σ_1 such that for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$E\left[\exp(\boldsymbol{x}^T\boldsymbol{\epsilon})|\boldsymbol{X}, \boldsymbol{Z}\right] \leq e^{\|\boldsymbol{x}\|^2\sigma_1^2/2}.$$

As in Cheng et al. (2016), this assumption allows us to give the smaller order for quadratic forms of error terms. Even though this assumption is only used for the evaluation of these terms, it is important for our procedure to have the screening consistency property.

Finally, we state an assumption on covariates.

Assumption 6. There exists a positive constant C_X such that

$$\max_{i,j} |X_{ij}| \leq C_X.$$

Under this assumption, covariates are not relevant in evaluating approximation errors of the orthonormal basis. Hence, this assumption allows us to deal with approximation errors in a simple way. We can see this uniformly bound assumption in other papers on varying coefficient models such as Wang et al. (2008).

1.3.2 Main theorems

Now we present main theoretical results for our forward selection procedure. An implication of Theorem 1 is that differences of mean squared errors are large when not all relevant covariates are selected yet. This result theoretically ensures that the forward selection procedure continues variable selection while there is a relevant covariate which is not selected yet.

Theorem 1. *Suppose that Assumptions 1-6 hold with $M = O(n^{\xi})$ for some $0 < \xi < 1/10$. Then, there exist some positive constant D^* and some sequence of positive numbers $\{\delta_{2n}\}$ tending to zero sufficiently slowly such that with probability tending to one, we have*

$$\max_{l \in \mathcal{F} \cap S^c} \{n\hat{\sigma}_S^2 - n\hat{\hat{\sigma}}^2_{S \cup \{l\})}\} \geq n(1 - \delta_{2n})D^*, \tag{1.6}$$

for any $S \subset \mathcal{F}$ satisfying $|S \cup \mathcal{S}_0| \leq M$ and $\mathcal{S}_0 \not\subset S$.

Theorem 2 establishes the screening consistency for our procedure. Namely, our procedure can select all relevant covariates consistently.

Theorem 2. *Suppose that Assumptions 1-6 hold with $M = O(n^{\xi})$ for some $0 < \xi < 1/10$. Let T^* be the smallest integer larger than or equal to $\{(1 - \delta_{2n})D^*\}^{-1} Var(Y_i)$. If $2T^* \leq M - 2p_0$, then, with probability tending to one, we have*

$$\mathcal{S}_c \subset E_k \text{ and } \mathcal{S}_v \subset F_k \text{ for some } 1 \leq k \leq T^*. \tag{1.7}$$

Theorem 3 shows the validity of EBIC stopping rules for our forward selection procedure.

Theorem 3. *Suppose that Assumptions 1-6 hold with $M = O(n^{\xi})$ for some $0 < \xi < 1/10$. If $\mathcal{S}_0 \not\subset S_{k-1}$, $\mathcal{S}_0 \subset S_k$, and $k \leq M$, then our forward selection procedure stops at the kth step with probability tending to one.*

Remark 2. Note that Theorem 2 does not imply that $\mathcal{S}_c \cap \mathcal{S}_v^c \subset E_k \cap F_k^c$ for some k with probability tending to one. It means that the proposed procedure may identify constant coefficients as non-constant coefficients incorrectly. In addition, the proposed procedure may select moderately large models although it can remove the sufficiently large number of irrelevant covariates. We can apply the group SCAD or selection consistency procedures such as the method by Wang and Lin (2016) to models selected by our method if one needs further variable selection and/or structure identification.

1.4 Numerical Studies

In this section, we present the results of simulation studies and a real data example for our forward selection procedure. In subsection 1.4.1, we show the results of simulation studies. In subsection 1.4.2, we show the result of applying our procedure to microarray data. We employed EBIC with $\eta = 0.06$ as the stopping criterion of our procedure. We set the dimension of the orthonormal basis as $L = [2n^{1/5}]$, where $[\cdot]$ is the function rounding to the nearest integer. In practice, some data would fluctuate EBIC values. In that case, our forward selection procedure may stop running a little too early and miss some relevant covariates. To avoid stopping the selection too early due to the interference of such small fluctuations, we ended the operation of our forward selection procedure only if EBIC increases for five consecutive steps as in Cheng et al. (2016).

1.4.1 Simulation Studies

In this subsection, we compare the finite sample performances of our forward selection procedure with those of the group Lasso (gLasso) and the group SCAD (gSCAD). We used R x64 4.0.2 and the R package *grpreg* of Breheny and Huang (2015) to implement gLasso and gSCAD. We chose their tuning parameters by the BIC, which is the default method. We set $(n, p) = (300, 1000)$, and the repetition number is 1000. As in Fan et al. (2014), the varying coefficient model in simulation studies is defined as

$$Y_i = \sum_{j=1}^{p} g_j(Z_i) X_{ij} + \epsilon_i,$$

where coefficients are given by

$$g_1(z) = 2, \quad g_2(z) = 3z, \quad g_3(z) = (z+1)^2,$$

$$g_4(z) = \frac{4\sin(2\pi z)}{2 - \sin(2\pi z)}, \quad g_j(z) = 0, \ j = 5, \ldots, p.$$

Covariates and the index variable are defined as

$$X_{ij} = \frac{T_{ij} + t_1 U_{i1}}{1 + t_1}, \quad Z_i = \frac{U_{i2} + t_2 U_{i1}}{1 + t_2}, \ j = 1, \ldots, p.$$

$\{T_{ij}\}_{j=1}^{p}, \{U_{ij}\}_{j=1}^{2}$, and ϵ_i were independently generated from $N(0,1), U(0,1)$, and $N(0,1)$, respectively. t_1 and t_2 are tuning parameters which determine correlation among X_{ij} and correlation between X_{ij} and Z_i. As in Cheng et al. (2016), we can show that for $j \neq k$,

$$\mathrm{corr}(X_{ij}, X_{ik}) = \frac{t_1^2}{(12 + t_1^2)}, \quad \mathrm{corr}(X_{ij}, Z_i) = \frac{t_1 t_2}{\sqrt{(12 + t_1^2)(1 + t_2^2)}}.$$

For the correspondence between (t_1, t_2) and correlations, see Table 1 of Cheng et al. (2016). Note that $\mathrm{corr}(X_{ij}, X_{ik})$ and $\mathrm{corr}(X_{ij}, Z_i)$ increase as t_1 and t_2 increase. As correlations are higher, variable selection and structure identification are more challenging. In this model, X_{i1} is a relevant covariate with a constant coefficient, and X_{i2}, X_{i3}, and X_{i4} are relevant covariates with non-constant coefficients. On the other hand, $X_{i5}, \dots, X_{ip}$ are irrelevant covariates. Our aim is not only to select relevant covariates, but also to identify constant coefficients and non-constant coefficients. To this aim, let us define

$$\begin{aligned} C_{sj} &= 1\{g_j(z) \text{ is identified as a constant function at the } s\text{th repetition}\}, \\ NC_{sj} &= 1\{g_j(z) \text{ is identified as a non-constant function at the } s\text{th repetition}\}, \\ NS_{sj} &= 1\{g_j(z) \text{ is identified as a zero function at the } s\text{th repetition}\}, \end{aligned}$$

as in Honda et al. (2019), where $1\{A\}$ is the indicator function for an event A. Note that $C_{sj} + NC_{sj} + NS_{sj} = 1$ for $j = 1, \dots, p$ and $s = 1, \dots, 1000$. Then, performance measures for structure identification are

$$C_j = \frac{1}{1000} \sum_{s=1}^{1000} C_{sj}, \quad NC_j = \frac{1}{1000} \sum_{s=1}^{1000} NC_{sj}, \quad NS_j = \frac{1}{1000} \sum_{s=1}^{1000} NS_{sj}.$$

C_j, NC_j, and NS_j represent the frequencies where $g_j(z)$ is identified as a constant function, a non-constant function, and a zero function, respectively. We also calculated the average numbers of true positive (TP) and false positive (FP) over repetition numbers as performance measures for variable selection. The numbers of TP and FP show the numbers of relevant covariates and irrelevant covariates selected by an approach, respectively. Since we also consider feature screening, higher the number of TP with the small or moderate number of FP is better.

Table 1 shows the simulation results of gLasso, gSCAD, and our forward selection procedure. Our method identifies the constant coefficient well when correlations are low or moderate. However, as in Remark 2, our method sometimes identifies the constant coefficient as a non-constant coefficient, particularly when correlations are high. One reason for this incorrect identification is that the covariate for the non-constant component of X_1 also improves the prediction of current models, and selecting it decreases or does not increase EBIC significantly because of correlations. Compared with our procedure, gLasso and gSCAD identify the constant coefficient more frequently. However, this may be because these methods are hard to select covariates for non-constant components. These methods hardly identify non-constant coefficients in this model even when correlations are small. It may be because covariates for non-constant components have relatively weak signals compared with the tuning parameters selected by the BIC. However, if we employ other criteria such as the AIC or the cross validation for the tuning parameters selection, then these methods select too many irrelevant covariates.

Table 1 also shows that our method selects many relevant covariates and less irrelevant covariates stably in general. gLasso does not perform well for variable selection of this model, since it selects fewer relevant covariates in many cases and many irrelevant covariates in all cases. gSCAD performs better than gLasso, but it selects relatively fewer relevant covariates than our procedure. Finally, our method is relatively more computationally intensive than gLasso and gSCAD, because our method is iterative. gLasso and gSCAD tuned by the BIC are too fast due to the coordinate descent algorithms. However, our method only takes about three to five seconds for one implementation on average, even though there are one thousand covariates. Hence, our method is also practical in computational cost. Note that if we employ other information criteria such as the cross validation for the tuning parameters selection of gLasso and gSCAD, they need much more computational time than our method.

In conclusion, our method relatively performs well for feature screening and structure identification with practical computational costs. For over identification issues of our procedure, we recommend applying finite-dimensional structure identification methods developed by previous studies to the selected model by our procedure.

1.4.2 Real data example

In this subsection, we show the result of microarray data analysis to illustrate the use of our procedure. We used the R packages *Biobase* of Huber et al. (2015) and *seventyGeneData* of Marchionni et al. (2013) to apply our forward selection procedure to the gene expression data set of van't Veer et al. (2002). The data set consists of not only the 24481 gene expression status but also clinical information like the age of patients, the follow-up year of the patients, the diameter of the tumor, the tumor grade, the presence or non presence of the angioinvasion, the estrogen receptor alpha (ERα) status, the progesterone receptor (PR) status, the presence of the lymphocytic infiltrate, and the existence or non existence of the BRCA1 mutation for 117 lymph node negative patients. Among them, 34 patients developed distant metastases within five years (the bad prognosis group), 44 patients remained to be free from the metastases more than five years (the good prognosis

Table 1: (C_j, NC_j, NS_j) for $j = 1, \cdots, 4$, the average numbers of TP and FP, and computational times to finish 1000 repetitions.

(t_1, t_2)	Approach	C_1	NC_1	NS_1	C_2	NC_2	NS_2	C_3	NC_3	NS_3	C_4	NC_4	NS_4	TP	FP	Time (min)
(0,0)	forward	0.99	0.01	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	4.00	4.00	88.28
	gLasso	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	4.00	13.66	10.73
	gSCAD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	4.00	11.63	11.69
(2,0)	forward	1.00	0.00	0.00	0.17	0.83	0.00	0.19	0.81	0.00	0.00	1.00	0.00	4.00	2.67	72.51
	gLasso	1.00	0.00	0.00	0.98	0.02	0.00	0.99	0.01	0.00	0.28	0.32	0.40	3.60	7.44	11.04
	gSCAD	1.00	0.00	0.00	0.93	0.07	0.00	0.93	0.07	0.00	0.09	0.63	0.28	3.72	5.11	12.30
(2,1)	forward	0.99	0.01	0.00	0.49	0.51	0.00	0.65	0.35	0.00	0.00	1.00	0.00	4.00	3.09	68.29
	gLasso	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.31	0.33	0.36	3.64	7.18	11.03
	gSCAD	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.06	0.76	0.18	3.82	3.87	12.24
(3,1)	forward	0.95	0.05	0.00	0.49	0.50	0.01	0.87	0.13	0.00	0.00	1.00	0.00	3.98	2.82	61.60
	gLasso	1.00	0.00	0.00	0.99	0.00	0.01	1.00	0.00	0.00	0.28	0.23	0.49	3.50	9.39	11.34
	gSCAD	1.00	0.00	0.00	0.99	0.00	0.01	1.00	0.00	0.00	0.08	0.58	0.34	3.65	3.48	12.75
(4,5)	forward	0.96	0.03	0.01	0.50	0.33	0.17	0.86	0.14	0.00	0.00	0.98	0.02	3.80	2.68	58.91
	gLasso	1.00	0.00	0.00	0.98	0.00	0.02	1.00	0.00	0.00	0.36	0.15	0.49	3.49	12.62	11.43
	gSCAD	1.00	0.00	0.00	0.92	0.00	0.08	1.00	0.00	0.00	0.10	0.46	0.44	3.48	4.42	12.70
(6,8)	forward	0.80	0.02	0.18	0.31	0.12	0.58	0.88	0.06	0.06	0.00	0.66	0.34	2.84	3.23	54.42
	gLasso	0.98	0.00	0.02	0.84	0.00	0.16	0.99	0.00	0.01	0.21	0.03	0.76	3.05	15.29	11.46
	gSCAD	0.86	0.00	0.14	0.45	0.00	0.55	0.94	0.00	0.06	0.02	0.24	0.74	2.52	2.20	12.48

group), 18 patients had BRCA1 mutations, two patients had BRCA2 mutations, and 19 patients are for the test data.

There are two aims in this real data example. One is to select genes which are possibly relevant with developments of the tumor size. The other is to identify which genes possibly have constant effects and which genes possibly have non-constant effects on developments of the tumor size. We used samples of the bad prognosis group and the good prognosis group in this study. Thus, the sample size is 78, and we set the dimension of the orthonormal basis as five. Before applying our forward selection procedure to the microarray data, we have to deal with missing values. We removed genes which have too many missing values, and 24188 genes remained. To impute missing values, we implemented the R package *missMDA* of Josse and Husson (2016).

Researchers consider that ERα is significantly related to the development of breast cancer since the majority of breast cancer shows its amplification. Thus, we used ERα as the index variable to associate the interaction between its expressions and gene expressions with the tumor size. Note that we divided values of ERα by 100 to be in the unit interval. Cheng et al. (2016) also considered similar varying coefficient models and tried to select relevant genes. However, they did not try to reveal which genes have constant effects and which genes have non-constant effects on developments of the tumor size because structure identification is not their purpose.

To select possibly relevant genes and identify their effects on developments of the tumor size, we applied our forward selection procedure to the varying coefficient model

$$Y_i = \sum_{j=1}^{24188} g_j(Z_i)X_{ij} + \epsilon_i,$$

where Y_i is the tumor size, X_{ij} is the jth gene expression, Z_i is ERα expression, $g_j(z)$ is an unknown smooth function, and ϵ_i is the error term.

Our forward selection procedure selected 15 genes, and one gene and the others are classified into the group of constant effects and the group of non-constant effects, respectively. Table 2 lists the result of our procedure. These genes are possibly good candidates for further scientific research. For comparison, we also applied gLasso and gSCAD tuned by the BIC to the same model. gLasso selected 71 genes, and two of them are classified into the group of non-constant effects. gSCAD selected 38 genes, and all genes are classified into the group of constant effects. These two methods do not select genes which are selected by our procedure. Compared with our procedure, gLasso and gSCAD selected

Table 2: Genes selected by our forward selection procedure.

Constant effect		Non-constant effect	
NM_002457	NM_016096	Contig42370_RC	Contig11067_RC
	Contig21077_RC	Contig34430_RC	NM_004625
	NM_006159	NM_004054	NM_014694
	Contig29788_RC	AF070582	NM_001017
	NM_021055	Contig51414_RC	

many genes. However, if we choose their tuning parameters via EBIC, then no gene is selected by them. The results of these two methods are significantly dependent on their tuning parameters, and it is hard to select tuning parameters which can control the false positive and the false negative appropriately in practice, as this example shows.

1.5 Conclusion

In this paper, we proposed a forward selection procedure for simultaneous feature screening and structure identification in ultra-high dimensional varying coefficient models. Unlike existing feature screening procedures and forward selection procedures, our method can explicitly specify partially linear varying coefficient models. We established the screening consistency for our procedure with EBIC stopping rules. Simulation studies show that our procedure performs well in finite sample, and its computational cost is low. We also applied our procedure to real data which have 24188 covariates. These results support the utility of our procedure in real practice.

However, in this paper, we do not consider additive models, which are other important and widely used nonparametric models. Since varying coefficient models and additive models have similar structures, forward selection procedures for feature screening and structure identification in additive models may work well. Studying these methods may be an important topic for future research. In addition, as stated in Remark 1 of this paper, feature screening and structure identification in spatio-temporal varying coefficient models are also remarkable, and we can modify our method in this paper for achieving these two things by employing the bivariate orthonormal basis and the group Lasso with suitably separated penalties at the variable selection step. Investigating theoretical and numerical properties of the method may be also another important topic for future research.

1.6 Appendix

1.6.1 Proofs

To prove main theorems, we need Lemma 1.

Lemma 1. *Suppose that Assumptions 1-3 hold with $M = O(n^\xi)$ for some $0 < \xi < 1/10$. Then, there exists some sequence of positive numbers $\{\delta_{1n}\}$ tending to zero sufficiently slowly such that with probability tending to one, we have*

$$\frac{\tau_{\min}}{L}(1 - \delta_{1n}) \leq \lambda_{\min}(\widehat{\mathbf{\Sigma}}_S) \leq \lambda_{\max}(\widehat{\mathbf{\Sigma}}_S) \leq \frac{\tau_{\max}}{L}(1 + \delta_{1n}), \tag{1.8}$$

for any $S \subset \mathcal{F}$ satisfying $|S| \leq M$.

Proof of Lemma 1. Note that there exists some positive constant α such that $\xi + \alpha < 1/10$. Let $\delta_{1n} = n^{-\alpha}$. We will show

$$P\left(|\lambda_{\min}(\widehat{\mathbf{\Sigma}}_S) - \lambda_{\min}(\mathbf{\Sigma}_S)| > \frac{\tau_{\min}}{L}\delta_{1n}\right) \to 0. \tag{1.9}$$

Note that eigenvalues of $\widehat{\mathbf{\Sigma}}_S$ and $\mathbf{\Sigma}_S$ are nonnegative because these matrices are positive semidefinite. We also note that the kth largest singular value is equal to the kth largest eigenvalue for any positive semidefinite matrix. Hence, from Corollary 7.3.5 of Horn and Johnson (2012), we get

$$\left|\lambda_k(\widehat{\mathbf{\Sigma}}_S) - \lambda_k(\mathbf{\Sigma}_S)\right| \leq \|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\| \quad \text{for } k = 1, \ldots, |\mathcal{F}_c \cap S| + |\mathcal{F}_v \cap S|(L-1), \tag{1.10}$$

where $\{\lambda_k(\widehat{\mathbf{\Sigma}}_S)\}$ and $\{\lambda_k(\mathbf{\Sigma}_S)\}$ are non increasingly ordered eigenvalues of $\widehat{\mathbf{\Sigma}}_S$ and $\mathbf{\Sigma}_S$, respectively. Thus, we obtain

$$|\lambda_{\min}(\widehat{\mathbf{\Sigma}}_S) - \lambda_{\min}(\mathbf{\Sigma}_S)| \leq \|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|. \tag{1.11}$$

Note that the numbers of row and column of $\widehat{\mathbf{\Sigma}}_S$ and $\mathbf{\Sigma}_S$ are equal to $|\mathcal{F}_c \cap S| + |\mathcal{F}_v \cap S|(L-1)$. Since $|\mathcal{F}_v \cap S| \leq M$ and $|\mathcal{F}_c \cap S| \leq M$, we have

$$\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\| \leq ML\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_\infty$$

by the Cauchy Schwarz inequality, and

$$P\left(|\lambda_{\min}(\widehat{\mathbf{\Sigma}}_S)-\lambda_{\min}(\mathbf{\Sigma}_S)|>\frac{\tau_{\min}}{L}\delta_{1n}\right)\leq P\left(\|\widehat{\mathbf{\Sigma}}_S-\mathbf{\Sigma}_S\|_\infty>\frac{\tau_{\min}}{ML^2}\delta_{1n}\right), \tag{1.12}$$

for any S satisfying $|S|\leq M$. Under Assumption 3, Lemma A.1 and Lemma A.2 of Fan et al. (2014) hold. Thus, for $k,j=1,\ldots,p$, $s,l=1,\ldots,L$, and $m\geq 2$, we have

$$\begin{aligned} &E\left(|X_{ik}B_s(Z_i)B_l(Z_i)X_{ij}-E[X_{ik}B_s(Z_i)B_l(Z_i)X_{ij}]|^m\right)\\ &\qquad\leq 2^m eK^m m!E[|B_s(Z_i)B_l(Z_i)|^m], \end{aligned} \tag{1.13}$$

where K is a positive constant. Recall that $\boldsymbol{B}(z)=\boldsymbol{A}_0\boldsymbol{B}_0(z)$, and $C_3\leq\lambda_{\max}(\boldsymbol{A}_0^T\boldsymbol{A}_0)\leq C_4$ for some positive constants C_3 and C_4. See Appendix A of Honda and Yabe (2017). Hence, for $l=1,\ldots,L$, we can denote $B_l(z)=\sum_{s=1}^L a_{ls}B_{0s}(z)$ where a_{ls} is the (l,s) element of $\boldsymbol{A}_0$. Note that $\sum_{k=1}^L B_{0k}(z)=1$ due to the properties of the B-spline basis. Therefore, for $s,l=1,\ldots,L$, we obtain

$$\begin{aligned} |B_s(Z_i)B_l(Z_i)| &= \left|\left\{\sum_{j=1}^L a_{sj}B_{0j}(Z_i)\right\}\left\{\sum_{k=1}^L a_{lk}B_{0k}(Z_i)\right\}\right|\\ &\leq \max_{1\leq j\leq L}|a_{sj}|\times\max_{1\leq k\leq L}|a_{lk}|\\ &\leq \|\boldsymbol{A}_0\|_\infty^2\leq\|\boldsymbol{A}_0\|^2\leq C_4. \end{aligned}$$

As a result, we get $E[|B_s(Z_i)B_l(Z_i)|^m]\leq {C_4}^m$ and the right hand side of the last inequality in (1.13) can be bounded by $\leq(2KC_4)^{m-2}m!8K^2{C_4}^2e2^{-1}$ for any $m\geq 2$. Thus, from Bernstein's inequality (see Lemma 2.2.11 of van der Vaart and Wellner (1996)), we get

$$\begin{aligned} &P\left(\left|\frac{1}{n}\sum_{i=1}^n\{X_{ik}B_s(Z_i)B_l(Z_i)X_{ij}-E[X_{ik}B_s(Z_i)B_l(Z_i)X_{ij}]\}\right|>\frac{\tau_{\min}}{ML^2}\delta_{1n}\right)\\ &\qquad\leq 2\exp\left(\frac{-n\tau_{\min}\delta_{1n}}{16M^2L^4K^2C_4^2e(\tau_{\min}\delta_{1n})^{-1}+4ML^2KC_4}\right). \end{aligned} \tag{1.14}$$

Since $M=O(n^\xi)$, $M\leq\nu_m n^\xi$ for some $\nu_m>0$. Thus, the right hand side of (1.14) can be bounded by

$$\leq 2\exp\left(\frac{-\tau_{\min}n^{1/5-2\xi-2\alpha}}{K^*\{1+n^{-2/5-\xi-\alpha}\}}\right), \tag{1.15}$$

where $K^*=\max\{16K^2C_4^2e\nu_m^2C_L^4\tau_{\min}^{-1},4KC_4\nu_mC_L^2\}$. Let $A=\{j:1j\in S\}$. Note that

$\{j : -1j \in S\} \subset \{j : 1j \in S\}$. Since $|A| \leq |S| \leq M$, the right hand side of (1.12) can be further bounded by

$$\leq \sum_{k,j \in A} \sum_{1 \leq s,l \leq L} 2 \exp\left(\frac{-\tau_{\min} n^{1/5-2\xi-2\alpha}}{K^*\{1 + n^{-2/5-\xi-2\alpha}\}}\right)$$
$$\leq 2\nu_m^2 C_L^2 n^{2\xi+2/5} \exp\left(\frac{-\tau_{\min} n^{1/5-2\xi-2\alpha}}{K^*\{1 + n^{-2/5-\xi-\alpha}\}}\right). \tag{1.16}$$

The right hand side of the last inequality in (1.16) converges to zero as $n \to \infty$. Thus, (1.9) holds. We can also get

$$P\left(|\lambda_{\max}(\widehat{\mathbf{\Sigma}}_S) - \lambda_{\max}(\mathbf{\Sigma}_S)| > \frac{\tau_{\max}}{L}\delta_{1n}\right) \to 0, \tag{1.17}$$

in the same way. (1.9), (1.17), and Assumption 2 imply

$$P\left(\frac{\tau_{\min}}{L}(1-\delta_{1n}) \leq \lambda_{\min}(\widehat{\mathbf{\Sigma}}_S) \leq \lambda_{\max}(\widehat{\mathbf{\Sigma}}_S) \leq \frac{\tau_{\max}}{L}(1+\delta_{1n})\right) \to 1, \tag{1.18}$$

for any S satisfying $|S| \leq M$. This completes the proof of Lemma 1. $\square$

Corollary 1 is derived by Lemma 1 immediately.

Corollary 1. *Suppose that Assumptions 1-3 hold with $M = O(n^\xi)$ for some $0 < \xi < 1/10$. Then, with probability tending to one, we have*

$$\frac{\tau_{\min}}{L}(1-\delta_{1n}) \leq \lambda_{\min}\left(\frac{1}{n}\boldsymbol{W}_E^T \boldsymbol{Q}_S \boldsymbol{W}_E\right) \leq \lambda_{\max}\left(\frac{1}{n}\boldsymbol{W}_E^T \boldsymbol{Q}_S \boldsymbol{W}_E\right) \leq \frac{\tau_{\max}}{L}(1+\delta_{1n}), \tag{1.19}$$

for any $S \subset \mathcal{F}$ and $E \subset \mathcal{F}$ satisfying $|E \cup S| \leq M$, where $\boldsymbol{Q}_S$ is the orthogonal projection defined in (1.26).

Proof of Corollary 1. Note that $\boldsymbol{W}_{E\cup S} = (\boldsymbol{W}_E, \boldsymbol{W}_S)$. By (A-74) in Greene (2012), $(\boldsymbol{W}_E^T \boldsymbol{Q}_S \boldsymbol{W}_E)^{-1}$ is the upper left block of $(\boldsymbol{W}_{E\cup S}^T \boldsymbol{W}_{E\cup S})^{-1}$. Notice that $\lambda_{\max}(\boldsymbol{A}^{-1}) = \lambda_{\min}(\boldsymbol{A})^{-1}$ and $\lambda_{\min}(\boldsymbol{A}^{-1}) = \lambda_{\max}(\boldsymbol{A})^{-1}$ for any symmetric and invertible matrix $\boldsymbol{A}$. Hence, we get

$$\lambda_{\min}(\boldsymbol{W}_{E\cup S}^T \boldsymbol{W}_{E\cup S}) \leq \lambda_{\min}(\boldsymbol{W}_E^T \boldsymbol{Q}_S \boldsymbol{W}_E),$$

and

$$\lambda_{\max}(\boldsymbol{W}_E^T \boldsymbol{Q}_S \boldsymbol{W}_E) \leq \lambda_{\max}(\boldsymbol{W}_{E\cup S}^T \boldsymbol{W}_{E\cup S}).$$

By Lemma 1, we have

$$\frac{n\tau_{\min}}{L}(1-\delta_{1n}) \le \lambda_{\min}(\boldsymbol{W}_{E\cup S}^T \boldsymbol{W}_{E\cup S}) \le \lambda_{\max}(\boldsymbol{W}_{E\cup S}^T \boldsymbol{W}_{E\cup S}) \le \frac{n\tau_{\max}}{L}(1+\delta_{1n}) \quad (1.20)$$

with probability tending to one. Hence, the proof of Corollary 1 is completed. □

Proof of Theorem 1. We can prove Theorem 1 almost in the same way as in Cheng et al. (2016). However, since there are some differences because we are also considering structure identification, we give the proof of Theorem 1 below. Firstly, we can represent (1.1) by

$$\begin{aligned} Y_i &= \sum_{j=1}^{p} X_{ij} g_{cj} + \sum_{j=1}^{p} X_{ij} g_{vj}(Z_i) + \epsilon_i \\ &= \sum_{j\in\mathcal{S}_c} X_{ij} g_{cj} + \sum_{j\in\mathcal{S}_v} X_{ij} g_{vj}(Z_i) + \epsilon_i \\ &= \sum_{j\in\mathcal{S}_c} X_{ij} B_1(Z_i)\gamma_{1j} + \sum_{j\in\mathcal{S}_v} X_{ij} \boldsymbol{B}_{-1}(Z_i)^T \boldsymbol{\gamma}_{-1j} + \delta_i + \epsilon_i, \end{aligned} \quad (1.21)$$

where

$$\delta_i = \sum_{j\in\mathcal{S}_v} X_{ij}\{g_{vj}(Z_i) - \boldsymbol{B}_{-1}(Z_i)^T \boldsymbol{\gamma}_{-1j}\}, \quad (1.22)$$

as in Honda et al. (2019). If $\boldsymbol{\gamma}_j$ is suitably chosen, there exists some positive C_g' such that

$$\sum_{j=1}^{p} \|g_j - \boldsymbol{B}^T \boldsymbol{\gamma}_j\|_\infty \le C_g' L^{-2}, \quad (1.23)$$

under Assumption 4. Note that $|g_{cj}|^2 = |\gamma_{1j}|^2 L^{-1}$ and $\|g_{vj}\|^2 = \|\boldsymbol{\gamma}_{-1j}\|^2 L^{-1} + O(L^{-4})$. See Appendix A of Honda and Yabe (2017) for properties of the orthonormal basis. Then, there exist some positive constants C_1 and C_2 such that $C_1 L\|g_{vj}\|^2 \le \|\boldsymbol{\gamma}_{-1j}\|^2 \le C_2 L\|g_{vj}\|^2$. Under Assumption 6, we have

$$|\delta_i| \le C_X C_g' L^{-2} \to 0. \quad (1.24)$$

The matrix form of (1.21) is represented by

$$\boldsymbol{Y} = \sum_{j\in\mathcal{S}_c} \boldsymbol{W}_{1j}\gamma_{1j} + \sum_{j\in\mathcal{S}_v} \boldsymbol{W}_{-1j}\boldsymbol{\gamma}_{-1j} + \boldsymbol{\delta} + \boldsymbol{\epsilon}$$

$$= \boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0} + \boldsymbol{\delta} + \boldsymbol{\epsilon}.$$

Let us denote

$$\begin{aligned}
\boldsymbol{P}_S &= \boldsymbol{W}_S(\boldsymbol{W}_S^T\boldsymbol{W}_S)^{-1}\boldsymbol{W}_S^T, \\
\boldsymbol{Q}_S &= \boldsymbol{I} - \boldsymbol{P}_S, \\
\widetilde{\boldsymbol{W}}_{lS} &= \boldsymbol{Q}_S\boldsymbol{W}_l, && (1.25)\\
\boldsymbol{H}_{lS} &= \widetilde{\boldsymbol{W}}_{lS}(\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1}\widetilde{\boldsymbol{W}}_{lS}^T, && (1.26)
\end{aligned}$$

for any S satisfying $|S \cup \mathcal{S}_0| \leq M$ and $\mathcal{S}_0 \not\subset S$. We can easily show that $\boldsymbol{H}_{lS}\boldsymbol{Q}_S = \boldsymbol{H}_{lS}$ and $n\hat{\sigma}_S^2 - n\hat{\sigma}_{S(l)}^2 = \|\boldsymbol{H}_{lS}\boldsymbol{Y}\|^2$. From the triangle inequality, we have

$$\max_{l\in\mathcal{S}_0\cap S^c} \|\boldsymbol{H}_{lS}\boldsymbol{Y}\| \geq \max_{l\in\mathcal{S}_0\cap S^c} \|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\| - \|\boldsymbol{H}_{lS}\boldsymbol{\delta}\| - \|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\|. \quad (1.27)$$

We evaluate each part in the right hand side of (1.27). Since $\boldsymbol{H}_{lS}$ is a projection matrix, we have

$$\|\boldsymbol{H}_{lS}\boldsymbol{\delta}\| \leq \|\boldsymbol{\delta}\| \leq \sqrt{n}C_X C_g' L^{-2}. \quad (1.28)$$

Hence, we have $\|\boldsymbol{H}_{lS}\boldsymbol{\delta}\| = O_P(n^{1/10})$. Next, we evaluate the third term in (1.27). Under Assumption 5, Proposition 3 of Zhang (2010) holds. Thus, we have

$$P\left(\frac{\|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\|^2}{m\sigma_1^2} \geq \frac{1+x}{\{1 - 2/(e^{x/2}\sqrt{1+x} - 1)\}_+^2}\right) \leq e^{-mx/2}(1+x)^{m/2}, \quad (1.29)$$

for all $x > 0$ and for any $l \in \mathcal{F} \cap S^c$, where $m = \mathrm{rank}(\boldsymbol{H}_{lS})$. We take $x = \log n$. If $l \in \mathcal{F}_c$, then $\mathrm{rank}(\boldsymbol{H}_{lS}) = 1$ and $\|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\|^2 = O_P(\log n)$, which implies $\|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\|^2 = O_P(L \log n)$. If $l \in \mathcal{F}_v$, then $\mathrm{rank}(\boldsymbol{H}_{lS}) = (L-1)$ and $\|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\|^2 = O_P(L\log n)$. Thus, we have

$$\|\boldsymbol{H}_{lS}\boldsymbol{\epsilon}\| = O_P(L^{1/2}(\log n)^{1/2}) = O_P(n^{1/10}(\log n)^{1/2}). \quad (1.30)$$

Next, we evaluate $\max_{l\in\mathcal{S}_0\cap S^c} \|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|$. Notice that

$$\|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \geq \lambda_{\min}((\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1})\|\widetilde{\boldsymbol{W}}_{lS}^T\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2. \quad (1.31)$$

Note that $\lambda_{\min}((\widetilde{\boldsymbol{W}}_{lS}^T\widetilde{\boldsymbol{W}}_{lS})^{-1}) = \lambda_{\min}((\boldsymbol{W}_l^T\boldsymbol{Q}_S\boldsymbol{W}_l)^{-1})$ and $|\{l\}\cup S| \leq M$. From Corollary

1, we get

$$\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \geq \frac{L}{n\tau_{\max}(1+\delta_{1n})}\max_{l\in\mathcal{S}_0\cap S^c}\|\widetilde{\boldsymbol{W}}_{lS}^T\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2, \tag{1.32}$$

with probability tending to one. We evaluate $\max_{l\in\mathcal{S}_0\cap S^c}\|\widetilde{\boldsymbol{W}}_{lS}^T\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2$. We can see that

$$\max_{l\in\mathcal{S}_0\cap S^c}\|\widetilde{\boldsymbol{W}}_{lS}^T\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 = \max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{W}_l^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2.$$

Notice that $|g_{cj}|^2+\|g_{vj}\|_2^2=\|g_j\|_2^2\leq\|g_j\|_\infty^2$. From Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&\|\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \leq \|\boldsymbol{\gamma}_{\mathcal{S}_0}\|\|\boldsymbol{W}_{\mathcal{S}_0}^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\| \\
&= \left(\sum_{j\in\mathcal{S}_c}|\gamma_{1j}|^2+\sum_{j\in\mathcal{S}_v}\|\boldsymbol{\gamma}_{-1j}\|^2\right)^{1/2}\left(\sum_{j\in\mathcal{S}_0\cap S^c}\|\boldsymbol{W}_j^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2\right)^{1/2} \\
&\leq \left(L\sum_{j\in\mathcal{S}_c}|g_{cj}|^2+C_2L\sum_{j\in\mathcal{S}_v}\|g_{vj}\|_2^2\right)^{1/2}\left(\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{W}_l^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 2p_0\right)^{1/2} \\
&\leq 2^{1/2}p_0^{1/2}L^{1/2}C_g(1+C_2)^{1/2}\left(\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{W}_l^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2\right)^{1/2},
\end{aligned}$$

with probability tending to one. Thus, we get

$$\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{W}_l^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \geq \frac{\|\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^4}{2p_0LC_g^2(1+C_2)}. \tag{1.33}$$

Since $|\mathcal{S}_0\cup S|\leq M$, we have

$$\begin{aligned}
\|\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 &\geq \lambda_{\min}\left(\boldsymbol{W}_{\mathcal{S}_0}^T\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\right)\|\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \\
&\geq \frac{n\tau_{\min}}{L}(1-\delta_{1n})\|\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2,
\end{aligned}$$

by Corollary 1. Notice that

$$\begin{aligned}
\|\boldsymbol{Q}_S\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^4 &\geq \frac{n^2\tau_{\min}^2}{L^2}(1-\delta_{1n})^2\|\boldsymbol{\gamma}_{\mathcal{S}_0}\|^4 \\
&\geq \frac{n^2\tau_{\min}^2}{L^2}(1-\delta_{1n})^2\left(L\max_{j\in\mathcal{S}_c}|g_{cj}|^2+LC_1\max_{j\in\mathcal{S}_v}\|g_{vj}\|_2^2\right)^2 \\
&\geq n^2\tau_{\min}^2(1-\delta_{1n})^2C_{\min}^2(1+C_1)^2,
\end{aligned} \tag{1.34}$$

with probability tending to one. Note that $\min\{\max_{j\in\mathcal{S}_c}|g_{cj}|^2, \max_{j\in\mathcal{S}_v}\|g_{vj}\|_2^2\} > C_{\min}$ under Assumption 4. From (1.32), (1.33), and (1.34), we get

$$\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2 \geq \frac{n(1-\delta_{1n})^2}{(1+\delta_{1n})}\frac{\tau_{\min}^2 C_{\min}^2(1+C_1)^2}{\tau_{\max}C_g^2(1+C_2)2p_0}, \tag{1.35}$$

with probability tending to one. (1.35) implies that the first term in the right hand side of (1.27) dominates the other two terms. Thus, we have

$$\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{H}_{lS}\boldsymbol{Y}\|^2 \geq 4^{-1}\max_{l\in\mathcal{S}_0\cap S^c}\|\boldsymbol{H}_{lS}\boldsymbol{W}_{\mathcal{S}_0}\boldsymbol{\gamma}_{\mathcal{S}_0}\|^2, \tag{1.36}$$

with probability tending to one. By taking $\{\delta_{2n}\}$ as $\delta_{2n} = 1-(1-\delta_{1n})^2(1+\delta_{1n})^{-1}$, we get

$$\begin{aligned}\max_{l\in\mathcal{F}\cap S^c}\{n\hat{\sigma}_S^2 - n\hat{\sigma}_{S\cup\{l\}}^2\} &\geq n(1-\delta_{2n})\frac{\tau_{\min}^2 C_{\min}^2(1+C_1)^2}{4\tau_{\max}C_g^2(1+C_2)2p_0}\\ &= n(1-\delta_{2n})D^*,\end{aligned}$$

for any S satisfying $|S\cup\mathcal{S}_0| \leq M$ with probability tending to one, and the proof of Theorem 1 is completed. □

Proof of Theorem 2. Assume Theorem 2 is false. Then, we have $\mathcal{S}_v \not\subset F_k$ or $\mathcal{S}_c \not\subset E_k$ for $k = 1,\cdots,T^*$. This implies that $\mathcal{S}_0 \not\subset S_k$ and $|S_k\cup\mathcal{S}_0| \leq M$ for $k=1,\cdots,T^*$. Note that we have $n^{-1}\sum_{i=1}^n Y_i^2 \geq \hat{\sigma}_{S_k(l)}^2$ and $\hat{\sigma}_{S_k}^2 - \hat{\sigma}_{S_k\cup\{l\}}^2 \leq n^{-1}\sum_{i=1}^n Y_i^2$ for any $l\in\mathcal{F}\cap S_k^c$, and $\log(1+x) \geq x(1+x)^{-1}$ for all $x > -1$. Hence, we have

$$\begin{aligned}&\mathrm{EBIC}(S_k) - \mathrm{EBIC}(S_k\cup\{l\}) = n\log\left(\frac{\hat{\sigma}_{S_k}^2}{\hat{\sigma}_{S_k\cup\{l\}}^2}\right) - (\log n + 2\eta\log p)\\ &\geq n\log\left(1+\frac{\hat{\sigma}_{S_k}^2 - \hat{\sigma}_{S_k\cup\{l\}}^2}{\frac{1}{n}\sum_{i=1}^n Y_i^2}\right) - (\log n + 2\eta\log p)\\ &\geq \frac{n\hat{\sigma}_{S_k}^2 - n\hat{\sigma}_{S_k\cup\{l\}}^2}{\frac{2}{n}\sum_{i=1}^n Y_i^2} - (\log n + 2\eta\log p),\end{aligned}$$

for any $l\in\mathcal{F}_c\cap S_k^c$. Hence, if $j^*\in\mathcal{F}_c\cap S_k^c$, we obtain

$$\mathrm{EBIC}(S_k) - \mathrm{EBIC}(S_k\cup\{j^*\}) \geq \frac{\max_{l\in\mathcal{F}\cap S_k^c}\{n\hat{\sigma}_{S_k}^2 - n\hat{\sigma}_{S_k\cup\{l\}}^2\}}{\frac{2}{n}\sum_{i=1}^n Y_i^2} - (\log n + 2\eta\log p), \tag{1.37}$$

since $j^* = \underset{l \in \mathcal{F} \cap S_k^c}{\operatorname{argmin}} \hat{\sigma}^2_{S_k \cup \{l\}}$. If $j^* \in \mathcal{F}_v \cap S_k^c$, we can also show that

$$\mathrm{EBIC}(S_k) - \mathrm{EBIC}(S_k \cup \{j^*\}) \geq \frac{\max_{l \in \mathcal{F} \cap S_k^c} \{n\hat{\sigma}^2_{S_k} - n\hat{\sigma}^2_{S_k \cup \{l\}}\}}{\frac{2}{n}\sum_{i=1}^n Y_i^2} - (L-1)(\log n + 2\eta \log p), \tag{1.38}$$

in the same way. The second terms of the right hand side in (1.37) and (1.38) are $o(n)$ under Assumption 1. In addition, we have $\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{P} E[Y^2]$. Therefore, Theorem 1 implies that the first terms of the right hand side in (1.37) and (1.38) dominate the second terms, and we get

$$\mathrm{EBIC}(S_k) \geq \mathrm{EBIC}(S_k \cup \{j^*\}) \quad \text{for } k = 1, \cdots, T^*, \tag{1.39}$$

with probability tending to one. (1.39) implies that our forward selection procedure has selected some index at each kth step for $k = 1, \cdots, T^* + 1$. Thus, we have the monotone increasing sequence and the non increasing sequence $\{S_k\}_{k=1}^{T^*+1}$ and $\{\hat{\sigma}_{S_k}\}_{k=1}^{T^*+1}$, respectively. Note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 &\geq \hat{\sigma}^2_{S_1} - \hat{\sigma}^2_{S_{T^*+1}} \\
&= \sum_{k=1}^{T^*} (\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_{k+1}}) \\
&= \sum_{k=1}^{T^*} \max_{l \in \mathcal{F} \cap S_k^c} \{\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{l\}}\},
\end{aligned}$$

where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$. Theorem 1 implies $\mathrm{Var}(Y) \geq \sum_{k=1}^{T^*}(1-\delta_{2n})D^* = T^*(1-\delta_{2n})D^*$ with probability tending to one. Thus, we have

$$T^* \leq \frac{\mathrm{Var}(Y)}{(1-\delta_{2n})D^*} < \frac{\mathrm{Var}(Y)}{(1-2\delta_{2n})D^*} \leq T^*, \tag{1.40}$$

with probability tending to one, and a contradiction occurs. This completes the proof. $\square$

Proof of Theorem 3. Since $\mathcal{S}_0 \subset S_k$, we have

$$\hat{\sigma}^2_{S_k} = n^{-1}\|\boldsymbol{Q}_{S_k}\boldsymbol{\delta}\|^2 + 2n^{-1}\boldsymbol{\delta}^T\boldsymbol{Q}_{S_k}\boldsymbol{\epsilon} + n^{-1}\|\boldsymbol{Q}_{S_k}\boldsymbol{\epsilon}\|^2.$$

We can see that the first term and the second term of the right hand side are $o_P(1)$ by

(1.28) and Cauchy-Schwarz inequality. Note that

$$\|\boldsymbol{Q}_{S_k}\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{\epsilon}\|^2 - \|\boldsymbol{P}_{S_k}\boldsymbol{\epsilon}\|^2,$$

and

$$\|\boldsymbol{P}_{S_k}\boldsymbol{\epsilon}\|^2 = O_P(m \log n),$$

by Proposition 3 of Zhang (2010), where $m = \text{rank}(\boldsymbol{P}_{S_k}) = |\mathcal{F}_c \cap S_k| + |\mathcal{F}_v \cap S_k|(L-1)$. Notice that $|S_k| \le 2k \le 2M$, and $M \le \nu_m n^{\xi}$ for some $\nu_m > 0$, since $M = O(n^{\xi})$. Thus, we have

$$n^{-1} m \log n \le n^{-1} 2ML \log n \le 2C_L \nu_m n^{\xi+1/5-1} \log n \to 0$$

as $n \to \infty$, and we get $\|\boldsymbol{P}_{S_k}\|^2/n = o_P(1)$. Since $\|\boldsymbol{\epsilon}\|^2/n \xrightarrow{P} \sigma^2$, we obtain $\hat{\sigma}^2_{S_k} = \sigma^2 + o_P(1)$. We can also show that $\hat{\sigma}^2_{S_k \cup \{l\}} = \sigma^2 + o_P(1)$ for any $l \in \mathcal{F} \cap S_k^c$ in the same way. Hence, we have $\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{l\}} = o_P(1)$ for any $l \in \mathcal{F} \cap S_k^c$.

Now, let $j^* \in \mathcal{F}_c \cap S_k^c$. Since $-1 < -(\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{j^*\}})(\hat{\sigma}^2_{S_k})^{-1}$ and $\log(1+x) \ge x(1+x)^{-1}$ for all $x > -1$, we get

$$\begin{aligned}
&\text{EBIC}(S_k \cup \{j^*\}) - \text{EBIC}(S_k) \qquad (1.41)\\
&= n \log\left(1 - \frac{\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{j^*\}}}{\hat{\sigma}^2_{S_k}}\right) + (\log n + 2\eta \log p)\\
&\ge \left(-\frac{n\hat{\sigma}^2_{S_k} - n\hat{\sigma}^2_{S_k \cup \{j^*\}}}{\hat{\sigma}^2_{S_k}}\right)\left(1 - \frac{\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{j^*\}}}{\hat{\sigma}^2_{S_k}}\right)^{-1} + (\log n + 2\eta \log p)\\
&= \left(-\frac{\max_{l \in \mathcal{F} \cap S_k^c}\{n\hat{\sigma}^2_{S_k} - n\hat{\sigma}^2_{S_k \cup \{l\}}\}}{\hat{\sigma}^2_{S_k}}\right)\left(1 - \frac{\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{j^*\}}}{\hat{\sigma}^2_{S_k}}\right)^{-1} + (\log n + 2\eta \log p).
\end{aligned} \tag{1.42}$$

Note that $\|\boldsymbol{H}_{lS_k}\boldsymbol{Y}\|^2 \le 2\|\boldsymbol{H}_{lS_k}\boldsymbol{\delta}\|^2 + 2\|\boldsymbol{H}_{lS_k}\boldsymbol{\epsilon}\|^2$. From (1.28) and Proposition 3 of Zhang (2010), we have

$$\max_{l \in \mathcal{F} \cap S_k^c}\{n\hat{\sigma}^2_{S_k} - n\hat{\sigma}^2_{S_k \cup \{l\}}\} = O_P(n^{1/5} \log n). \tag{1.43}$$

The first term of the right hand side in the last equality of (1.42) is $O_P(n^{1/5} \log n)$ because its denominator converges to one in probability. On the other hand, the second term in the same equality is $O_P(n^{\xi_0})$, and $O_P(n^{1/5-\xi_0} \log n) = o_P(1)$. Hence, we have

$$
\begin{aligned}
\mathrm{EBIC}(S_k \cup \{j^*\}) - \mathrm{EBIC}(S_k) &\geq (\log n + 2\eta \log p)\{1 - O_P(n^{1/5-\xi_0}\log n)\} \\
&= (\log n + 2\eta \log p)\{1 - o_P(1)\}, \qquad (1.44)
\end{aligned}
$$

when $j^* \in \mathcal{F}_c \cap S_k^c$. Next, let $j^* \in \mathcal{F}_v \cap S_k^c$. By the same argument, we get

$$
\begin{aligned}
&\mathrm{EBIC}(S_k \cup \{j^*\}) - \mathrm{EBIC}(S_k) \\
&\geq \left(-\frac{\max_{l \in \mathcal{F} \cap S_k^c}\{n\hat{\sigma}^2_{S_k} - n\hat{\sigma}^2_{S_k \cup \{l\}}\}}{\hat{\sigma}^2_{S_k}}\right)\left(1 - \frac{\hat{\sigma}^2_{S_k} - \hat{\sigma}^2_{S_k \cup \{j^*\}}}{\hat{\sigma}^2_{S_k}}\right)^{-1} \\
&+ (L-1)(\log n + 2\eta \log p) \\
&= (L-1)(\log n + 2\eta \log p)\{1 - O_P(n^{-\xi_0}\log n)\} \\
&= (L-1)(\log n + 2\eta \log p)\{1 - o_P(1)\}. \qquad (1.45)
\end{aligned}
$$

(1.44) and (1.45) imply

$$
\mathrm{EBIC}(S_k \cup \{j^*\}) - \mathrm{EBIC}(S_k) > 0 \quad \text{for } j^* \in \mathcal{F} \cap S_k^c, \qquad (1.46)
$$

with probability tending to one. (1.46) implies that our forward selection procedure stops variable selection at the kth step. The proof of Theorem 3 is completed. $\square$

Chapter 2

Small Tuning Parameter Selection for the Debiased Lasso

2.1 Introduction

In this study, we make inferences about individual coefficients of the linear regression model:

$$Y = X\beta_0 + \epsilon, \tag{2.1}$$

where $Y \in \mathbb{R}^n$ denotes the response, $X = (X_1, \ldots, X_p) \in \mathbb{R}^{n \times p}$ denotes the random design matrix, $\beta_0 \in \mathbb{R}^p$ denotes the unknown regression coefficient, and $\epsilon \in \mathbb{R}^n \sim N(0, \sigma_\epsilon^2 I)$ denotes the unobservable noise that is independent of X. We consider $p > n$ and assume that X has zero mean and a positive definite covariance matrix $\Sigma = E[X^T X/n]$. We focus on statistical inferences for β_{01}, the first component of β_0, although the inferences are equally valid for other components.

Much research has been conducted on high-dimensional linear regression models. The Lasso (Tibshirani 1996) is one of the most popular methods for estimating their coefficients. The popularity is due to variable selection properties, the oracle inequalities (Bickel et al. 2009, Bühlmann and van de Geer 2011), and useful computational algorithms (Efron et al. 2004, Friedman et al. 2010). Bühlmann and van de Geer (2011) gave excellent overviews of the Lasso. However, the asymptotic distribution of the Lasso is generally intractable because the Lasso is biased and the distribution is not continuous (Knight and Fu 2000). Therefore, it is difficult to perform statistical inference based on the Lasso itself, and there are currently three main approaches of performing statistical inference in high-dimensional linear regression models.

The first approach of inference is the so-called selective inference. See Berk et al. (2013), Lee et al. (2016), Tibshirani et al. (2016), and Tibshirani et al. (2018), among others. Statistical inferences after model selection are distorted if the selected model is treated as if it is a true one. Rather than making inferences on coefficients of the true regression model, the selective inference makes inferences on coefficients in a model selected via data-driven procedures. The validity of the inferences does not depend on the correctness of the selected model.

The second approach is the orthogonalization approach, which is developed in econometrics. Some contributions include the post-double selection method of Belloni et al. (2014), projection onto the double selection method of Wang et al. (2020), and double machine learning method of Chernozhukov et al. (2018). Such methods perform inference on low-dimensional parameters when high-dimensional nuisance parameters appear in regression models. The estimators of the parameter of interest are derived through the orthogonal score function, which is insensitive to the estimation of nuisance parameters.

The third approach is to debias or desparsify the Lasso, which we consider in this study. The debiased Lasso was proposed by Zhang and Zhang (2014) and further developed by van de Geer et al. (2014) and Javanmard and Montanari (2014). Unlike selective inference, this approach is intended to make inferences on coefficients in the true linear regression model. If the true data generating process is sufficiently sparse, the estimator is asymptotically normally distributed, so that inference is possible even when $p > n$.

The performance of the debiased Lasso depends on the degree of sparsity and the choice of tuning parameters. To obtain the debiased Lasso estimator of β_{01}, we need an estimate of Θ_1, the first column of Σ^{-1}, and the node-wise Lasso (Meinshausen and Bühlmann 2006) is commonly used to obtain it. Therefore, the debiased Lasso estimator depends on the tuning parameters of the Lasso λ_0 and node-wise Lasso λ_1. Let s_0 and s_1 be the numbers of nonzero components in β_0 and Θ_1, respectively. If both λ_0 and λ_1 are of order $\sqrt{\log p/n}$, the asymptotic normality of the debiased Lasso is shown for $s_0 = o(\sqrt{n}/\log p)$ and $s_1 = o(n/\log p)$ (van de Geer et al. 2014, Zhang and Zhang 2014) or for $s_0 = o(n/\|\Theta_1\|_1^2(\log p)^2)$ and $\min\{s_0, s_1\} = o(\sqrt{n}/\log p)$ (Javanmard and Montanari 2018). Recently, Bellec and Zhang (2022) showed asymptotic normality of the debiased Lasso when $\max\{s_0, s_1\} = o(n/\log p)$ and $\min\{s_0, s_1\} = o(\sqrt{n}/\log p)$ using degrees of freedom adjustments.

Although the above studies significantly contributed to the development of the debiased Lasso, there are two main issues to be addressed. First, the bias of the debiased Lasso may not be asymptotically negligible if Θ_1 and β_0 are not sufficiently sparse. If no sparsity condition is imposed on Θ_1, s_0 must satisfy $s_0 = o(\sqrt{n}/\log p)$ for zero-mean asymptotic normality, which is rather restrictive given that only $s_0 = o(n/\log p)$ is need for consistent estimation. Second, existing tuning parameter selection procedures of λ_1, such as the cross-validation (Chetverikov et al. 2021) and the methods of Zhang and Zhang (2014) and Dezeure et al. (2015), often produce a rather large bias in the debiased Lasso, yielding a poor coverage of confidence intervals and large size distortion of t tests.

We address these two issues by selecting a smaller tuning parameter of the node-wise Lasso than in prior studies. First, we show that by setting the order of λ_1 to $1/\sqrt{n}$, the bias of the debiased Lasso can be reduced without diverging the asymptotic variance, so that asymptotic normality holds provided $s_0 = o(\sqrt{n/\log p})$. The novelty of our method is that we intentionally overfit the node-wise Lasso to reduce the bias of the debiased Lasso. Further, our analysis does not impose any sparsity assumption on Θ_1. To the best of our knowledge, there is no asymptotic normality result of the debiased Lasso when the order of λ_1 is $1/\sqrt{n}$. Moreover, with the exception of van de Geer (2019), no prior study, including Belloni et al. (2014) and Chernozhukov et al. (2018), has established asymptotic normality without imposing any sparsity condition on Θ_1. A sufficient condition for our result is that each row vector of X is independent Gaussian and $p \leq Cn$ for some $C < \infty$. Although the condition that p and n be of the same order excludes some high-dimensional data, such as genomic data, various data fit the condition in many fields, such as economics.

Second, we propose a data-driven procedure to select λ_1, which we call the small tuning parameter selector (STPS). We modify the selection method of Zhang and Zhang (2014) so that the selector chooses a tuning parameter with a reasonably small value. Figure 1 compares the distributions of the debiased Lasso whose λ_1 is selected by cross-validation and STPS, respectively, indicating that STPS successfully reduces the bias. Simulation studies show that the debiased Lasso tuned by our procedure outperforms other inference methods, such as the double machine learning, post-double selection, and debiased Lasso with other tuning parameter selection methods. This is particularly true when β_0 and Θ_1 are dense. We also show a real economic data example to see the performance of our method. Because the true sparsity of β_0 and Θ_1 is unknown, it is essential to propose a selection method that is robust to the violation of the sparsity condition.

We emphasize that the aim of this study is not to obtain an asymptotically efficient estimator (see, e.g., van de Geer et al. 2014, Jankova and van de Geer 2018, van de Geer 2019, and Bellec and Zhang 2022 about the asymptotic efficiency of the debiased Lasso). The aim is to construct confidence intervals with good coverage properties or test statistics with the correct size. There is a bias–variance tradeoff with respect to λ_1.

(a) 10-fold cross-validation. (b) STPS.

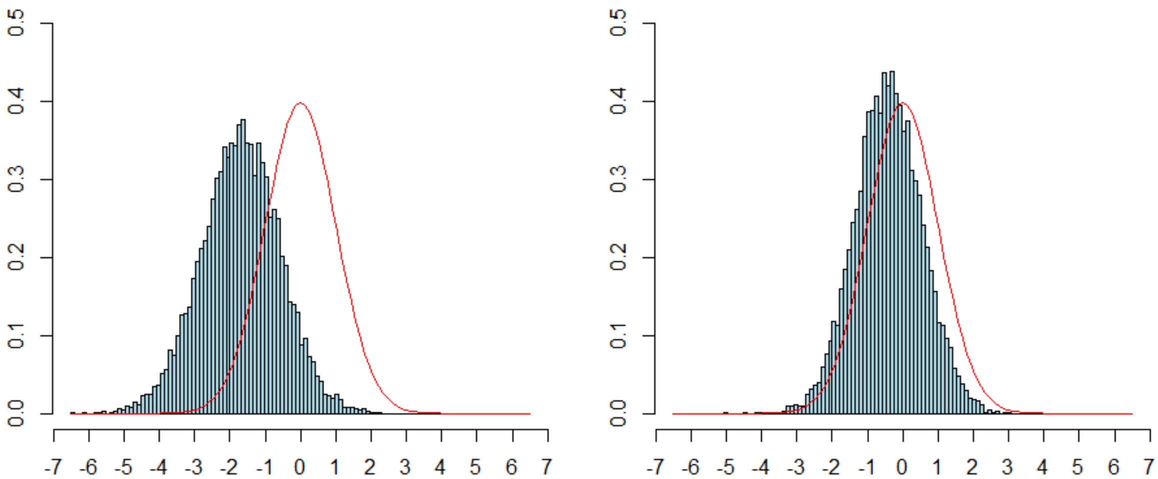

Figure 2.1: Distributions of the studentized debiased Lasso (histogram) and the standard normal distribution (solid line). We employ the simulation design II of Table II in Kozbur (2020) with $n = 200$ and $p = 400$ and repeat experiments 10,000 times. The tuning parameter λ_0 of the Lasso to be debiased is selected by the 10-fold cross-validation. The error variance is estimated by the sample mean of squared residuals from the Lasso tuned by the one standard error rule. For panel (a), the 10-fold cross-validation selects 0.13 on average, and the coverage of 95% confidence interval is 59.7% with the average length of 0.208. For panel (b), STPS selects 0.001 on average, and the coverage of 95% confidence interval is 94.3% with the average length of 0.459.

Our method reduces the bias at the cost of slightly increasing the variance. Setting the order of λ_1 to $1/\sqrt{n}$ maintains the $\sqrt{n}$-consistency of the debiased Lasso, although the asymptotic variance may exceed the efficiency bound.

The remainder of this article is organized as follows. In Section 2.2, we formally state an asymptotic normality result of the debiased Lasso when the order of λ_1 is $1/\sqrt{n}$. In Section 2.3, we propose a data-driven method to select λ_1. In Section 2.4, we show results of extensive simulation studies for the debiased Lasso tuned by our procedure and other statistical inference methods. We present a real economic data example based on the considered methods in Section 2.5. Finally, concluding remarks are presented in Section 2.6. All proofs are given in Section 2.7.

2.2 Theoretical Results

We define $|S|$ as the cardinality of a set S. Let a^T be the transpose of a vector a. Let $\|a\|$, $\|a\|_1$, and $\|a\|_\infty$ be the Euclidean norm, ℓ_1 norm, and maximum element in the absolute value of a vector a, respectively. For any sequences of real numbers $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means $C_1 \leq a_n/b_n \leq C_2$ for all n and some $C_1, C_2 > 0$. Let $a_n \ll b_n$ and $b_n \gg a_n$ denote $\lim_{n\to\infty} |a_n/b_n| = 0$. Let $a_n \lesssim b_n$ and $b_n \gtrsim a_n$ denote $a_n \leq Cb_n$ for all n and some $C > 0$. Let us denote $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum eigenvalues for a matrix A, respectively.

2.2.1 Bias of the Debiased Lasso

First, we introduce the debiased Lasso. The Lasso for β_0 is defined as follows:

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|^2 + 2\lambda_0\|\beta\|_1, \tag{2.2}$$

where $\lambda_0 > 0$ denotes the tuning parameter. The debiased Lasso $\hat{b}_1$ for β_{01} is given by

$$\hat{b}_1 = \hat{\beta}_1 + \frac{1}{n}\hat{\Theta}_1^T X^T (Y - X\hat{\beta}), \tag{2.3}$$

where $\hat{\beta}_1$ denotes the first component of $\hat{\beta}$, and $\hat{\Theta}_1$ denotes an estimate of Θ_1.

Although it is not a requisite, $\hat{\Theta}_1$ is commonly constructed by the node-wise Lasso. Let X_{-1} be a submatrix of X without X_1. We define

$$\hat{\gamma}_1 := \operatorname*{argmin}_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{n}\|X_1 - X_{-1}\gamma\|^2 + 2\lambda_1\|\gamma\|_1,$$

where $\lambda_1 > 0$ is another tuning parameter. Let τ_1^2 denote the error variance in regressing X_1 on X_{-1}. Moreover, let us define $\tilde{\tau}_1^2$ and $\hat{\tau}_1^2$ as

$$\tilde{\tau}_1^2 := \frac{1}{n}\|X_1 - X_{-1}\hat{\gamma}_1\|^2, \quad \hat{\tau}_1^2 := \tilde{\tau}_1^2 + \lambda_1\|\hat{\gamma}_1\|_1.$$

The Karush-Kuhn–Tucker (KKT) conditions for the node-wise Lasso yield $\hat{\tau}_1^2 = X_1^T(X_1 - X_{-1}\hat{\gamma}_1)/n$. Then, $\hat{\Theta}_1$ is given by

$$\hat{\Theta}_1 := \frac{1}{\hat{\tau}_1^2}(1, -\hat{\gamma}_1^T)^T.$$

By a simple calculation, we get

$$\sqrt{n}(\hat{b}_1 - \beta_{01}) = \frac{1}{\sqrt{n}}\hat{\Theta}_1^T X^T \epsilon + \Delta_1, \tag{2.4}$$

where

$$\Delta_1 := \sqrt{n}(\hat{\Sigma}\hat{\Theta}_1 - e_1)^T(\beta_0 - \hat{\beta}),$$

and e_1 is the unit vector whose first element is unity.

From (2.4), the proof of asymptotic normality hinges on showing that the bias term Δ_1 is asymptotically negligible. For $\min\{s_0, s_1\} = o(\sqrt{n}/\log p)$, Javanmard and Montanari (2018) and Bellec and Zhang (2022) developed skillful techniques to evaluate the bias. Otherwise, it is common to rely on the ℓ_1-ℓ_∞ Hölder inequality:

$$|\Delta_1| \leq \sqrt{n}\|\hat{\Sigma}\hat{\Theta}_1 - e_1\|_\infty \|\hat{\beta} - \beta_0\|_1 \leq \sqrt{n}\frac{\lambda_1}{\hat{\tau}_1^2}\|\hat{\beta} - \beta_0\|_1. \tag{2.5}$$

The last inequality in (2.5) is due to the KKT conditions for the node-wise Lasso. Under certain conditions, the Lasso satisfies $\|\hat{\beta} - \beta_0\|_1 = O_p(s_0\sqrt{\log p/n})$ and the convergence rate cannot be improved. Thus, the order of the bias significantly depends on $\lambda_1/\hat{\tau}_1^2$. It has been shown that $\lambda_1/\hat{\tau}_1^2 = O_p(\sqrt{\log p/n})$ for the choice of $\lambda_1 \asymp \sqrt{\log p/n}$ in previous studies, such as van de Geer et al. (2014). They have shown that $1/\hat{\tau}_1^2 = O_p(1)$ using the fact that $\hat{\gamma}_1$ converges to its population counterpart if $\lambda_1 \asymp \sqrt{\log p/n}$ and $s_1 = o(n/\log p)$.

In this study, we present a new evaluation method of the bias by showing that $1/\hat{\tau}_1^2 = O_p(1)$ for $\lambda_1 \asymp 1/\sqrt{n}$. Different from previous studies, we do not need consistency of the node-wise Lasso. Instead, we bound $1/\hat{\tau}_1^2$ using the properties of the node-wise Lasso derived from the general position assumption and conditions on the restricted eigenvalues of $\hat{\Sigma}$. This allows us not to restrict the sparsity of Θ_1. A crucial point to bound $1/\hat{\tau}_1^2$ is bounding $1/\tilde{\tau}_1^2$, which is an upper bound of $1/\hat{\tau}_1^2$ and the variance factor of the debiased Lasso. Because Zhang and Zhang (2014) showed that $1/\tilde{\tau}_1^2$ is a nonincreasing function in λ_1, selecting too small λ_1 may make $1/\tilde{\tau}_1^2$ and the variance diverge. We show that the debiased Lasso is $\sqrt{n}$-consistent even for $\lambda_1 \asymp 1/\sqrt{n}$, although the asymptotic variance may be greater than when $\lambda_1 \asymp \sqrt{\log p/n}$.

2.2.2 Asymptotic Normality with Small Tuning Parameter

To obtain the new bound for the bias, we first introduce four generic assumptions. Then, we show that the assumptions are satisfied under rather simple conditions. Roughly speaking, our assumptions are satisfied if p is bounded by a constant multiple of n.

The first assumption is the general position assumption of Tibshirani (2013).

Assumption 1. For any $k < \min\{n, p\}$, the affine span of $\sigma_1 X_{i_1}, \ldots, \sigma_{k+1} X_{i_{k+1}}$ for arbitrary signs $\sigma_1, \ldots, \sigma_{i_{k+1}} \in \{-1, 1\}$ does not include any element of $\{\pm X_i : i \neq i_1, \ldots, i_{k+1}\}$.

Tibshirani (2013) showed that the Lasso is unique and selects at most $\min\{n, p\}$ covariates if each column vector of X is in general position. This assumption also implies that the covariates selected by the Lasso are linearly independent with probability one. A sufficient condition for the general position assumption is that the distribution of each element in X is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{np}$. See Lemma 4 in Tibshirani (2013).

Before introducing the next assumption, we define the restricted maximum and minimum eigenvalues of the sample covariance of X:

$$\phi_{\max}(s) := \max_{1 \leq \|z\|_0 \leq s} \frac{\|Xz\|^2}{n\|z\|^2}, \quad \phi_{\min}(s) := \min_{1 \leq \|z\|_0 \leq s} \frac{\|Xz\|^2}{n\|z\|^2}.$$

These two quantities also play crucial roles in evaluating $1/\tilde{\tau}_1^2$.

Assumption 2. There are a positive constant $C_{\min}$ and a positive constant K such that

$$\lim_{n \to \infty} P\left(\phi_{\min}\left(\frac{n}{K}\right) \geq C_{\min}\right) = 1.$$

We assume that n is divisible by K to simplify the notation. Under this assumption, the minimum eigenvalues of all sub-sample covariance matrices formed by less than n/K covariates are bounded below with probability tending to one. Assumption 1 only tells us that the node-wise Lasso selects less than or equal to n covariates. Assumption 2 plays the role of bounding $1/\tilde{\tau}_1^2$ above when the node-wise Lasso selects less than n/K covariates.

Assumption 3. There is a positive constant $C_{\max}$ such that

$$\lim_{n \to \infty} P(\phi_{\max}(n) \leq C_{\max}) = 1.$$

Under this assumption, the maximum eigenvalues of all sub-sample covariance matrices formed by no greater than n covariates are bounded above with probability tending to one. This assumption is to bound $1/\tilde{\tau}_1^2$ above when the node-wise Lasso selects at least n/K covariates.

Under Assumptions 1–3, we can bound $1/\tilde{\tau}_1^2$ above even when $\lambda_1 \asymp 1/\sqrt{n}$.

Lemma 1. *Suppose that Assumptions 1–3 hold and let λ_1 satisfy $\lambda_1 \geq C_1/\sqrt{n}$ for some $C_1 > 0$. Then, we have*

$$\lim_{n\to\infty} P\left(\frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} + \frac{C_{\max}K}{C_1^2}\right) = 1.$$

Lemma 1 implies that the convergence rate of the bias is improved over previous studies. van de Geer et al. (2014) and Javanmard and Montanari (2018) obtain $\lambda_1/\hat{\tau}_1^2 = O_p(\sqrt{\log p/n})$ under the assumption that $s_1 = o_p(n/\log p)$. van de Geer (2019) does not assume the sparsity of Θ_1 but obtains the same convergence rate of the bias. Meanwhile, we have $\lambda_1/\hat{\tau}_1^2 = O_P(1/\sqrt{n})$, and our derivation does not need any sparsity on Θ_1. Therefore, if the Lasso satisfies $\|\hat{\beta} - \beta_0\|_1 = O_p(s_0\sqrt{\log p/n})$, zero-mean asymptotic normality is satisfied for $s_0 = o(\sqrt{n/\log p})$. Lemma 1 also indicates that if we allow $1/C_{\min}$, $C_{\max}$, or K to diverge, the order of $\lambda_1/\hat{\tau}_1^2$ will be greater than $O_P(1/\sqrt{n})$ and s_0 should be less than $o(\sqrt{n/\log p})$ for asymptotic normality.

Before we formally state the main result, we define the restricted eigenvalue of the sample covariance introduced in Bickel et al. (2009):

$$\kappa(s, c_0)^2 := \min_{\substack{S\subset\{1,\ldots,p\}\\|S|\leq s}} \min_{\substack{z\neq 0\\\|z_{S^c}\|_1\leq c_0\|z_S\|_1}} \frac{\|Xz\|^2}{n\|z_S\|^2}.$$

We put the following assumptions on $\phi_{\max}(1)$ and the restricted eigenvalue.

Assumption 4. (i) There exists a positive constant C^* such that

$$P(\phi_{\max}(1) > C^*) = O\left(\frac{1}{p}\right).$$

(ii) For $s_0 = o(\sqrt{n/\log p})$, there exists a positive constant $\kappa_{\min}$ such that

$$\lim_{n\to\infty} P(\kappa(s_0, 3)^2 \geq \kappa_{\min}^2) = 1.$$

We use these assumptions to derive the oracle inequalities of the Lasso, which are required to evaluate the bias Δ_1. The first assumption holds, for instance, if each row vector of X is independent Gaussian. Further imposing $n \gtrsim s_0 \log(p/s_0)$, the second assumption is also satisfied. See Theorem 6 of Rudelson and Zhou (2013).

Now, we present a result of asymptotic normality for the debiased Lasso when $\lambda_1 \asymp 1/\sqrt{n}$.

Theorem 1. *Suppose Assumptions 1–4 hold and s_0 satisfies $s_0 = o(\sqrt{n/\log p})$. Then, for suitably selected $\lambda_0 \asymp \sqrt{\log p/n}$ and $\lambda_1 \asymp 1/\sqrt{n}$, we have*

$$\sqrt{n}(\hat{b}_1 - \beta_{01}) = W_1 + \Delta_1,$$

$$W_1|X \sim N(0, \sigma_\epsilon^2 \hat{\Omega}_{1,1}),$$

$$\hat{\Omega}_{1,1} := \hat{\Theta}_1^T \hat{\Sigma} \hat{\Theta}_1 = O_P(1),$$

$$\Delta_1 = o_P(1).$$

Our variance of the debiased Lasso has the same form as that of van de Geer et al. (2014). However, our variance is typically larger than theirs because our $\hat{\Theta}_1$ may not be sparse due to small λ_1. Moreover, $\hat{\Omega}_{1,1}$ may not converge in probability to the $(1,1)$ component of Σ^{-1}.

We can also establish asymptotic normality for the studentized estimator.

Corollary 1. *Suppose that the assumption of Theorem 1 holds. Let $\hat{\sigma}_\epsilon^2$ be a consistent estimator for σ_ϵ^2. Then, we have*

$$\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \xrightarrow{d} N(0,1).$$

There are some known consistent estimators for σ_ϵ^2. For example, we can set $\hat{\sigma}_\epsilon^2 = \|Y - X\hat{\beta}\|^2/n$ with $\lambda_0 \asymp \sqrt{\log p/n}$ if $s_0 = o(n/\log p)$. We can also use the scaled Lasso proposed by Sun and Zhang (2012) to obtain a consistent estimate of σ_ϵ^2.

Corollary 1 states that even if $\sqrt{n}/\log p \lesssim s_0 \ll \sqrt{n/\log p}$ and Θ_1 is not sparse, the studentized debiased Lasso is still valid to perform the t test or construct confidence intervals. We can confirm that the length of confidence intervals given by the debiased Lasso with $\lambda_1 \asymp 1/\sqrt{n}$ is $O_P(1/\sqrt{n})$ because variances are bounded.

Finally, we give a set of simple conditions under which Assumptions 1–4 hold.

Proposition 1. *Suppose that each row vector of X is independently drawn from $N(0,\Sigma)$ where Σ satisfies $K_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq K_{\max}$ for some positive constants $K_{\min}$ and $K_{\max}$. Moreover, we assume that $s_0 = o(n/\log p)$ and $p \leq C_0 n$ for some positive constant $C_0 < \infty$. Then, Assumptions 1–4 hold.*

Without the condition on p, we only have $\phi_{\max}(n) \lesssim \sqrt{\log p}$ and $\phi_{\min}(s_n) \gtrsim 1$ for $s_n = o(n/\log p)$ with high probability (see, for instance, the proof of Lemma 3.5 of Javanmard and Montanari 2018). The condition on p is used to give a tighter upper bound for $\phi_{\max}(n)$ and lower bound for $\phi_{\min}(n/K)$.

Remark 1. It has not been shown that other inference methods, such as the post-double selection and double machine learning, can have asymptotic normality under similar conditions to ours. Original proof techniques of these methods depend on convergence rates of machine learning estimators, but these rates are unavailable if no sparsity is assumed on Θ_1.

Remark 2. Ren et al. (2015), Cai and Guo (2017), and Javanmard and Montanari (2018) showed that confidence intervals with length of order $1/\sqrt{n}$ cannot be constructed when $s_0 \gtrsim \sqrt{n}/\log p$ and the population covariance matrix is unknown. However, our results do not contradict their results because they only considered the case where $p > s_0^\alpha$ for some $\alpha > 2$. If $p \leq Cn$, we have $p/s_0^\alpha \leq Cn^{1-\alpha/2}(\log p)^{\alpha/2} \to 0$ for s_0 such that $s_0 \lesssim \sqrt{n/\log p}$. Therefore, we obtain $p \leq s_0^\alpha$ for large n and their condition does not hold.

2.3 Tuning Parameter Selection Procedure

From Theorem 1 and Corollary 1, it is better to select a small tuning parameter such that $\lambda_1 \asymp 1/\sqrt{n}$ for statistical inference on β_{01} if you are unsure of how sparse β_0 and Θ_1 are. However, our theoretical results do not indicate how to select specific λ_1 in practice. Existing selection procedures may not select a sufficiently small tuning parameter needed to perform accurate statistical inference. For example, the cross-validation method is designed to minimize the prediction mean squared error, so it may be unsuitable for statistical inference. The method of Dezeure et al. (2015) selects a slightly smaller value than the cross-validation method, but it is not theoretically justified. Therefore, we propose a tuning parameter selection procedure for the node-wise Lasso that yields stable confidence intervals.

Now, we investigate the bias term of the studentized debiased Lasso. Let Λ be a set of candidate values of λ_1. For each $\lambda \in \Lambda$, we define

$$f(\lambda) := \frac{\|\hat{\Sigma}\hat{\Theta}_1(\lambda) - e_1\|_\infty}{\hat{\Omega}_{1,1}^{1/2}(\lambda)},$$

where $\hat{\Theta}_1(\lambda)$ and $\hat{\Omega}_{1,1}(\lambda)$ denote $\hat{\Theta}_1$ and $\hat{\Omega}_{1,1}$ when $\lambda_1 = \lambda$, respectively. The ℓ_1-ℓ_∞ Hölder inequality for the bias term of the studentized debiased Lasso $\tilde{\Delta}_1$ yields

$$|\tilde{\Delta}_1(\lambda)| \leq f(\lambda)\sqrt{n}\hat{\sigma}_\epsilon^{-1}\|\hat{\beta} - \beta_0\|_1.$$

We can trace the bias factor $f(\lambda)$ for each $\lambda \in \Lambda$ because $\hat{\Theta}_1$ and $\hat{\Omega}_{1,1}$ only depend on X. Note that $\sqrt{n}\hat{\sigma}_\epsilon^{-1}\|\hat{\beta} - \beta_0\|_1$ is independent of the tuning parameter of the node-wise Lasso. If $f(\lambda)$ is small, $\tilde{\Delta}_1(\lambda)$ is also small, so that the distribution of the studentized debiased Lasso is well-approximated by the standard normal distribution. However, minimizing $f(\lambda)$ is inappropriate because it is nondecreasing in λ (see Proposition 1 of Zhang and Zhang 2014). Too small tuning parameter unnecessarily increases the variance of the debiased Lasso.

To select a suitably small tuning parameter for the node-wise Lasso, we propose the following method, a simple and conservative modification of the method described in Table 2 of Zhang and Zhang (2014).

1. Let $\Lambda = \{\lambda_{\min}, \ldots, \lambda_{\max}\}$ be a candidate set of λ_1 with $\lambda_{\min} = \kappa/\sqrt{n}$ for some $\kappa > 0$ and $\lambda_{\max} = \max_{k \neq 1} |X_k^T X_1|/n$. Let

$$\eta_1^* = \frac{\tilde{\tau}_1(\lambda_{\mathrm{CV}})}{\sqrt{n}},$$

   where $\tilde{\tau}_1^2(\lambda_{\mathrm{CV}})$ denotes the sample mean of squared residuals from the cross-validated node-wise Lasso.

2. Select λ_1 with the following rule:

   (a) If $f(\lambda) > \eta_1^*$ for all $\lambda \in \Lambda$, then set $\lambda_1 = \lambda_{\min}$.

   (b) Otherwise,

$$\begin{aligned}
\lambda^* &= \operatorname{argmin}\{\hat{\Omega}_{1,1}^{1/2}(\lambda) : f(\lambda) \leq \eta_1^*\}, \\
\hat{\Omega}_{1,1}^{*1/2} &= 1.1 \times \hat{\Omega}_{1,1}^{1/2}(\lambda^*), \\
\lambda_1 &= \operatorname{argmin}\{f(\lambda) : \hat{\Omega}_{1,1}^{1/2}(\lambda) \leq \hat{\Omega}_{1,1}^{*1/2}\}.
\end{aligned}$$

We call this proposed method STPS. The idea of STPS is simple. It minimizes the bias with a constraint on the upper bound of the variance. When $f(\lambda) \leq \eta_1^*$ for some $\lambda \in \Lambda$,

λ_1 is also equal to the smallest value in $\{\lambda : \hat{\Omega}_{1,1}^{1/2}(\lambda) \le \hat{\Omega}_{1,1}^{*1/2}\}$. Simulation studies show that $\kappa = 0.001$ is an appropriate value to obtain accurate coverage.

The main difference between our method and the method of Zhang and Zhang (2014) is that our upper bound of $f(\lambda)$ is much smaller than theirs, which is given by $\eta_1^* = \sqrt{2\log p/n}$. Therefore, our method can select a smaller tuning parameter than theirs, and we can show that $f(\lambda_1) = O_P(1/\sqrt{n})$. Moreover, STPS does not increase the standard error overly because we can show that $\hat{\Omega}_{1,1}^{*1/2}$ is bounded. Meanwhile, if we select λ_1 using their method, the standard error may be smaller than ours; however, we obtain $f(\lambda_1) = O_P(\sqrt{\log p/n})$, meaning that the bias may be large and coverage of confidence intervals may be small if β_0 and Θ_1 are not sufficiently sparse.

Finally, we present an asymptotic justification for the proposed procedure.

Theorem 2. *Suppose that the assumption of Theorem 1 holds, and λ_1 is selected by STPS. Then, we have*

$$\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \xrightarrow{d} N(0,1).$$

Remark 3. For our selection method to work well, Λ must contain small candidate values. However, well-known software packages typically do not generate sufficiently small values as candidates. Thus, it is necessary to manually include small values to Λ. An example of the choice of Λ is illustrated in the next section.

2.4 Simulation Studies

In this section, we compare finite sample performances of the debiased Lasso tuned by our procedure and those of other statistical inference methods. We use the R packages `glmnet` and `hdm`. In all examples, we set $p = 2n$ and repeat experiments 1,000 times.

2.4.1 Simulation Design

We employ the settings of Wang et al. (2020). For a given value of c_y, we set β_0 as one of the following $p+1$-dimensional vectors.

$$\text{sparse} : (1.5, \underbrace{c_y, \ldots, c_y}_{5}, 0, \ldots, 0)^T,$$

$$\text{moderately sparse} : (1.5, \underbrace{5c_y, \ldots, 5c_y}_{10}, \underbrace{c_y, \ldots, c_y}_{10}, 0, \ldots, 0)^T,$$

$$\text{dense} : (1.5, c_y, c_y/\sqrt{2}, \ldots, c_y/\sqrt{p})^T.$$

In addition, for a given value of c_x, we set γ_0 as one of the following p-dimensional vectors.

$$\text{sparse} : (\underbrace{0, \ldots, 0}_{4}, \underbrace{c_x, \ldots, c_x}_{4}, 0, \ldots, 0)^T,$$
$$\text{dense} : (c_x, c_x/\sqrt{2}, \ldots, c_x/\sqrt{p})^T.$$

The covariance matrix Σ is given by one of the following matrices:

$$\text{Independent} : \Sigma = I, \quad \text{Toeplitz} : \Sigma_{jk} = \rho^{|j-k|},$$
$$\text{Equal correlation} : \Sigma_{jk} = \rho^{1(j \neq k)}, \quad \rho \in \{0.3, 0.9\}.$$

We consider the following two settings.

- Setting 1. The data generating process is given by

$$Y_i = 1 + X_i^T \beta_0 + \epsilon_i \quad \text{for } i = 1, \ldots, n,$$

  where $X_i \in \mathbb{R}^{p+1} \sim N(0, \Sigma)$, $\epsilon_i \sim N(0, 1)$, and X_i and ϵ_i are independent. We select c_y so that R^2 is 0.8.

- Setting 2. The data generating process is given by

$$Y_i = 1 + X_i^T \beta_0 + \epsilon_i,$$
$$X_{i,1} = 0.5 + X_{i,-1}^T \gamma_0 + \eta_i,$$

  where $X_{i,-1} \in \mathbb{R}^p \sim N(0, \Sigma)$, $(\epsilon_i, \eta_i) \sim N(0, I_2)$, and $X_i = (X_{i,1}, X_{i,-1}^T)^T$ and (ϵ_i, η_i) are independent. We select c_y so that R^2 of the first equation is 0.8 and select c_x so that R^2 of the second equation is one of {0.3, 0.5, 0.8, 0.9}.

The true value of the parameter of interest, the first element of β_0, is 1.5. In Setting

1, some calculations yield

$$\tau_1^2 = \begin{cases} 1 & \text{if} \quad \Sigma_{jk} = 1(j = k), \\ 0.701 & \text{if} \quad \Sigma_{jk} = 0.3^{1(j \neq k)}, \\ 0.19 & \text{if} \quad \Sigma_{jk} = 0.9^{|j-k|}, \\ 0.1 & \text{if} \quad \Sigma_{jk} = 0.9^{1(j \neq k)}. \end{cases}$$

In Setting 2, we have $\tau_1^2 = 1$. When the covariance matrix has the structure of the Toeplitz or equal correlation, Θ_1 is dense for Setting 1. In Setting 2, γ_0 controls the sparsity of Θ_1 because we have $\Theta_1 = (1, -\gamma_0^T)^T$.

The following procedures are included in the comparisons.

- "Oracle" refers to OLS derived from true models.
- "STPS" refers to the debiased Lasso with λ_1 selected by STPS.
- "CV" refers to the debiased Lasso with λ_1 selected by 10-fold cross-validation.
- "Univ" refers to the debiased Lasso with $\lambda_1 = \tau_1\sqrt{2\log p/n}$, the universal tuning parameter.
- "DBMM" refers to the debiased Lasso with λ_1 selected by the method of Dezeure et al. (2015).
- "ZZ" represents the debiased Lasso with λ_1 selected by the method of Zhang and Zhang (2014).
- "Double" refers to the post-double selection of Belloni et al. (2014).
- "PODS" refers to the projection onto the double selection of Wang et al. (2020).
- "DML" refers to the double machine learning of Chernozhukov et al. (2018) with 5-fold sample splitting.

We compared the biases, standard deviations, coverages, and lengths of 95% confidence intervals for all methods.

To implement the debiased Lasso, we need to determine two tuning parameters. For all debiased Lasso-based methods, including STPS, the tuning parameter of the Lasso is determined by the 10-fold cross-validation, which is implemented using the `lambda.min` option of `glmnet`. The tuning parameter of the node-wise Lasso is selected from a common set Λ. As we note in the previous section, the default set generated by `glmnet` does not contain sufficiently small values. Thus, we augment the default set by adding some small values. Specifically, we set $\Lambda = \Lambda_1 \cup \Lambda_2$ where $\Lambda_1 = \{0.001/\sqrt{n}, 0.002/\sqrt{n}, \ldots, 0.02/\sqrt{n}\}$ consists of 20 elements, and Λ_2 is the set of 100 elements generated by `glmnet`. Note that we do not multiply λ_0 and λ_1 by 2 in our simulation studies. We estimate the error variance $\sigma_\epsilon^2 = 1$ by $\hat{\sigma}_\epsilon^2 = \|Y - X\hat{\beta}(\lambda_{1se})\|^2/n$, where $\hat{\beta}(\lambda_{1se})$ is the Lasso whose tuning parameter is selected by the one standard error rule with `lambda.1se` option.

We briefly explain other estimation methods. Double is implemented by the function `rlassoEffects` of the R package `hdm`. In PODS, we use the Lasso implemented by the function `rlasso` of the R package `hdm` as the model selection procedure. To implement DML, we follow Definition 3.2 of Chernozhukov et al. (2018) and use the variance estimator defined in Theorem 3.2 to construct confidence intervals. The orthogonal score function ψ is defined as follows:

$$\psi(W_i, \eta_0, \beta_{01}) = (Y_i - X_{i,1}\beta_{01} - g_0(X_{i,-1}))(X_{i,1} - m_0(X_{i,-1})),$$

where $\eta_0 = (g_0, m_0)$ is a nuisance parameter and $W_i = (X_i, Y_i)$. We estimate g_0 and m_0 by the post-Lasso, which is implemented by the function `rlasso`. Tuning parameters for the post-Lasso are selected using the method of Belloni et al. (2012), which is also implemented by `rlasso`.

2.4.2 Simulation Results

Table 2.1 summarizes the results for Setting 1 with $n = 100$ and $p = 200$. In all cases, the coverage rates of STPS are close to the nominal level. On the other hand, the debiased Lasso-based confidence intervals have smaller coverage rates if λ_1 is selected by other methods. This is because they select relatively larger tuning parameters so that the debiased Lasso tends to have a large bias and small variance. In particular, the universal penalty, which is often used for theoretical analysis of the Lasso, may not be suitable for the node-wise Lasso. CV also yields low coverage, although it is the most well-known method for tuning parameter selection. Double and PODS are also unstable, and their coverages are generally low. If the correlation among covariates is not too high, DML performs well and its standard deviations are smaller than those of STPS. However, when the correlation among covariates is the highest, the coverage of DML is much higher than the nominal level.

Table 2.2 shows the results for Setting 1 when $n = 500$ and $p = 1000$. STPS gives confidence intervals with accurate coverage. Even in this case, Univ still has large biases, and its coverage is small. Although the results of CV, DBMM, and ZZ are improved compared with the small sample size case, their coverage is still small, particularly when β_0 is dense and Σ has the Toeplitz structure. DML performs well, but its coverage is still large when the correlation is the highest. The performances of Double and PODS are unstable, even in the large sample case.

Next, we compare the results of all procedures for Setting 2. Table 2.3 shows the simulation results for this setting when $n = 100$ and $p = 200$. Only the coverage rates of STPS are close to the nominal coverage in all cases. Other debiased Lasso-based methods do not cover β_{01} well in this setting. In particular, when β_0 and γ_0 are dense, they have large biases. These results are consistent with our theoretical analysis. Because the sparsity assumptions of van de Geer et al. (2014) and Javanmard and Montanari (2018) are violated when both β_0 and γ_0 are dense, existing selection methods cannot remove the bias. Although DML performs well when either β_0 or γ_0 is not dense, its performance is significantly aggravated when both of them are dense. The theory of DML allows that one nuisance parameter is dense to make the bias small if the other nuisance parameter is quite sparse. However, because both nuisance parameters are dense, a large bias occurs for DML.

Table 2.4 shows simulation results for Setting 2 when $n = 500$ and $p = 1,000$. When β_0 and γ_0 are dense, the biases for all methods other than STPS are so large that confidence intervals yield much small coverage. This is because the other methods miss more relevant covariates than the case of $n = 100$.

We also performed simulation for $p > 2n$. The results are similar to the case of $p = 2n$ and are not reported here. We confirmed that STPS outperforms other methods in terms of coverage for $p \leq 10n$. However, when $p = 10n$, the bias of the debiased Lasso cannot be sufficiently removed, even by selecting the smallest value in Λ. Therefore, the coverage of STPS also tends to be smaller than the nominal level.

In summary, provided p is not significantly greater than n, STPS gives accurate confidence intervals regardless of whether β_0 and Θ_1 are sparse or dense. Meanwhile, other methods perform poorly, particularly when β_0 and Θ_1 are dense. Our simulation results indicate that the debiased Lasso tuned by our method would enable us to perform accurate statistical inference in general settings when p is reasonably greater than n.

Table 2.1: Simulation results for Setting 1 with $n = 100$ and $p = 200$.

	Oracle	STPS	Univ	CV	DBMM	ZZ	Double	PODS	DML
	β_0 is sparse, $\Sigma = I$								
Bias	0.001	-0.039	-0.052	-0.051	-0.045	-0.043	-0.162	-0.006	-0.003
SD	0.103	0.165	0.121	0.122	0.13	0.139	0.161	0.163	0.127
Cover	0.932	0.94	0.875	0.886	0.909	0.919	0.757	0.831	0.938
Length	0.392	0.671	0.427	0.435	0.485	0.532	0.545	0.442	0.462
	β_0 is sparse, $\Sigma_{j,k} = 0.3^{1(j \neq k)}$								
Bias	-0.006	-0.026	-0.046	-0.042	-0.038	-0.036	-0.041	0.019	0.006
SD	0.111	0.192	0.133	0.139	0.145	0.145	0.159	0.213	0.153
Cover	0.949	0.944	0.867	0.899	0.921	0.919	0.904	0.775	0.95
Length	0.434	0.757	0.424	0.482	0.537	0.542	0.569	0.511	0.597
	β_0 is sparse, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$								
Bias	0.002	-0.067	-0.311	-0.114	-0.087	-0.218	0.331	0.014	0.14
SD	0.291	0.489	0.325	0.347	0.357	0.332	0.367	0.566	0.387
Cover	0.944	0.935	0.509	0.886	0.917	0.739	0.848	0.732	0.984
Length	1.109	1.841	0.652	1.201	1.309	0.894	1.425	1.28	2.037
	β_0 is sparse, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$								
Bias	0.001	-0.011	0.014	0.009	0.004	0.005	-0.064	-0.004	-0.001
SD	0.24	0.384	0.224	0.234	0.241	0.232	0.267	0.284	0.25
Cover	0.934	0.945	0.826	0.886	0.913	0.901	0.908	0.905	0.932
Length	0.9	1.525	0.601	0.734	0.814	0.749	0.953	0.932	0.931
	β_0 is moderately sparse, $\Sigma = I$								
Bias	0	-0.066	-0.087	-0.086	-0.077	-0.073	-0.265	-0.009	-0.002
SD	0.111	0.179	0.147	0.148	0.152	0.158	0.186	0.183	0.173
Cover	0.92	0.909	0.79	0.797	0.833	0.86	0.622	0.834	0.926
Length	0.392	0.7	0.446	0.454	0.506	0.555	0.653	0.496	0.632
	β_0 is moderately sparse, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$								
Bias	0.002	-0.066	-0.31	-0.113	-0.085	-0.217	0.333	0.014	0.138
SD	0.348	0.488	0.325	0.346	0.356	0.331	0.369	0.565	0.386
Cover	0.908	0.938	0.51	0.885	0.92	0.743	0.845	0.732	0.986
Length	1.203	1.837	0.651	1.199	1.306	0.893	1.424	1.277	2.028
	β_0 is moderately sparse, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$								
Bias	0.003	-0.015	-0.008	-0.009	-0.012	-0.013	-0.077	-0.003	0.002
SD	0.265	0.383	0.214	0.227	0.236	0.226	0.267	0.288	0.253
Cover	0.913	0.949	0.835	0.893	0.911	0.898	0.899	0.904	0.934
Length	0.899	1.515	0.597	0.729	0.809	0.744	0.951	0.933	0.951
	β_0 is dense, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$								
Bias	-	-0.065	-0.31	-0.112	-0.085	-0.216	0.334	0.014	0.138
SD	-	0.488	0.325	0.346	0.357	0.331	0.369	0.567	0.387
Cover	-	0.939	0.516	0.891	0.92	0.74	0.85	0.728	0.985
Length	-	1.836	0.65	1.198	1.306	0.892	1.423	1.278	2.028
	β_0 is dense, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$								
Bias	-	-0.021	-0.037	-0.032	-0.03	-0.034	-0.152	-0.013	0.002
SD	-	0.381	0.225	0.235	0.242	0.234	0.264	0.297	0.289
Cover	-	0.939	0.785	0.842	0.876	0.862	0.865	0.873	0.944
Length	-	1.459	0.575	0.702	0.779	0.716	0.955	0.931	1.09

Table 2.2: Simulation results for Setting 1 with $n = 500$ and $p = 1000$.

	Oracle	STPS	Univ	CV	DBMM	ZZ	Double	PODS	DML
					β_0 is sparse, $\Sigma = I$				
Bias	-0.002	-0.005	-0.008	-0.008	-0.006	-0.006	-0.1	-0.001	-0.002
SD	0.046	0.076	0.048	0.048	0.052	0.057	0.062	0.061	0.046
Cover	0.94	0.964	0.945	0.943	0.957	0.953	0.536	0.862	0.942
Length	0.176	0.314	0.185	0.185	0.207	0.23	0.211	0.185	0.176
					β_0 is sparse, $\Sigma_{j,k} = 0.3^{1(j \neq k)}$				
Bias	0.001	-0.004	-0.01	-0.008	-0.006	-0.006	0.03	0.02	0.021
SD	0.049	0.096	0.054	0.055	0.059	0.063	0.062	0.078	0.057
Cover	0.959	0.954	0.923	0.945	0.95	0.953	0.902	0.81	0.961
Length	0.195	0.379	0.196	0.211	0.235	0.248	0.229	0.214	0.242
					β_0 is sparse, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$				
Bias	0.002	-0.008	-0.102	-0.023	-0.017	-0.044	0.242	0.027	0.037
SD	0.127	0.224	0.141	0.144	0.144	0.143	0.152	0.22	0.147
Cover	0.95	0.953	0.721	0.938	0.947	0.908	0.636	0.786	0.975
Length	0.497	0.893	0.384	0.544	0.556	0.5	0.589	0.555	0.667
					β_0 is sparse, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$				
Bias	0	0.007	0.013	0.009	0.004	0.003	-0.06	-0.001	0
SD	0.108	0.18	0.106	0.106	0.108	0.109	0.115	0.118	0.109
Cover	0.941	0.958	0.85	0.89	0.92	0.925	0.892	0.924	0.941
Length	0.402	0.707	0.314	0.341	0.38	0.392	0.419	0.409	0.404
					β_0 is moderately sparse, $\Sigma = I$				
Bias	-0.002	-0.008	-0.014	-0.014	-0.011	-0.01	-0.106	0	-0.002
SD	0.047	0.077	0.05	0.05	0.054	0.058	0.063	0.063	0.048
Cover	0.938	0.963	0.928	0.926	0.945	0.956	0.532	0.866	0.952
Length	0.176	0.323	0.19	0.19	0.212	0.236	0.22	0.19	0.185
					β_0 is moderately sparse, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$				
Bias	0.003	-0.008	-0.101	-0.022	-0.017	-0.043	0.242	0.026	0.037
SD	0.138	0.224	0.14	0.144	0.144	0.143	0.151	0.22	0.147
Cover	0.947	0.953	0.723	0.938	0.945	0.907	0.632	0.777	0.978
Length	0.541	0.891	0.383	0.543	0.555	0.499	0.587	0.554	0.665
					β_0 is moderately sparse, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$				
Bias	0	0.006	0.007	0.004	0	0	-0.064	-0.002	0
SD	0.109	0.18	0.102	0.103	0.107	0.108	0.115	0.119	0.109
Cover	0.94	0.959	0.871	0.896	0.924	0.928	0.892	0.919	0.943
Length	0.402	0.706	0.313	0.341	0.38	0.392	0.419	0.41	0.407
					β_0 is dense, $\Sigma_{j,k} = 0.9^{1(j \neq k)}$				
Bias	-	-0.008	-0.101	-0.022	-0.016	-0.043	0.241	0.024	0.038
SD	-	0.224	0.14	0.143	0.143	0.142	0.151	0.221	0.146
Cover	-	0.95	0.72	0.937	0.947	0.912	0.639	0.785	0.976
Length	-	0.89	0.382	0.542	0.554	0.499	0.587	0.554	0.664
					β_0 is dense, $\Sigma_{j,k} = 0.9^{\lvert j-k \rvert}$				
Bias	-	0.004	-0.009	-0.01	-0.01	-0.009	-0.16	-0.003	-0.002
SD	-	0.179	0.108	0.108	0.11	0.111	0.115	0.124	0.128
Cover	-	0.943	0.822	0.863	0.899	0.904	0.661	0.903	0.944
Length	-	0.67	0.297	0.324	0.361	0.372	0.421	0.409	0.481

Table 2.3: Simulation results for Setting 2 with $n = 100$ and $p = 200$.

	Oracle	STPS	Univ	CV	DBMM	ZZ	Double	PODS	DML
				β_0 and γ_0 are sparse, $R^2 = 0.5$					
Bias	-0.002	-0.014	-0.003	-0.007	-0.009	-0.009	-0.043	-0.003	0
SD	0.09	0.191	0.101	0.105	0.109	0.109	0.117	0.122	0.108
Cover	0.931	0.962	0.89	0.913	0.936	0.933	0.907	0.908	0.938
Length	0.341	0.754	0.33	0.368	0.41	0.414	0.439	0.416	0.415
				β_0 and γ_0 are sparse, $R^2 = 0.9$					
Bias	-0.003	-0.024	-0.111	-0.084	-0.071	-0.085	-0.007	0.007	0.029
SD	0.053	0.175	0.08	0.087	0.091	0.084	0.125	0.122	0.128
Cover	0.935	0.964	0.534	0.734	0.823	0.75	0.937	0.903	0.98
Length	0.198	0.709	0.22	0.284	0.315	0.279	0.484	0.415	0.664
				β_0 is moderately sparse and γ_0 is dense, $R^2 = 0.3$					
Bias	0.003	-0.013	0.019	0.006	-0.002	-0.003	-0.061	-0.014	0.008
SD	0.101	0.192	0.102	0.109	0.115	0.112	0.125	0.134	0.11
Cover	0.921	0.948	0.893	0.928	0.937	0.937	0.901	0.889	0.922
Length	0.359	0.744	0.342	0.39	0.434	0.427	0.466	0.437	0.398
				β_0 is moderately sparse and γ_0 is dense, $R^2 = 0.8$					
Bias	0.003	-0.005	0.006	0.003	0.002	0.004	-0.007	-0.007	0.007
SD	0.067	0.187	0.074	0.099	0.106	0.08	0.149	0.172	0.092
Cover	0.921	0.962	0.867	0.931	0.94	0.898	0.923	0.827	0.937
Length	0.234	0.748	0.214	0.367	0.406	0.264	0.551	0.473	0.329
				β_0 and γ_0 are dense, $R^2 = 0.8$					
Bias	-	0.002	0.113	0.05	0.038	0.087	-0.032	-0.014	0.122
SD	-	0.187	0.097	0.106	0.111	0.096	0.149	0.173	0.092
Cover	-	0.967	0.41	0.877	0.917	0.66	0.911	0.819	0.657
Length	-	0.773	0.221	0.379	0.42	0.273	0.547	0.472	0.338

Table 2.4: Simulation results for Setting 2 with $n = 500$ and $p = 1000$.

	Oracle	STPS	Univ	CV	DBMM	ZZ	Double	PODS	DML
	β_0 and γ_0 are sparse, $R^2 = 0.5$								
Bias	-0.001	0.001	0.002	0	-0.001	-0.001	-0.026	0	-0.002
SD	0.04	0.085	0.045	0.045	0.047	0.049	0.051	0.052	0.047
Cover	0.942	0.961	0.918	0.93	0.95	0.956	0.88	0.911	0.94
Length	0.152	0.34	0.156	0.164	0.183	0.196	0.187	0.179	0.177
	β_0 and γ_0 are sparse, $R^2 = 0.9$								
Bias	0	-0.002	-0.046	-0.035	-0.025	-0.025	-0.024	0	0
SD	0.023	0.083	0.037	0.039	0.042	0.041	0.051	0.053	0.047
Cover	0.952	0.96	0.682	0.798	0.871	0.871	0.897	0.914	0.956
Length	0.089	0.341	0.123	0.138	0.154	0.154	0.191	0.179	0.189
	β_0 is moderately sparse and γ_0 is dense, $R^2 = 0.3$								
Bias	-0.002	0.001	0.012	0.004	0.001	0.001	-0.029	-0.005	-0.001
SD	0.042	0.089	0.044	0.046	0.048	0.049	0.053	0.056	0.044
Cover	0.933	0.959	0.91	0.936	0.952	0.953	0.891	0.91	0.942
Length	0.156	0.352	0.154	0.169	0.189	0.191	0.199	0.189	0.165
	β_0 is moderately sparse and γ_0 is dense, $R^2 = 0.8$								
Bias	-0.001	0.002	0.006	0.002	0.002	0.004	-0.011	-0.003	0
SD	0.025	0.093	0.029	0.04	0.043	0.032	0.061	0.067	0.037
Cover	0.944	0.95	0.898	0.95	0.953	0.924	0.942	0.877	0.937
Length	0.095	0.37	0.097	0.156	0.173	0.116	0.231	0.209	0.134
	β_0 and γ_0 are dense, $R^2 = 0.8$								
Bias	-	0.004	0.16	0.068	0.049	0.123	-0.035	-0.012	0.144
SD	-	0.093	0.041	0.043	0.045	0.04	0.062	0.067	0.038
Cover	-	0.954	0.015	0.597	0.813	0.068	0.893	0.868	0.029
Length	-	0.382	0.1	0.161	0.179	0.12	0.229	0.209	0.14

2.5 Real Data Example

In this section, we present an example of applying our method to real economic data. We revisit the study of Harrison and Rubinfeld (1978), who investigated methodological problems in using housing data to estimate the demand for clean air. They used data for census tracts in the Boston Standard Metropolitan Statistical Area. We examine the effect of nitrogen oxide concentration on housing prices by adding new control variables to the original regression model.

Harrison and Rubinfeld (1978) considered the following linear regression model:

$$\begin{aligned}\log \mathrm{MV}_i = \alpha &+ \beta_1 \mathrm{NOX}_i^2 + \beta_2 \mathrm{RM}_i^2 + \beta_3 \log \mathrm{DIS}_i + \beta_4 \mathrm{AGE}_i + \beta_5 \log \mathrm{RAD}_i + \beta_6 \mathrm{TAX}_i \\ &+ \beta_7 \mathrm{PTRATIO}_i + \beta_8 (B_i - 0.63)^2 + \beta_9 \log \mathrm{LSTAT}_i + \beta_{10} \mathrm{CRIM}_i \\ &+ \beta_{11} \mathrm{ZN}_i + \beta_{12} \mathrm{INDUS}_i + \beta_{13} \mathrm{CHAS}_i + \epsilon_i,\end{aligned}$$

where MV denotes the median value of houses and NOX denotes the nitrogen oxide concentration in each census tract[1]. See Gilley and Pace (1996) for explanations of other covariates. Harrison and Rubinfeld (1978) showed that the OLS estimate of the coefficient of NOX^2 has a negative sign, being highly significant. Because the original data of Harrison and Rubinfeld (1978) contained some incorrectly coded observations, Gilley and Pace (1996) performed sensitivity analysis based on corrected data and showed that the conclusions of the empirical study do not change.

We performed sensitivity analysis by adding new control variables. We considered the following linear regression model:

$$\log \mathrm{MV}_i = \alpha + \beta_1 \mathrm{NOX}_i^2 + X_{i,-1}^T \beta_{-1} + \epsilon_i, \tag{2.6}$$

where $X_{i,-1}$ denotes the vector of 78 control variables, including all covariates except for NOX^2 in the original model and their first order interaction terms. We used the data of Gilley and Pace (1996), which is available from the R package `mlbench`. The sample size was 506. We performed statistical inference on β_1 using methods listed in our simulation studies. Because we could obtain the OLS estimate of β_1, we also compared the results of methods with OLS estimates in the original model and the model (2.6). To obtain the estimator of DML, we employed the median method using 100 sample splits (see

[1] In Harrison and Rubinfeld (1978), the exponent 2 of NOX is an estimated value rather than a predetermined value.

Table 2.5: Estimates of β_1 and their standard errors for the model (2.6) with $n = 506$ and $\dim(X_{i,-1}) = 78$. OLS refers to the OLS results for β_1 in the original model, which are given in Gilley and Pace (1996). HOLS refers to the OLS results for β_1 in the model (2.6).

	OLS	HOLS	STPS	CV	Univ
Est	-0.63724	-0.65051	-0.65184	-0.63743	-0.56759
SE	0.11003	0.11512	0.14784	0.13480	0.08642

	DBMM	ZZ	Double	PODS	DML
Est	-0.65184	-0.65184	-0.70252	-0.66724	-0.58113
SE	0.14784	0.14784	0.14019	0.10230	0.15045

Definition 3.3 of Chernozhukov et al. 2018). Because we do not know the true error variance in regressing NOX_i^2 on $X_{i,-1}$ for Univ, we estimate it by the sample mean of squared residuals from the node-wise Lasso tuned by the one standard error rule.

Table 2.5 shows the results of applying methods to the model (2.6). The results of all methods are consistent with those of previous studies: all methods rejected the null hypothesis $\beta_1 = 0$ at a 5% significance level. Because the high-dimensional OLS estimate does not differ from the original OLS estimate, it is plausible to think that control variables in the original model are sufficient. The estimates of the debiased Lasso tuned by cross-validation, methods of Zhang and Zhang (2014), Dezeure et al. (2015), and STPS, and the estimate of PODS are close to the OLS estimates. Meanwhile, other estimates are relatively different from the two OLS estimates. The estimates of Univ and DML are relatively large, whereas the estimate of Double is small. Because the simulation studies show that STPS always has small biases when n is large, estimates that significantly differ from that of STPS may have large biases in this real data example.

2.6 Conclusion

In this study, we analyzed theoretical and numerical properties of the debiased Lasso when the tuning parameter of the node-wise Lasso is much smaller than in previous studies. Moreover, we proposed a data-driven tuning parameter selection procedure, called STPS. Our analysis shows that selecting a moderately small tuning parameter can mitigate the bias of the debiased Lasso without making variances diverge if the number of covariates is not too large. This implies that asymptotic normality for the debiased

Lasso holds provided that the number of nonzero coefficients in linear regression models is $o(\sqrt{n/\log p})$. According to simulation studies and a real data example, STPS yields accurate confidence intervals. Therefore, STPS would be a useful tool for high-dimensional statistical inferences.

2.7 Appendix

2.7.1 Proofs

Proof of Lemma 1. For the node-wise Lasso estimate $\hat{\gamma}_1 = (\hat{\gamma}_{1,2}, \ldots, \hat{\gamma}_{1,p})^T \in \mathbb{R}^{p-1}$, let $\hat{S}_1 = \{j : \hat{\gamma}_{1,j} \neq 0\}$ and $\hat{s}_1 = |\hat{S}_1|$. Moreover, let $X_{\hat{S}_1}$ denote the submatrix of X whose columns correspond to the elements of $\hat{S}_1$. Because each column vector of X is in general position, all column vectors of $X_{\hat{S}_1}$ are linearly independent, and the node-wise Lasso selects at most n variables when $p > n$. Because $X_{\hat{S}_1}^T X_{\hat{S}_1}$ is invertible with probability one, for $\lambda_1 \geq C_1/\sqrt{n}$, we can write

$$\begin{aligned}
\tilde{\tau}_1^2 &= \frac{1}{n} X_1^T Q_{\hat{S}_1} X_1 + \lambda_1^2 \kappa_{\hat{S}_1}^T \left(\frac{1}{n} X_{\hat{S}_1}^T X_{\hat{S}_1} \right)^{-1} \kappa_{\hat{S}_1} \\
&\geq \lambda_{\min} \left(\frac{1}{n} X_{\{1\}\cup\hat{S}_1}^T X_{\{1\}\cup\hat{S}_1} \right) + \frac{\hat{s}_1}{n} \frac{C_1^2}{\phi_{\max}(\hat{s}_1)} \\
&\geq \phi_{\min}(1 + \hat{s}_1) + \frac{\hat{s}_1}{n} \frac{C_1^2}{\phi_{\max}(\hat{s}_1)},
\end{aligned}$$

where $Q_{\hat{S}_1} = I - X_{\hat{S}_1}(X_{\hat{S}_1}^T X_{\hat{S}_1})^{-1} X_{\hat{S}_1}^T$. The second inequality holds because of the definition of restricted maximum eigenvalues and the fact that $(X_1^T Q_{\hat{S}_1} X_1/n)^{-1}$ is the first diagonal element of $(X_{\{1\}\cup\hat{S}_1}^T X_{\{1\}\cup\hat{S}_1}/n)^{-1}$. See A-74 of Greene (2012). The last inequality holds by the definition of restricted minimum eigenvalues. Therefore, we obtain

$$\frac{1}{\tilde{\tau}_1^2} \leq \min \left\{ \frac{1}{\phi_{\min}(1 + \hat{s}_1)}, \frac{n}{\hat{s}_1} \frac{\phi_{\max}(\hat{s}_1)}{C_1^2} \right\}.$$

Let $A_1 = \{\phi_{\min}(n/K) \geq C_{\min}\}$, $A_2 = \{\phi_{\max}(n) \leq C_{\max}\}$, G=\{each column vector of X is in general position\}, and

$$T = \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \min \left\{ \frac{1}{\phi_{\min}(1 + \hat{s}_1)}, \frac{n}{\hat{s}_1} \frac{\phi_{\max}(\hat{s}_1)}{C_1^2} \right\} \right\}.$$

Notice that $G \subset T \cap \{\hat{s}_1 \leq n\}$. Let $B_1 = \{\hat{s}_1 < n/K\}$ and $B_2 = \{n \geq \hat{s}_1 \geq n/K\}$. Then, we have $\{\hat{s}_1 \leq n\} = B_1 \cup B_2$ and $B_1 \cap B_2 = \emptyset$. Thus, we have

$$A_1 \cap A_2 \cap G \subset A_1 \cap A_2 \cap T \cap (B_1 \cup B_2) \subset (A_1 \cap B_1 \cap T) \cup (A_2 \cap B_2 \cap T). \tag{2.7}$$

First, we focus on the event $A_1 \cap B_1 \cap T$. Recall that n is divisible by K so that $1 + \hat{s}_1 \leq n/K$. Recall also that $\phi_{\min}(s)$ is nonincreasing in s. Therefore, on the event $A_1 \cap B_1 \cap T$, we have $\phi_{\min}(1+\hat{s}_1) \geq \phi_{\min}(n/K) \geq C_{\min}$ and $1/\tilde{\tau}_1^2 \leq 1/\phi_{\min}(1+\hat{s}_1)$. Hence, we obtain

$$A_1 \cap B_1 \cap T \subset \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} \right\}. \tag{2.8}$$

Next, we focus on the event $A_2 \cap B_2 \cap T$. Recall that $\phi_{\max}(s)$ is a nondecreasing function in s. Therefore, on the event $A_2 \cap B_2 \cap T$, we have $n \geq \hat{s}_1 \geq n/K$, $\phi_{\max}(\hat{s}_1) \leq \phi_{\max}(n) \leq C_{\max}$, and $1/\tilde{\tau}_1^2 \leq n\phi_{\max}(\hat{s}_1)/(\hat{s}_1 C_1^2)$. Hence, we obtain

$$A_2 \cap B_2 \cap T \subset \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \frac{K C_{\max}}{C_1^2} \right\}. \tag{2.9}$$

By combining (2.7), (2.8), and (2.9), we have

$$A_1 \cap A_2 \cap G \subset \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} \right\} \cup \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \frac{K C_{\max}}{C_1^2} \right\} \subset \left\{ \frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} + \frac{K C_{\max}}{C_1^2} \right\}.$$

By the union bound and Assumption 1, we obtain

$$\begin{aligned} P\left(\frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} + \frac{C_{\max} K}{C_1^2} \right) &\geq P(A_1 \cap A_2 \cap G) = 1 - P(A_1^c \cup A_2^c \cup G^c) \\ &\geq 1 - P(A_1^c) - P(A_2^c). \end{aligned}$$

By Assumptions 2 and 3, we have

$$\liminf_{n \to \infty} P\left(\frac{1}{\tilde{\tau}_1^2} \leq \frac{1}{C_{\min}} + \frac{C_{\max} K}{C_1^2} \right) = 1.$$

This completes the proof of this lemma. □

Proof of Proposition 1. Assumptions 1 and 4 (ii) are satisfied from Lemma 4 of Tibshirani (2013) and Theorem 6 of Rudelson and Zhou (2013), respectively. We now show that the condition of Proposition 1 implies Assumptions 2, 3, and 4 (i). As in the appendix of Javanmard and Montanari (2018), who follow Remark 5.40 of Vershynin (2012), we have

$$P\left(\phi_{\max}(k) \geq K_{\max} + C\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\right) \leq 2\binom{p}{k}\exp(-ct^2),$$

for any $t \geq 0$ and $1 \leq k \leq n$, where positive constants C and c depend on $K_{\max}$ but not on n. Notice that

$$\binom{p}{k} = \frac{p!}{k!(p-k)!} \leq \frac{p^k}{k!} \leq \left(\frac{p}{k}\right)^k \exp(k)$$

because $\exp(k) = \sum_{i=0}^{\infty} k^i/i! \geq k^k/k!$ for any positive integer k. Thus, for $p \leq C_0 n$, we have

$$P\left(\phi_{\max}(n) \geq K_{\max} + C + \frac{t}{\sqrt{n}}\right) \leq 2\exp(n(\log C_0 + 1) - ct^2).$$

Let

$$t = \sqrt{\frac{2n(\log C_0 + 1)}{c}}.$$

Then,

$$n(\log C_0 + 1) - ct^2 = n(\log C_0 + 1) - 2n(\log C_0 + 1) = -n(\log C_0 + 1).$$

Hence, we have

$$P\left(\phi_{\max}(n) \geq K_{\max} + C + \sqrt{\frac{2(\log C_0 + 1)}{c}}\right) \leq 2\exp(-n(\log C_0 + 1)).$$

By taking $C_{\max} = K_{\max} + C + \sqrt{2(\log C_0 + 1)/c}$, the first part of the proof is completed. The analogous arguments show that Assumption 4 (ii) is also satisfied. Similarly, following Remark 5.40 of Vershynin (2012), we have

$$P\left(\phi_{\min}(k) \leq K_{\min} - C\sqrt{\frac{k}{n}} - \frac{t}{\sqrt{n}}\right) \leq 2\binom{p}{k}\exp(-ct^2).$$

Hence, we have

$$P\left(\phi_{\min}\left(\frac{n}{K}\right) \leq K_{\min} - \frac{C}{\sqrt{K}} - \frac{t}{\sqrt{n}}\right) \leq 2\exp\left(\frac{n}{K}(\log K + \log C_0 + 1) - ct^2\right)$$

for any K. Let

$$t = \frac{\sqrt{n} K_{\min}}{2}.$$

Then,

$$P\left(\phi_{\min}\left(\frac{n}{K}\right) \leq \frac{K_{\min}}{2} - \frac{C}{\sqrt{K}}\right) \leq 2\exp\left(-n\left(\frac{cK_{\min}^2}{4} - \frac{1}{K}\{\log K + \log C_0 + 1\}\right)\right).$$

By taking K so that

$$\frac{K_{\min}}{2} - \frac{C}{\sqrt{K}} > 0, \tag{2.10}$$

we have

$$\frac{cK_{\min}^2}{4} - \frac{1}{K}\{\log K + \log C_0 + 1\} > 0. \tag{2.11}$$

Notice that constants on left hand sides of (2.10) and (2.11) are independent of n. Thus, the second part of the proof is completed by taking $C_{\min} = K_{\min}/2 - C/\sqrt{K}$. □

Proof of Theorem 1. For a sufficiently large positive constant C such that $C \geq \sigma_\epsilon \sqrt{C^*}$, set $\lambda_0 = C\sqrt{2\log p/n}$. Under Assumption 4, standard arguments of normal distribution yield

$$P\left(\frac{1}{n}\|X^T \epsilon\|_\infty > \lambda_0\right) = O\left(\frac{1}{\sqrt{\log p}}\right),$$

and we have the oracle inequalities

$$\frac{1}{n}\|X(\hat{\beta} - \beta_0)\|^2 \leq \frac{8 s_0 \lambda_0^2}{\kappa_{\min}^2}, \quad \|\hat{\beta} - \beta_0\|_1 \leq \frac{4 s_0 \lambda_0}{\kappa_{\min}^2},$$

with probability tending to one. See Lemmas 6.1 and 6.3 and Theorem 6.1 in Bühlmann and van de Geer (2011). Recall that

$$\Delta_1 = \sqrt{n}(\hat{\Sigma}\hat{\Theta}_1 - e_1)^T(\beta_0 - \hat{\beta}).$$

Thus, it follows from Lemma 1 that for s_0 such that $s_0 = o(\sqrt{n/\log p})$

$$|\Delta_1| \leq \sqrt{n}\frac{\lambda_1}{\tilde{\tau}_1^2}\|\hat{\beta} - \beta_0\|_1 = o_P(1).$$

Lemma 1 also yields

$$\hat{\Omega}_{1,1} = \frac{\tilde{\tau}_1^2}{\hat{\tau}_1^4} \leq \frac{1}{\tilde{\tau}_1^2} = O_P(1).$$

This completes the proof. □

Proof of Corollary 1. The studentized debiased Lasso can be expressed as follows:

$$\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} = \frac{1}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \frac{1}{\sqrt{n}} \hat{\Theta}_1^T X^T \epsilon + \tilde{\Delta}_1,$$

where

$$\tilde{\Delta}_1 = \frac{\sqrt{n}(\hat{\Sigma}\hat{\Theta}_1 - e_1)^T(\beta_0 - \hat{\beta})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}}.$$

By the same argument in the proof of Theorem 1, we obtain $\tilde{\Delta}_1 = o_P(1)$. Let

$$Z_1 = \frac{1}{\sigma_\epsilon \hat{\Omega}_{1,1}^{1/2}} \frac{1}{\sqrt{n}} \hat{\Theta}_1^T X^T \epsilon \sim N(0,1).$$

By the same argument of Javanmard and Montanari (2014), we obtain

$$\begin{aligned} P\left(\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \le z\right) &= P\left(\frac{\sigma}{\hat{\sigma}_\epsilon} Z_1 + \tilde{\Delta}_1 \le z\right) \le P\left(Z_1 \le \frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon}(z+\delta)\right) + P\left(|\tilde{\Delta}_1| > \delta\right) \\ &\le \Phi(\delta(1+|z|) + \delta^2 + z) + P\left(\left|\frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon} - 1\right| > \delta\right) + P(|\tilde{\Delta}_1| > \delta) \end{aligned}$$

for any $\delta > 0$. By taking the limit supremum as $n \to \infty$ at first and the limit as $\delta \to 0$ subsequently, we obtain

$$\limsup_{n\to\infty} P\left(\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \le z\right) \le \Phi(z).$$

Next, for any $\delta > 0$ and $C > 0$, we have

$$\begin{aligned} \Phi(z-\delta) = P(Z_1 + \delta \le z) &= P\left(\frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon} \frac{1}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \frac{1}{\sqrt{n}} \hat{\Theta}_1^T X^T \epsilon + \delta \le z\right) \\ &\le P\left(\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \le z\right) + P\left(|\tilde{\Delta}_1| > \frac{\delta}{2}\right) + P\left(\left|\frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon} - 1\right| > \frac{\delta}{2C^2}\right) \\ &\quad + P\left(\left|\frac{1}{\hat{\Omega}_{1,1}^{1/2}} \frac{1}{\sqrt{n}} \hat{\Theta}_1^T X^T \epsilon\right| > C\right) + P\left(\frac{1}{\hat{\sigma}_\epsilon} > C\right). \end{aligned}$$

By taking the limit infimum as $n \to \infty$ at first, $C \to \infty$ subsequently, and $\delta \to 0$

afterward, we have

$$\Phi(z) \leq \liminf_{n\to\infty} P\left(\frac{\sqrt{n}(\hat{b}_1 - \beta_{01})}{\hat{\sigma}_\epsilon \hat{\Omega}_{1,1}^{1/2}} \leq z\right).$$

This completes the proof. $\square$

Proof of Theorem 2. It is sufficient to show that $\tilde{\Delta}_1(\lambda_1) = o_P(1)$ and $\hat{\Omega}_{1,1}(\lambda_1) = O_P(1)$. Recall that $f(\lambda) := \|\hat{\Sigma}\hat{\Theta}_1(\lambda) - e_1\|_\infty / \hat{\Omega}_{1,1}^{1/2}(\lambda)$ and $\eta_1^* := \tilde{\tau}_1(\lambda_{\mathrm{CV}})/\sqrt{n}$. First, suppose that $f(\lambda) > \eta_1^*$ for all $\lambda \in \Lambda$. Then, our procedure selects $\lambda_1 = \lambda_{\min}$, and the KKT conditions for the node-wise Lasso yield

$$f(\lambda_1) = \frac{\|\hat{\Sigma}\hat{\Theta}_1(\lambda_{\min}) - e_1\|_\infty}{\hat{\Omega}_{1,1}^{1/2}(\lambda_{\min})} \leq \frac{\lambda_{\min}}{\tilde{\tau}_1(\lambda_{\min})}.$$

Second, suppose that $f(\lambda) \leq \eta_1^*$ for some $\lambda \in \Lambda$. Because $\tilde{\tau}_1^2(\lambda)$ is a nondecreasing function in $\lambda \in \Lambda$, we obtain

$$f(\lambda_1) \leq f(\lambda^*) \leq \eta_1^* \leq \frac{\tilde{\tau}(\lambda_{\max})}{\sqrt{n}} = \frac{1}{\sqrt{n}}\left(\frac{1}{n}\|X_1\|^2\right)^{1/2},$$

where the fourth equality follows from $\hat{\gamma}_1(\lambda_{\max}) = 0$ under Assumption 1.

Consequently, for the tuning parameter λ_1 selected by our procedure, we obtain

$$f(\lambda_1) \leq \max\left\{\frac{\lambda_{\min}}{\tilde{\tau}_1(\lambda_{\min})}, \frac{1}{\sqrt{n}}\left(\frac{1}{n}\|X_1\|^2\right)^{1/2}\right\} = O_P\left(\frac{1}{\sqrt{n}}\right),$$

by Lemma 1. Therefore, by Hölder inequality,

$$|\tilde{\Delta}_1(\lambda_1)| \leq f(\lambda_1)\frac{1}{\hat{\sigma}_\epsilon}\|\hat{\beta} - \beta_0\|_1 = O_P\left(\sqrt{\frac{s_0^2 \log p}{n}}\right) = o_P(1).$$

Moreover, we obtain

$$\hat{\Omega}_{1,1}(\lambda_1) \leq \frac{1}{\tilde{\tau}_1^2(\lambda_1)} \leq \frac{1}{\tilde{\tau}_1^2(\lambda_{\min})} = O_P(1),$$

by Lemma 1. This completes the proof. $\square$

Chapter 3

Cross-Validation for High-Dimensional Ridge Regression under Misspecified Linear Models

3.1 Introduction

Recently, high dimensional prediction gets highly attention. Consider a linear regression, and let p be the number of covariates and n be the sample size. It is well-known that when $n > p$ the least squares estimator which includes many covariates gives the large out-of-sample prediction error. When $p > n$, the least squares objective functions have infinitely many minimizers. Among them, the minimum ℓ_2 norm (min-norm) least squares is investigated actively. If the design matrix is full row rank, the min-norm least squares estimator is an interpolator, which is an estimator yielding zero residuals. Surprisingly, recent studies such as Bartlett et al. (2020) showed that the min-norm least squares estimator yields the small out-of-sample prediction error. They call such phenomenon as "benign overfitting." Furthermore, Hastie et al. (2022) also showed that the so-called "double descent" phenomenon (Advani et al. 2020, Spigler et al. 2019, Belkin et al. 2019) occurs for the min-norm least squares estimator. That is, as p increases when $n > p$, the out-of-sample prediction error of the min-norm least squares first decreases and increases while p is closing to n. However, when p exceeds n, the out-of-sample prediction error decreases again as p increases.

The min-norm least squares estimator can be seen as the extended ridge regression (see Section 3.2) whose tuning parameter is zero. However, it is not true that choosing zero for the tuning parameter of the extended ridge regression always achieves the optimal prediction. For example, Hastie et al. (2022) show that when the coefficients of linear regression models follow the spherical prior and covariates are isotropic, the ridge regression with positive tuning parameters outperforms the min-norm least squares in prediction. On the other hand, Kobak et al. (2020) showed that zero or even negative regularization for the ridge regression can yield the optimal prediction, particularly when the coefficient of the linear regression model is aligned with the leading eigenvectors of the covraiance of covariates. Because we do not know which tuning parameter is optimal, we have to rely on data driven selection procedures in practice.

Among them, the cross-validation is one of the most popular methods. An excellent survey of the method is given by Arlot and Celisse (2010). There are various forms for the method, such as the 5 or 10-fold cross-validations. The most widely used form is the leave-one-out cross-validation. Henceforth, we call the leave-one-out cross-validation as CV for simplicity. However, CV is computationally intensive. To avoid the computational burden, the generalized cross-validation (GCV), which is a simple method and an approximation of CV is also used widely.

CV and GCV were studied from long decades in the $n > p$ setting for the tuning parameter selection of the ridge regression. For example, see Li (1986) and Golub et al. (1979). In the $p > n$ setting, Hastie et al. (2022) showed that CV gives asymptotically optimal tuning for the ridge regression. However, they make strict assumptions: covariates are independent and coefficients follow the spherical prior with mean zero.

Recently, Patil et al. (2021) examined GCV and CV in more general assumptions for linear regression models. They allow that covariance matrices can have general structures as long as their eigenvalues are bounded below and above by positive constants. They also allow that coefficients of linear regression models can be non random with bounded ℓ_2 norm. Under these assumptions, they showed that CV and GCV tuning are asymptotically optimal tuning, even when candidate sets of the tuning parameter have not only positive values but also zero and negative values.

However, assuming that the true model is a linear regression model may not be realistic. When p is large compared with n, it is highly possible that relationships between the response and covariates are nonlinear. Because linear regression models are often used for prediction, it is more realistic that we specify the true model as a linear regression model incorrectly.

Li (1986) investigated GCV under misspecified linear regression models with the $n > p$ setting, and showed that GCV tuning is optimal tuning in the sense of the in-sample prediction error for the ridge regression. However, as far as we know, no study investigated CV and GCV tuning for the ridge regression under misspecified linear regression models where p can be larger than n.

In this study, we investigate CV and GCV tuning for the ridge regression under misspecified linear regression models where p can be larger than n. We will show that if errors between the true regression function and the best linear approximation converge to zero, then CV and GCV tuning are still asymptotically equivalent to optimal tuning in the sense of the out-of-sample prediction error. As in Patil et al. (2021), we also allow that candidates for the tuning parameter include not only positive values but also zero and negative values.

The remainder of this article is organized as follows. In Section 3.2, we formally introduce the model, the extended ridge regression, the objective functions of CV and GCV, and the out-of-sample prediction error. In Section 3.3, we introduce the assumptions and show that the CV and GCV tuning are asymptotically equivalent with optimal tuning. In Section 3.4, we compared prediction performances of the ridge regression tuned by CV and GCV with those of the Lasso tuned by the 10-fold cross-validation. Section 3.5 concludes this article. Section 3.6 contains the proofs of theoretical results.

3.2 Model and Notations

First, we introduce some notations. Let a^T be the transpose of a vector a. Let $\|a\|$ be the ℓ_2 norm of a vector a. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and maximum eigenvalue of a matrix A, respectively. Let $\|A\|$ and $\mathrm{tr}[A]$ be the operator norm and the trace of a matrix A, respectively. Let C be a generic positive constant that can vary from line to line.

Consider a regression model

$$y_i = f(x_i) + \epsilon_i, \tag{3.1}$$

where $y_i \in \mathbb{R}$ is the response variable, $x_i = (x_{i,1}, \ldots, x_{i,p})^T \in \mathbb{R}^p$ is the p dimensional covariate vector, $f : \mathbb{R}^p \to \mathbb{R}$ is a regression function, and ϵ_i is an error term. The samples $\{(y_1, x_1^T)^T, \ldots, (y_n, x_n^T)^T\}$ are independently and identically distributed. We consider the proportional asymptotics: $p/n \to \gamma \in (0, \infty)$. We also assume that the model (3.1) is misspecified as the linear regression model. That is, we estimate

$$y_i = x_i^T \beta_0 + \xi_i, \tag{3.2}$$

where $\beta_0 \in \mathbb{R}^p$ is some non random p dimensional vector, ξ_i is an error term which is given by $\xi_i = \delta_i + \epsilon_i$, and δ_i is a specification error: $\delta_i = f(x_i) - x_i^T \beta_0$.

We introduce the matrix representation of the misspecified model (3.2). Let $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$, and $\xi = (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$. Then, the matrix form of (3.2) is given by

$$Y = X\beta_0 + \xi, \tag{3.3}$$

where $\xi = \delta + \epsilon$, $\delta = (\delta_1, \ldots, \delta_n)^T \in \mathbb{R}^n$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$. We assume that X is full row rank: $\mathrm{rank}(X) = n$.

We estimate the coefficient of the misspecified linear model β_0 by the extended ridge regression. To define the extended ridge estimators, first we introduce the usual ridge estimators. Let $\lambda > 0$ be a tuning parameter for the ridge regression. The ridge estimator $\hat{\beta}_\lambda$ is given by

$$\hat{\beta}_\lambda := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \|Y - X\beta\|^2 + \lambda \|\beta\|^2. \tag{3.4}$$

We can write $\hat{\beta}_\lambda$ explicitly as

$$\hat{\beta}_\lambda = (\hat{\Sigma} + \lambda I_p)^{-1} \frac{1}{n} X^T Y, \tag{3.5}$$

where $\hat{\Sigma} = \frac{1}{n} X^T X$. If we use (3.5) as the ridge estimator, then the min-norm least squares estimator is given by $\lim_{\lambda \to 0} \hat{\beta}_\lambda$. This can be shown by applying the singular value decomposition to X. As in Patil et al. (2021), to extend the range of λ including zero and negative values, we define the extended ridge estimator as

$$\hat{\beta}_\lambda = (\hat{\Sigma} + \lambda I_p)^{+} \frac{1}{n} X^T Y, \tag{3.6}$$

where A^+ is the Moore-Penrose pseudoinverse of a matrix A. Henceforth, we call $\hat{\beta}_\lambda$ in (3.6) as the ridge estimator. Note that when A is invertible, then $A^+ = A^{-1}$. In addition, note that we will set the range of λ so that the pseudoinverse is defined only when $\lambda = 0$ asymptotically. This is because if the absolute value of λ is less than the asymptotic smallest positive eigenvalue of $\hat{\Sigma}$, then $\hat{\Sigma} + \lambda I_p$ will be invertible even when λ is negative. However, notice that when λ is negative, $\hat{\Sigma} + \lambda I_p$ is not positive semidefinite because $p - n$ eigenvalues of the matrix are negative.

First, we investigate GCV because it is simpler than CV. After that, we investigate CV by using the fact that GCV is an approximation of CV. Let $\Lambda \subset (\lambda_{\min}, \infty)$ be a closed interval which is the set of candidates for λ. $\lambda_{\min}$ is defined in Section 3.3, and it can be negative. To define objective functions of GCV and CV, let S_λ be the ridge smoother matrix, which is defined as

$$S_\lambda = X(\hat{\Sigma} + \lambda I_p)^{+}\frac{1}{n}X^T, \tag{3.7}$$

so that we get $S_\lambda Y = X\hat{\beta}_\lambda$. Then, the objective function of GCV is written as

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\hat{\beta}_\lambda}{1 - \frac{1}{n}\mathrm{tr}[S_\lambda]}\right)^2. \tag{3.8}$$

However, if we plug in $\lambda = 0$ for $\mathrm{GCV}(\lambda)$ in (3.8), we get $0/0$ because $\hat{\beta}_0$ is interpolating and $S_0 = I_n$ when X is full row rank. Therefore, as in Hastie et al. (2022) and Patil et al. (2021), we define GCV(0) as the limit of $\mathrm{GCV}(\lambda)$ as $\lambda \to 0$. It is given by

$$\mathrm{GCV}(0) := \lim_{\lambda \to 0}\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{Y^T(\frac{1}{n}XX^T)^{+2}Y}{(\frac{1}{n}\mathrm{tr}(\hat{\Sigma}^{+}))^2}.$$

On the other hand, the objective function of CV is given by

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta}_{-i,\lambda})^2,$$

where $\hat{\beta}_{\lambda,-i}$ is the ridge estimator derived by the training data (X, Y) without the ith observation. If we use this objective function, we have to compute the ridge regression n times. It is computationally intensive particularly when n and p are large. However, it is well-known that we can express this objective function as

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\hat{\beta}_\lambda}{1 - (S_\lambda)_{i,i}}\right)^2, \tag{3.9}$$

by matrix algebra, where $(S_\lambda)_{i,i}$ is the ith diagonal element of S_λ. For example, see Section 7 of Hastie et al. (2022). Similar to GCV, if we plug in $\lambda = 0$ for $\mathrm{CV}(\lambda)$ in (3.9), we get $0/0$. Therefore, we also define CV(0) as the limit of $\mathrm{CV}(\lambda)$ as $\lambda \to 0$. It is given by

$$\mathrm{CV}(0) := \lim_{\lambda \to 0}\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left[(\frac{1}{n}XX^T)^{+}Y\right]_i}{[(\frac{1}{n}XX^T)^{+}]_{i,i}}\right)^2.$$

The tuning parameters selected by GCV and CV from Λ are given by

$$\hat{\lambda} = \operatorname*{argmin}_{\lambda \in \Lambda} \mathrm{GCV}(\lambda), \tag{3.10}$$

and

$$\bar{\lambda} = \operatorname*{argmin}_{\lambda \in \Lambda} \mathrm{CV}(\lambda), \tag{3.11}$$

respectively.

Next, we introduce an optimal tuning parameter of the ridge regression. As in Patil et al. (2021), we use the conditional out-of-sample prediction error as a measure of the prediction accuracy for the ridge regression. Henceforth, we call the out-of-sample prediction error as the risk for simplicity. The risk is given by

$$R(\lambda) = E[(y_0 - x_0^T \hat{\beta}_\lambda)^2 | X, Y], \tag{3.12}$$

where the expectation is taken over the test data $(x_0, y_0) \in \mathbb{R}^{p+1}$ conditionally on the training data (X, Y). Notice that the test data (x_0, y_0) are independent of the training data (X, Y). The optimal tuning parameter of the ridge regression is given by the minimizer of the risk from Λ. Namely, the optimal tuning parameter λ^* is given by

$$\lambda^* = \operatorname*{argmin}_{\lambda \in \Lambda} R(\lambda).$$

Note that λ^* is a random variable because it depends on the training data (X, Y).

Our purpose is to show that CV and GCV tuning are asymptotically optimal tuning. To achieve the purpose, our goals are to show that the risks by CV and GCV and the optimal risk are asymptotically equivalent even when the model is misspecified. That is, we will show that

$$R(\hat{\lambda}) - R(\lambda^*) \xrightarrow{a.s.} 0, \tag{3.13}$$

and

$$R(\bar{\lambda}) - R(\lambda^*) \xrightarrow{a.s.} 0, \tag{3.14}$$

under misspecified models.

To finish the goals, we firstly analyse properties of GCV. Actually, Patil et al. (2021) showed that GCV and CV are asymptotically equivalent even when the model is misspecified. Therefore, it is enough to show the asymptotic equivalence between GCV and the risk.

3.3 Theoretical Results

To analyse CV and GCV for the ridge regression, we present assumptions. First, we put the following assumption on the error terms.

Assumption 1. Error term ϵ_i, which is independent of x_i, is independently and identically distributed with mean zero, variance σ^2, and finite $4+\eta$ moment for some $\eta > 0$.

Assumptions of homoskedasticity on error variances are often used in high-dimensional statistics.

Next, we make the following assumption on the covariates.

Assumption 2. The row vectors x_i of X is expressed as $x_i = \Sigma^{1/2} z_i$, where $\Sigma \in \mathbb{R}^{p\times p}$ is a positive definite matrix, and z_i is a random vector whose elements are independently and identically distributed with mean zero, variance one, and finite $4+\eta$ moment for some $\eta > 0$.

This assumption is essential to bound the maximum eigenvalue and the smallest positive eigenvalue of the sample covariance. See Theorem 1 of Bai and Yin (1993).

Next, we make the following assumption on eigenvalues of Σ.

Assumption 3. There exists positive constants $C_{\min}$ and $C_{\max} < \infty$ such that $C_{\min} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_{\max}$.

This assumption is also important to bound eigenvalues of the sample covariance. Assumptions of bounded eigenvalues of the population covariance Σ are also popular in high-dimensional statistics.

Note that Assumptions 1–3 are also used in Patil et al. (2021). Now we introduce the next assumption on the regression function, which is a new one.

Assumption 4. (i) β_0 is given by

$$\beta_0 = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} E[(f(x_i) - x_i^T\beta)^2].$$

Furthermore, there is a positive constant C such that $\|\beta_0\|^2 \leq C$.
(ii) $f(x_i)$ satisfies $E[f(x_i)^2] \leq C$ for some $C > 0$. Furthermore, the squared approximation error is bounded and asymptotically negligible. That is, $E[(f(x_i) - x_i^T\beta_0)^2] \leq C$ for some $C > 0$ and

$$E[(f(x_0) - x_0^T\beta_0)^2] \to 0,$$

as $p \to \infty$.

Notice that Patil et al. (2021) assume that the true models is a linear regression model. Therefore, in their study, there is no approximation error: $E[(f(x_i) - x_i^T \beta_0)^2] = 0$ because $f(x_i) = x_i^T \beta_0$ for any i. On the other hand, Assumption 4 is weaker than their assumption: the squared approximation error is not needed to be zero for fixed n. Note that we do not need sparsity on the projection coefficient β_0.

Let $\lambda_{\min}$ be $\lambda_{\min} = -C_{\min}(1 - \sqrt{\gamma})^2$. This $\lambda_{\min}$ is the lower bound for the tuning parameter λ which makes the matrix $\hat{\Sigma} + \lambda I_p$ invertible asymptotically.

Now we give a main result on GCV under the misspecified linear model.

Theorem 1. *Under Assumptions 1–4, we have*

$$\mathrm{GCV}(\lambda) - R(\lambda) \xrightarrow{a.s.} 0, \tag{3.15}$$

for all $\lambda \in (\lambda_{\min}, \infty)$. *Furthermore, the uniform convergence holds on any compact sub interval* $\Lambda \subset (\lambda_{\min}, \infty)$.

Theorem 1 implies that GCV is asymptotically optimal tuning.

Theorem 2. *Under Assumptions 1–4, we get*

$$R(\hat{\lambda}) - R(\lambda^*) \xrightarrow{a.s.} 0.$$

Therefore, GCV chooses the optimal tuning parameter asymptotically even when the regression model is misspecified as a linear regression model.

To prove Theorem 1, we also take the similar approaches of Patil et al. (2021). That is, we decompose the risk and the objective function of GCV into bias components, variance components, and cross components of bias and variance. That is, we rewrite $R(\lambda)$ and $\mathrm{GCV}(\lambda)$ as

$$R(\lambda) = R_b(\lambda) + R_v(\lambda) + R_c(\lambda),$$

and

$$\mathrm{GCV}(\lambda) = \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} + \frac{\mathrm{GCV}_v(\lambda)}{\mathrm{GCV}_d(\lambda)} + \frac{\mathrm{GCV}_c(\lambda)}{\mathrm{GCV}_d(\lambda)},$$

respectively, where $\mathrm{GCV}_d(\lambda)$ is the denominator of $\mathrm{GCV}(\lambda)$: $\mathrm{GCV}_d(\lambda) = (1 - \frac{1}{n}\mathrm{tr}[S_\lambda])^2$. These components are defined precisely in Section 3.6. First, we will show that these cross components are asymptotically negligible. Then, we will show that bias components and variance components of the risk and GCV are asymptotically equivalent, respectively.

First, we show that cross components of the risk and GCV are asymptotically negligible.

Lemma 1. *Suppose that Assumptions 1–4 hold. Then for all* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$R_c(\lambda) \xrightarrow{a.s.} 0. \tag{3.16}$$

Lemma 2. *Suppose that Assumptions 1–4 hold. Then for all* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$\frac{\mathrm{GCV}_c(\lambda)}{\mathrm{GCV}_d(\lambda)} \xrightarrow{a.s.} 0. \tag{3.17}$$

We note that these cross components are more complex than cross components of Patil et al. (2021) due to the model misspecification.

Next, we will show that bias components are asymptotically equivalent.

Lemma 3. *Suppose that the Assumptions 1–4 hold. Then for all* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$R_b(\lambda) - \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} \xrightarrow{a.s.} 0. \tag{3.18}$$

Actually, the most different part between us and Patil et al. (2021) is to show the asymptotic equivalence of the bias components. Because the model is correctly specified in Patil et al. (2021), their bias components are relatively simple. On the other hand, because the model is incorrectly specified in this study, the bias components are much more complex. However, Assumption 4 enables us to show the asymptotic equivalence of the bias components.

We define variance components as quadratic forms with respect to the error terms. Therefore, the variance components do not change from those of Patil et al. (2021) so that their results still hold under the misspecification.

Lemma 4. (Lemma 5.4 in Patil et al. 2021.) *Suppose that Assumptions 1–3 hold. Then for all* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$R_v(\lambda) - \frac{\mathrm{GCV}_v(\lambda)}{\mathrm{GCV}_d(\lambda)} \xrightarrow{a.s.} 0. \tag{3.19}$$

Lemmas 1–4 imply the pointwise convergence part of Theorem 1. The uniform convergence part can be established by showing that the risk and the objective function of

GCV and their derivatives are bounded over any compact sub interval of $\Lambda \subset (\lambda_{\min}, \infty)$ so that the Arzela-Ascoli theorem holds.

Now we will examine CV. Actually, regardless of whether the model is correctly specified or incorrectly specified, the asymptotic equivalence between CV and GCV holds as long as the response vector Y has a finite second moment under Assumptions 2 and 3. See Theorem 4.2 of Patil et al. (2021).

Lemma 5. (Theorem 4.2 of Patil et al. 2021.) *Suppose that Assumptions 1–4 hold. Then for all* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$\mathrm{GCV}(\lambda) - \mathrm{CV}(\lambda) \xrightarrow{a.s.} 0. \tag{3.20}$$

Furthermore, the uniform convergence holds on any compact sub interval $\Lambda \subset (\lambda_{\min}, \infty)$*.*

3.4 Simulation Studies

In this section, we investigate finite sample prediction performances of the ridge regression tuned by CV and GCV. As a reference, we compare them with prediction performances of the Lasso (Tibshirani 1996). To implement the Lasso, we employ the R package `glmnet` (Friedman et al. 2010). We choose the tuning parameter of the Lasso by the 10-fold cross-validation with the `lambda.min` option. The candidates of the tuning parameter of the Lasso are also generated by `glmnet`. On the other hand, we set the candidates of the tuning parameter for the ridge regression as Λ, which includes 100 breakpoints of the interval $[\lambda_1, \lambda_2]$ and zero, where $\lambda_1 = \lambda_{\min} + 0.01$, $\lambda_{\min} = -\lambda_{\min}(\Sigma)(1 - \sqrt{\gamma})^2$, and $\lambda_2 = 5$. In all settings, we set the covariance of covariates as $\Sigma = I + \rho \mathbf{1}\mathbf{1}^T$ with $\rho = 0.1$ as in Kobak et al. (2020). Note that the eigen equation of this Σ is given by $\phi(\lambda) = (1 + p\rho - \lambda)(1 - \lambda)^{p-1}$. Therefore, $\lambda_{\min}(\Sigma) = 1$ and $\lambda_{\max}(\Sigma) = 1 + p\rho$. We also set $n = 50$ and $p = \gamma n$ with $\gamma \in \{0.1, 0.2, 0.5, 0.9, 1, 2, 5, 10, 20, 50\}$.

We consider the following three settings.

- Setting 1. We borrow the simulation setting from Kobak et al. (2020). The data generating process is given by

$$y_i = x_i^T \beta + \epsilon_i,$$

  where $x_i \sim N(0, \Sigma)$, and $\epsilon_i \sim N(0, 1)$. β is given by $\beta = (b, \dots, b)^T \in \mathbb{R}^p$, with $b = \sqrt{\alpha/(p + p^2\rho)}$, $\alpha = 10$, and $\rho = 0.1$.

- Setting 2. We consider the following nonlinear data generating process with respect to x_i which includes a linear component:

$$y_i = f(x_i) + \epsilon_i,$$

where $x_i \sim N(0, \Sigma)$, $\epsilon_i \sim N(0,1)$, and

$$f(x_i) = x_i^T \beta + \frac{1}{\sqrt{(1+\rho)p \log n}} \|x_i\|.$$

Σ and β are the same as in Setting 1. Note that simple calculations yield

$$E[(f(x_i) - x_i^T \beta_0)^2] \leq \frac{1}{(1+\rho)p \log n} E[\|x_i\|^2] = \frac{1}{\log n},$$

for the linear projection coefficient β_0. Therefore, the linear approximation error for $f(x_i)$ is getting smaller with the speed of $1/\log n$.

- Setting 3. We consider the following nonlinear data generating process with respect to x_i:

$$y_i = f(x_i) + \epsilon_i,$$

where $x_i \sim N(0, \Sigma)$, $\epsilon_i \sim N(0,1)$, and

$$f(x_i) = \sum_{j=1}^{p} \sin(x_{i,j}) \beta_j.$$

Σ and β is the same as in Setting 1.

To compare the prediction accuracy of the three methods, we calculated the empirical risk, which is given by

$$\frac{1}{m} \sum_{k=1}^{m} \left(y_0^{(k)} - x_0^{(k)T} \hat{\beta}^{(k)} \right)^2,$$

where $m = 100$, $(y_0^{(k)}, x_0^{(k)})$ are the test data on the kth repetition, and $\hat{\beta}^{(k)}$ is an estimator of the coefficient of the misspecified linear regression model derived by the training data on the kth repetition.

Now we check the simulation results in Setting 1. Table 3.1 shows the empirical risks of the ridge regression and the Lasso. When $\gamma \leq 0.2$, their empirical risks are similar. However, differences of the empirical risks occur when $\gamma \geq 0.5$. As γ increases, the empirical risks of the Lasso tend to be large. It seems that the double descent phenomenon cannot be observed for the Lasso tuned by the 10-fold cross-validation in this setting. On the other hand, the empirical risks of the ridge regression tend to increase as γ is close to one, but they turn to decrease when γ exceeds one. Furthermore, the empirical risks of the ridge regression continue to decrease as γ is getting larger than one, indicating that the double descent phenomenon is observed for the ridge regression. Therefore, as γ increases, the empirical risks of the ridge regression are getting significantly smaller than those of the Lasso. In particular, when $\gamma = 50$, the empirical risks of the ridge regression are approximately three times smaller than those of the Lasso.

Table 3.2 shows the average of selected tuning parameters by CV and GCV for the ridge regression, and the 10-fold cross-validation for the Lasso in Setting 1. CV and GCV in the ridge regression selected similar values. When $0.1 \leq \gamma \leq 2$, tuning parameters selected by CV and GCV seem to increase. But when $5 \leq \gamma \leq 50$, tuning parameters selected by CV and GCV are decreasing, and they turn to become negative when $\gamma \geq 10$. Furthermore, when $\gamma = 50$, CV and GCV choose negatively large values, but corresponding empirical risks are small. Kobak et al. (2020) found that when $n = 64$ and $p = 1000$, an empirically optimal tuning parameter is about -2.344 (Note that their objective function of the ridge regression is not divided by n). Although we do not report the empirical risks of CV and GCV when $n = 64$ and $p = 1000$ here, average tuning parameters selected by them are -2.975 and -2.916, respectively. Therefore, CV and GCV would select optimal or nearly optimal tuning parameters in this setting. On the other hand, the average tuning parameters of the Lasso selected by the 10-fold cross-validation do not change significantly over all γ. This and denseness of β may be reasons why the prediction performances of the Lasso are not well particularly when $\gamma \geq 0.5$.

Next, we check the simulation results in Setting 2. Table 3.3 presents the empirical risks in Setting 2. Due to misspecification, the empirical risks of the three methods become larger than those of Setting 1. However, the behaviors of the empirical risks are similar. The empirical risks of the ridge regression are increasing when $0.1 \leq \gamma \leq 1$, and they tend to decrease as γ is getting larger than one. Note that when $\gamma = 2$, the ridge regression tuned by GCV has the relatively larger empirical risks than that of CV. This is because GCV selected a negatively large tuning parameter $\lambda = -0.16$ at 44th repetition, which may not be nearly optimal, and the corresponding error at the repetition is 141.25, making the mean of the errors large. On the other hand, CV selects -0.06 at the repetition, and the corresponding error is 10.01, which is much smaller than the error of the Lasso: 25.85. Meanwhile, the empirical risks of the Lasso tend to increase as γ is getting large. The ridge regression tuned by CV tend to outperform the Lasso more than Setting 1.

Table 3.4 presents the mean of selected tuning parameters by CV and GCV in the ridge regression, and the 10-fold cross-validation in the Lasso, respectively in Setting 2. All mean values of the tuning parameters selected by all methods become slightly larger on the positive direction than Setting 1, but behaviors of them are similar. The tuning parameters selected by CV and GCV are still negative in Setting 2 when γ is large. On the other hand, as in Setting 1, variations of the tuning parameters of the Lasso selected by the 10-fold cross-validation are small.

Finally, we check the simulation results in Setting 3. Table 3.5 reports the empirical risks in Setting 3, indicating that the ridge regression tuned by CV outperforms the Lasso in all γ. Interestingly, when $\gamma \geq 0.2$, the empirical risks of the ridge regression tuned by CV and the Lasso become smaller than Setting 2. We can see the first descent of the empirical risks in $0.1 \leq \gamma \leq 2$, and they seem to be increasing in $0.2 \leq \gamma \leq 1$ for all methods. We can see the second descent when $\gamma \geq 1$ because the empirical risks of all methods seem to be decreasing slowly. We also see that the risks of the ridge regression tuned by GCV becomes large. At the 44th repetition, GCV selected -0.16 as the tuning parameter, and the corresponding error is 221.76. On the other hand, CV selected -0.11 as the tuning parameter, but the corresponding error is 0.1. We do not know why the big difference of the errors occurs between GCV and CV, although the difference of the selected tuning parameters is just about 0.05.

Table 3.6 reports the average of selected tuning parameters by CV and GCV in the ridge regression, and the 10-fold cross-validation in the Lasso, respectively in Setting 3. The tuning parameters selected by CV and GCV become larger on positive direction than those of Setting 2, but they also turn to be negative when γ is large. Although the tuning parameters of the Lasso becomes slightly larger than those of Settings 1 and 2, variation is still small. Furthermore, when $\gamma \geq 20$, the tuning parameters of the Lasso are much similar in all settings.

In conclusion, under the spiked covariance with low correlation, the ridge regression tuned by CV and GCV generally outperforms the Lasso tuned by the 10-fold cross-validation in correctly specified and incorrectly specified settings. The superiority of the ridge regression tuned by CV and GCV stands out particularly when p is much larger than n. We also observed that the ridge regression tuned by CV has stable prediction performances than GCV. GCV sometimes selects unnecessarily large values on negative direction particularly when p is moderately larger than n under the misspecification.

Table 3.1: Empirical risks in Setting 1 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	1.185	0.993	1.107	1.456	1.844	1.532	1.872	1.429	1.751	1.262
GCV	1.171	0.999	1.093	1.491	1.869	1.57	1.887	1.487	1.744	1.223
Lasso	1.184	1.01	1.64	2.606	2.843	2.712	3.639	3.895	4.11	4.101

Table 3.2: The average of selected tuning parameters in Setting 1 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	0.031	0.069	0.284	0.498	0.477	0.625	0.091	-1.638	-6.689	-25.092
GCV	0.038	0.067	0.291	0.514	0.497	0.65	0.17	-1.552	-6.521	-24.661
Lasso	0.015	0.014	0.043	0.064	0.057	0.05	0.049	0.043	0.049	0.053

3.5 Conclusion

In this study, we showed that if the approximation error between the regression function and the linear projection can be asymptotically negligible, the leave-one-out cross-validation and the generalized cross-validation still can select the asymptotically optimal tuning parameters of the ridge regression under the misspecified linear model. We also allow that the candidates of the tuning parameter of the ridge regression can have not only positive values, but also zero and negative values. Simulation studies showed that if the covariance of the covariates is a spiked covariance and correlation is low, the ridge regression tuned by the leave-one-out cross-vaidation and the generalized cross validation yields the better prediction performances than the Lasso tuned by the 10-fold cross-validation regardless of whether the model is correctly or incorrectly specified.

3.6 Appendix

3.6.1 Proofs

Let us define $z = \mu(X) - X\beta_0 \in \mathbb{R}^n$, where $\mu(X) = E[Y|X] = (f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n$. We decompose the risk $R(\lambda)$ into the bias components $R_b(\lambda)$, the variance component $R_v(\lambda)$, and the cross components of the bias and the variance $R_c(\lambda)$. That is, for $\lambda \in$

Table 3.3: Empirical risks in Setting 2 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	1.433	1.4	1.6	2.06	2.166	1.716	2.299	1.806	2.037	1.434
GCV	1.426	1.408	1.606	2.083	2.142	3.133	2.269	1.8	2.122	1.409
Lasso	1.439	1.443	2.263	3.39	3.462	3.18	3.621	3.915	4.689	4.590

Table 3.4: The average of selected tuning parameters in Setting 2 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	0.034	0.084	0.331	0.558	0.533	0.809	0.228	-1.461	-5.937	-24.027
GCV	0.039	0.085	0.328	0.564	0.55	0.815	0.329	-1.404	-5.809	-23.638
Lasso	0.014	0.015	0.005	0.081	0.067	0.065	0.054	0.052	0.055	0.051

Table 3.5: Empirical risks in Setting 3 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	2.507	1.213	1.472	1.383	1.603	1.358	1.662	1.307	1.598	1.291
GCV	2.501	1.221	1.508	1.479	1.602	3.469	1.674	1.459	1.564	1.280
Lasso	2.574	1.344	1.846	1.929	2.068	1.937	2.396	1.911	2.555	1.909

Table 3.6: The average of selected tuning parameters in Setting 3 with $p = \gamma n$ and $n = 50$.

γ	0.1	0.2	0.5	0.9	1	2	5	10	20	50
CV	0.154	0.262	0.646	1.028	0.973	1.404	1.515	0.601	-3.055	-17.961
GCV	0.096	0.21	0.621	1.037	0.99	1.414	1.646	0.628	-2.839	17.412
Lasso	0.034	0.036	0.099	0.115	0.098	0.098	0.067	0.057	0.051	0.037

$(\lambda_{\min}, \infty)$,

$$\begin{aligned}
R_b(\lambda) &= E[(f(x_0) - x_0^T\beta_0)^2] + \beta_0^T\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0 \\
&\quad + \frac{2}{n}z^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0 + \frac{1}{n}z^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma(\hat{\Sigma} + \lambda I_p)^+\frac{1}{n}X^T z, \\
R_v(\lambda) &= \sigma_\epsilon^2 + \frac{1}{n}\epsilon^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma(\hat{\Sigma} + \lambda I_p)^+\frac{1}{n}X^T\epsilon, \\
R_c(\lambda) &= \frac{2}{n}z^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma(\hat{\Sigma} + \lambda I_p)^+\frac{1}{n}X^T\epsilon + \frac{2}{n}\epsilon^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0.
\end{aligned}$$

Simple calculations show that

$$R(\lambda) = R_b(\lambda) + R_v(\lambda) + R_c(\lambda).$$

Next, we decompose the numerator of $\mathrm{GCV}(\lambda)$ into the bias components $\mathrm{GCV}_b(\lambda)$, the variance component $\mathrm{GCV}_v(\lambda)$, and the cross components of the bias and the variance $\mathrm{GCV}_c(\lambda)$.

$$\begin{aligned}
\mathrm{GCV}_b(\lambda) &= \frac{1}{n}\beta_0^T X^T(I_n - S_\lambda)^2 X\beta_0 + \frac{2}{n}\beta_0^T X^T(I_n - S_\lambda)^2 z + \frac{1}{n}z^T(I_n - S_\lambda)^2 z, \\
\mathrm{GCV}_v(\lambda) &= \frac{1}{n}\epsilon^T(I_n - S_\lambda)^2\epsilon, \\
\mathrm{GCV}_c(\lambda) &= \frac{2}{n}\beta_0^T X^T(I_n - S_\lambda)^2\epsilon + \frac{2}{n}z^T(I_n - S_\lambda)^2\epsilon.
\end{aligned}$$

The denominator of $\mathrm{GCV}_d(\lambda)$ can be written as

$$\mathrm{GCV}_d(\lambda) = \left(1 - \mathrm{tr}[S_\lambda]\frac{1}{n}\right)^2 = \left(1 - \mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^+\hat{\Sigma}]\frac{1}{n}\right)^2.$$

Simple calculations yield

$$\mathrm{GCV}(\lambda) = \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} + \frac{\mathrm{GCV}_v(\lambda)}{\mathrm{GCV}_d(\lambda)} + \frac{\mathrm{GCV}_c(\lambda)}{\mathrm{GCV}_d(\lambda)}.$$

Now we proceed proofs of theoretical results.

Proof of Lemma 1. First, we evaluate the first term in $R_c(\lambda)$.

$$\frac{2}{n}\left|z^T X(\hat{\Sigma} + \lambda I_p)^+\Sigma(\hat{\Sigma} + \lambda I_p)^+\frac{1}{n}X^T\epsilon\right| \leq \frac{2}{n^2}\|z\|\|X(\hat{\Sigma} + \lambda I_p)^+\Sigma(\hat{\Sigma} + \lambda I_p)^+X^T\epsilon\|$$

$$\leq \frac{2}{n^2}\|z\|\|X(\hat{\Sigma}+\lambda I_p)^+\|^2\|\Sigma\|\|\epsilon\|$$
$$= (I).$$

Note that $\lambda_{\min}(\frac{1}{n}XX^T) \geq \lambda_{\min}(\Sigma)\lambda_{\min}(\frac{1}{n}ZZ^T) \geq C_{\min}\lambda_{\min}(\frac{1}{n}ZZ^T)$. The Bai-Yin theorem (Theorem 1 in Bai and Yin 1993) yields that

$$\lambda_{\min}\left(\frac{1}{p}ZZ^T\right) \xrightarrow{a.s.} \left(1-\frac{1}{\sqrt{\gamma}}\right)^2,$$

so that

$$\lambda_{\min}\left(\frac{1}{n}ZZ^T\right) = \frac{p}{n}\lambda_{\min}\left(\frac{1}{p}ZZ^T\right) \xrightarrow{a.s.} \gamma\left(1-\frac{1}{\sqrt{\gamma}}\right)^2 = (1-\sqrt{\gamma})^2.$$

Therefore, for sufficiently large n, we get

$$\lambda_{\min}\left(\frac{1}{n}XX^T\right) \geq C_{\min}(1-\sqrt{\gamma})^2,$$

almost surely. Similar arguments also yield that for sufficiently large n,

$$\lambda_{\max}\left(\frac{1}{n}X^TX\right) = \lambda_{\max}\left(\frac{1}{n}XX^T\right) \leq C_{\max}(1-\sqrt{\gamma})^2,$$

almost surely. Note that

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^+\right\|^2 &= \lambda_{\max}\left(\left(\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^+\right)^T\left(\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^+\right)\right)\\ &= \lambda_{\max}\left(\hat{\Sigma}+\lambda I_p)^+\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^+\right).\end{aligned}$$

Let

$$\hat{\Sigma} = V\operatorname{diag}(\lambda_1,\ldots,\lambda_n,0,\ldots,0)V^T \tag{3.21}$$

be the eigen decomposition of $\hat{\Sigma}$, where $V \in \mathbb{R}^{p\times p}$ is an orthogonal matrix, and $\lambda_1 \geq \ldots \geq \lambda_n > 0$ are decreasing ordered eigenvalues of $\hat{\Sigma}$. Note that $\lambda_1 = \lambda_{\max}(\hat{\Sigma})$, and $\lambda_n = \lambda_{\min}(\frac{1}{n}XX^T)$. Note also that $\hat{\Sigma}^+$ can be expressed as $\hat{\Sigma}^+ = V\operatorname{diag}(\lambda_1^{-1},\ldots,\lambda_n^{-1},0,\ldots,0)V^T$.

Then we can write

$$(\hat{\Sigma}+\lambda I_p)^{+}\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+} = V\mathrm{diag}\left(\frac{\lambda_1}{(\lambda_1+\lambda)^2},\ldots,\frac{\lambda_n}{(\lambda_n+\lambda)^2},0,\ldots,0\right)V^T.$$

Note that for $i=1,\ldots,n$, sufficiently large n, and all $\lambda \in (\lambda_{\min},\infty)$,

$$\frac{\lambda_i}{(\lambda_i+\lambda)^2} \le \frac{\lambda_1}{(\lambda_n+\lambda)^2} \le \frac{C_{\max}(1-\sqrt{\gamma})^2}{(C_{\min}(1-\sqrt{\gamma})^2+\lambda)^2},$$

almost surely. Therefore, for sufficiently large n and all $\lambda \in (\lambda_{\min},\infty)$, we get

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\|^2 &= \lambda_{\max}\left((\hat{\Sigma}+\lambda I_p)^{+}\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+}\right)\\ &\le \frac{C_{\max}(1-\sqrt{\gamma})^2}{(C_{\min}(1-\sqrt{\gamma})^2+\lambda)^2},\end{aligned}$$

almost surely. Note that when $\gamma=1$, $\lambda_{\min}=0$, so that $\lambda>0$. Therefore, for sufficiently large n and all $\lambda \in (\lambda_{\min},\infty)$, we get

$$\begin{aligned}(I) &\le \frac{2C_{\max}^2(1-\sqrt{\gamma})^2}{(C_{\min}(1-\sqrt{\gamma})^2+\lambda)^2}\frac{1}{\sqrt{n}}\|z\|\frac{1}{\sqrt{n}}\|\epsilon\|\\ &\le \frac{C}{\sqrt{n}}\|z\| \xrightarrow{a.s.} 0.\end{aligned}$$

The same arguments in Patil et al. (2021) show that

$$\frac{2}{n}\epsilon^T X(\hat{\Sigma}+\lambda I_p)^{+}\Sigma\{\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+}-I_p\}\beta_0 \xrightarrow{a.s.} 0.$$

Therefore, we get for all $\lambda \in (\lambda_{\min},\infty)$,

$$R_c(\lambda) \xrightarrow{a.s.} 0,$$

and this completes the proof of this lemma. □

Proof of Lemma 2. The same arguments in Patil et al. (2021) show that the first term in $\mathrm{GCV}_c(\lambda)/\mathrm{GCV}_d(\lambda)$ converges to zero almost surely for all $\lambda \in (\lambda_{\min},\infty)$:

$$\frac{\frac{2}{n}\beta_0^T X^T(I_n-S_\lambda)^2\epsilon}{(1-\mathrm{tr}[(\hat{\Sigma}+\lambda I_p)^{+}\hat{\Sigma}]\frac{1}{n})^2} \xrightarrow{a.s.} 0.$$

We evaluate the second term. Consider the case $\lambda \neq 0$. Note that by the Cauchy–Schwarz inequality, we have

$$\frac{2}{n}\left|z^T(I_n - S_\lambda)^2\epsilon\right| \leq \frac{2\lambda^2}{n}\left\|\left(\frac{1}{n}XX^T + \lambda I_n\right)^{+2}\epsilon\right\|\|z\|.$$

Therefore, we have

$$\begin{aligned}
\left|\frac{\frac{2}{n}z^T(I_n - S_\lambda)^2\epsilon}{(1-\mathrm{tr}[(\hat{\Sigma}+\lambda I_p)^+\hat{\Sigma}\frac{1}{n}])^2}\right| &= \frac{2}{n}\left|\frac{z^T(\frac{1}{n}XX^T+\lambda I_n)^{+2}\epsilon}{(\frac{1}{n}\mathrm{tr}[(\frac{1}{n}XX^T+\lambda I_n)^+])^2}\right| \\
&\leq \frac{2}{n}\left(\lambda_{\max}\left(\frac{1}{n}XX^T\right)+\lambda\right)^2\left|z^T\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}\epsilon\right| \\
&\leq \frac{2}{n}\left(\lambda_{\max}\left(\frac{1}{n}XX^T\right)+\lambda\right)^2\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}\epsilon\right\|\|z\| \\
&= (II).
\end{aligned}$$

Note that for sufficiently large n,

$$\begin{aligned}
\frac{1}{n}\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}\epsilon\right\|^2 &= \frac{1}{n}\epsilon^T\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+4}\epsilon \\
&\leq \frac{1}{(\lambda_{\min}\left(\frac{1}{n}XX^T\right)+\lambda)^4}\frac{1}{n}\|\epsilon\|^2 \\
&\leq C,
\end{aligned}$$

almost surely. Furthermore, we have $\lambda_{\max}(\frac{1}{n}XX^T)+\lambda \leq C_{\max}(1-\sqrt{\gamma})^2+\lambda$ for sufficiently large n. Therefore, we have

$$\begin{aligned}
(II) &= \left(\lambda_{\max}\left(\frac{1}{n}XX^T\right)+\lambda\right)^2\left\{\frac{1}{n}\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}\epsilon\right\|^2\right\}^{1/2}\frac{2}{\sqrt{n}}\|z\| \\
&\leq \frac{2(\lambda_{\max}\left(\frac{1}{n}XX^T\right)+\lambda)^2}{(\lambda_{\min}\left(\frac{1}{n}XX^T\right)+\lambda)^2}\frac{1}{\sqrt{n}}\|\epsilon\|\frac{1}{\sqrt{n}}\|z\| \\
&\leq \frac{C}{\sqrt{n}}\|z\| \xrightarrow{a.s.} 0.
\end{aligned}$$

Next, we consider the case $\lambda = 0$. The limit of the second term is given by

$$\lim_{\lambda\to 0}\frac{\frac{2}{n}z^T(I_n - S_\lambda)^2\epsilon}{(1-\mathrm{tr}[(\hat\Sigma+\lambda I_p)^+\hat\Sigma\frac{1}{n}])^2} = \frac{2}{n}\frac{z^T(\frac{1}{n}XX^T)^{+2}\epsilon}{(\frac{1}{n}\mathrm{tr}[(\frac{1}{n}XX^T)^+])^2}.$$

Similar arguments of the case $\lambda \neq 0$ show that

$$\frac{2}{n}\frac{z^T(\frac{1}{n}XX^T)^{+2}\epsilon}{(\frac{1}{n}\mathrm{tr}[(\frac{1}{n}XX^T)^+])^2} \xrightarrow{a.s.} 0.$$

This completes the proof. □

Proof of Lemma 3. Let us write

$$R_b(\lambda) - \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} = (a) + (b) + (c),$$

where

$$\begin{aligned}
(a) &= \beta_0^T\{\hat\Sigma(\hat\Sigma+\lambda I_p)^+ - I_p\}\Sigma\{(\hat\Sigma+\lambda I_p)^+\hat\Sigma - I_p\}\beta_0 - \frac{\frac{1}{n}\beta_0^T X^T(I_n - S_\lambda)^2 X\beta_0}{(1-\mathrm{tr}[(\hat\Sigma+\lambda I_p)^+\hat\Sigma]\frac{1}{n})^2},\\
(b) &= \frac{1}{n}z^T X(\hat\Sigma+\lambda I_p)^+\Sigma(\hat\Sigma+\lambda I_p)^+\frac{1}{n}X^T z - \frac{\frac{1}{n}z^T(I_n-S_\lambda)^2 z}{(1-\mathrm{tr}[(\hat\Sigma+\lambda I_p)^+\hat\Sigma]\frac{1}{n})^2} + E[(f(x_0) - x_0^T\beta_0)^2],\\
(c) &= 2\beta_0^T\{\hat\Sigma(\hat\Sigma+\lambda I_p)^+ - I_p\}\Sigma(\hat\Sigma+\lambda I_p)^+\frac{1}{n}X^T z - \frac{\frac{2}{n}z^T(I_n-S_\lambda)^2 X\beta_0}{(1-\mathrm{tr}[(\hat\Sigma+\lambda I_p)^+\hat\Sigma]\frac{1}{n})^2}.
\end{aligned}$$

We will evaluate each term. First, by the same arguments of Patil et al. (2021), we can show that (a) converges to zero almost surely for all $\lambda \in (\lambda_{\min}, \infty)$:

$$\begin{aligned}
(a) &= \beta_0^T\{\hat\Sigma(\hat\Sigma+\lambda I_p)^+ - I_p\}\Sigma\{(\hat\Sigma+\lambda I_p)^+\hat\Sigma - I_p\}\beta_0\\
&\quad - \frac{\beta_0^T\{\hat\Sigma(\hat\Sigma+\lambda I_p)^+ - I_p\}\hat\Sigma\{\hat\Sigma(\hat\Sigma+\lambda I_p)^+ - I_p\}\beta_0}{(1-\mathrm{tr}[(\hat\Sigma+\lambda I_p)^+\hat\Sigma]\frac{1}{n})^2} \xrightarrow{a.s.} 0.
\end{aligned}$$

Next, we evaluate (b). First, we focus on the first term in (b). Note that $X(\hat\Sigma+\lambda I_p)^+\Sigma(\hat\Sigma+\lambda I_p)^+X^T$ is positive semidefinite. Therefore, by the Cauchy–Schwarz inequality, we get for all $\lambda \in (\lambda_{\min}, \infty)$

$$0 \le \frac{1}{n}z^T X(\hat\Sigma+\lambda I_p)^+\Sigma(\hat\Sigma+\lambda I_p)^+\frac{1}{n}X^T z$$

$$
\begin{aligned}
&\leq \|\Sigma\| \frac{1}{n} z^T X(\hat{\Sigma} + \lambda I_p)^{+2} X^T \frac{1}{n} z \\
&\leq \|\Sigma\| \lambda_{\max} \left(\frac{1}{n} X(\hat{\Sigma} + \lambda I_p)^{+2} X^T \right) \frac{1}{n} \|z\|^2 \\
&= (III).
\end{aligned}
$$

Let

$$
\frac{1}{\sqrt{n}} X = U Q_0 V^T
$$

be the singular value decomposition of $\frac{1}{\sqrt{n}}X$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and $Q_0 \in \mathbb{R}^{n\times p}$ is the matrix whose jth diagonal element is $\sqrt{\lambda_j}$ for $j = 1, \ldots, n$, and all non-diagonal elements are zero. Recall that λ_j, $j = 1, \ldots, n$ are eigenvalues of $\hat{\Sigma}$, and V is the same as in (3.21). Then we can write

$$
\frac{1}{n} X(\hat{\Sigma} + \lambda I_p)^{+2} X^T = U \mathrm{diag} \left(\frac{\lambda_1}{(\lambda_1 + \lambda)^2}, \ldots, \frac{\lambda_n}{(\lambda_n + \lambda)^2} \right) U^T.
$$

Thus, for sufficiently large n, we get

$$
\lambda_{\max} \left(\frac{1}{n} X(\hat{\Sigma} + \lambda I_p)^{+2} X^T \right) \leq \frac{C_{\max}(1 - \sqrt{\gamma})^2}{(C_{\min}(1 - \sqrt{\gamma})^2 + \lambda)^2},
$$

almost surely. Therefore, we get

$$
(III) \leq \frac{C_{\max}^2(1 - \sqrt{\gamma})^2}{(C_{\min}(1 - \sqrt{\gamma})^2 + \lambda)^2} \frac{1}{n} \|z\|^2 \xrightarrow{a.s.} 0.
$$

Next, we evaluate the second term in (b). We first consider the case $\lambda \neq 0$. Note that

$$
\begin{aligned}
\left(I_n - X(\hat{\Sigma} + \lambda I_p)^{+} \frac{1}{n} X^T \right)^2 &= \lambda^2 \left(\frac{1}{n} X X^T + \lambda I_n \right)^{+2}, \\
\left(1 - \mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^{+} \hat{\Sigma}] \frac{1}{n} \right)^2 &= \lambda^2 \left(\frac{1}{n} \mathrm{tr} \left[\left(\frac{1}{n} X X^T + \lambda I_n \right)^{+} \right] \right)^2.
\end{aligned}
$$

We also have

$$
\left(\frac{1}{n} \mathrm{tr} \left[\left(\frac{1}{n} X X^T + \lambda I_n \right)^{+} \right] \right)^2 \geq \frac{1}{(\lambda_{\max}(\frac{1}{n} X X^T) + \lambda)^2},
$$

$$\frac{1}{n} z^T \left(\frac{1}{n} XX^T + \lambda I_n \right)^{+2} z \le \frac{1}{(\lambda_{\min}\left(\frac{1}{n}XX^T\right) + \lambda)^2} \frac{1}{n} \|z\|^2.$$

Hence, we get for all $\lambda \in (\lambda_{\min}, \infty) \cap \{0\}^c$,

$$\begin{aligned}
\frac{1}{n} \frac{z^T (I_n - X(\hat{\Sigma} + \lambda I_p)^+ \frac{1}{n} X^T)^2 z}{(1 - \mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^+ \hat{\Sigma}] \frac{1}{n})^2} &= \frac{1}{n} \frac{z^T \left(\frac{1}{n}XX^T + \lambda I_n\right)^{+2} z}{\left(\frac{1}{n}\mathrm{tr}\left[(\frac{1}{n}XX^T + \lambda I_n)^+\right]\right)^2} \\
&\le \frac{(\lambda_{\max}\left(\frac{1}{n}XX^T\right) + \lambda)^2}{(\lambda_{\min}(\frac{1}{n}XX^T) + \lambda)^2} \frac{1}{n} \|z\|^2 \\
&\le \frac{C}{n} \|z\|^2 \xrightarrow{a.s.} 0.
\end{aligned}$$

Thus, when $\lambda \neq 0$, the second term in (b) also converges to zero almost surely. Next, we consider the case $\lambda = 0$. The limit of the second term in (b) is given by

$$\lim_{\lambda \to 0} \frac{1}{n} \frac{z^T (I_n - X(\hat{\Sigma} + \lambda I_p)^+ \frac{1}{n} X^T)^2 z}{(1 - \mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^+ \hat{\Sigma}] \frac{1}{n})^2} = \frac{1}{n} \frac{z^T \left(\frac{1}{n}XX^T\right)^{+2} z}{\left(\frac{1}{n}\mathrm{tr}\left[(\frac{1}{n}XX^T)^+\right]\right)^2}.$$

Similar arguments of the case $\lambda \neq 0$ for the second term in (b) show that

$$\frac{1}{n} \frac{z^T \left(\frac{1}{n}XX^T\right)^{+2} z}{\left(\frac{1}{n}\mathrm{tr}\left[(\frac{1}{n}XX^T)^+\right]\right)^2} \xrightarrow{a.s.} 0.$$

Thus, the second term in (b) goes to zero for all $\lambda \in (\lambda_{\min}, \infty)$. The third term in (b) goes to zero by Assumption 4. Now we have shown that (b) goes to zero almost surely for all $\lambda \in (\lambda_{\min}, \infty)$.

Next, we evaluate (c). We focus on the first term. By the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
\frac{2}{n} \left| \beta_0^T \{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\} \Sigma (\hat{\Sigma} + \lambda I_p)^+ X^T z \right| &\le \frac{2}{n} \left\| X(\hat{\Sigma} + \lambda I_p)^+ \Sigma \{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\} \beta_0 \right\| \|z\| \\
&= (IV).
\end{aligned}$$

Notice that

$$\frac{1}{\sqrt{n}} \left\| X(\hat{\Sigma} + \lambda I_p)^+ \Sigma \{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\} \beta_0 \right\| \le \left\| \frac{1}{\sqrt{n}} X(\hat{\Sigma} + \lambda I_p)^+ \right\| \|\Sigma\| \|\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\| \|\beta_0\|.$$

Note that

$$\|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^+ - I_p\| \leq \|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^+\| + 1,$$

and for sufficiently large n,

$$\begin{aligned}
\|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^+\| &= \sqrt{\lambda_{\max}((\hat{\Sigma}+\lambda I_p)^+\hat{\Sigma}^2(\hat{\Sigma}+\lambda I_p)^+)} \\
&\leq \frac{C_{\max}(1-\sqrt{\gamma})^2}{C_{\min}(1-\sqrt{\gamma})^2+\lambda},
\end{aligned}$$

almost surely. Therefore, for sufficiently large n, we get

$$(IV) \leq \frac{C_{\max}^{3/2}|1-\sqrt{\gamma}|}{C_{\min}(1-\sqrt{\gamma})^2+\lambda}\left(\frac{C_{\max}(1-\sqrt{\gamma})^2}{C_{\min}(1-\sqrt{\gamma})^2+\lambda}+1\right)\frac{1}{\sqrt{n}}\|z\| \xrightarrow{a.s.} 0.$$

So the first term in (c) goes to zero almost surely for all $\lambda \in (\lambda_{\min}, \infty)$. Now we evaluate the second term in (c). Consider the case $\lambda \neq 0$. For the numerator, the Cauchy–Schwarz inequality yields

$$\begin{aligned}
\frac{2}{n}|z^T(I_n - S_\lambda)^2 X\beta_0| &= \frac{2\lambda^2}{n}\left|z^T\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}X\beta_0\right| \\
&\leq \frac{2\lambda^2}{n}\|z\|\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}X\beta_0\right\| \\
&= (V).
\end{aligned}$$

Note that for sufficiently large n, we get

$$\begin{aligned}
\frac{1}{n}\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+2}X\beta_0\right\|^2 &= \frac{1}{n}\beta_0^T X^T\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+4}X\beta_0 \\
&\leq \left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+4}\right\|\frac{1}{n}\beta_0^T XX^T\beta_0 \\
&\leq \|\hat{\Sigma}\|\left\|\left(\frac{1}{n}XX^T+\lambda I_n\right)^{+}\right\|^4\|\beta_0\|^2 \\
&\leq \frac{\|\hat{\Sigma}\|}{(\lambda_{\min}(\frac{1}{n}XX^T)+\lambda)^4}\|\beta_0\|^2
\end{aligned}$$

$$\leq C,$$

almost surely. Therefore, we obtain

$$(V) \leq \frac{C\lambda^2}{\sqrt{n}} \|z\|,$$

almost surely for sufficiently large n, and

$$\begin{aligned}
\left| \frac{\frac{2}{n} z^T (I_n - S_\lambda)^2 X \beta_0}{(1 - \mathrm{tr}[(\hat{\Sigma} + \lambda I_p)^+ \hat{\Sigma}] \frac{1}{n})^2} \right| &= \left| \frac{\frac{2}{n} z^T (\frac{1}{n} X X^T + \lambda I_n)^{+2} X \beta_0}{\left(\frac{1}{n} \mathrm{tr}\left[(\frac{1}{n} X X^T + \lambda I_n)^+ \right] \right)^2} \right| \\
&\leq \left(\lambda_{\max} \left(\frac{1}{n} X X^T \right) + \lambda \right)^2 \frac{C\lambda^2}{\sqrt{n}} \|z\| \\
&\leq \frac{C}{\sqrt{n}} \|z\| \xrightarrow{a.s.} 0,
\end{aligned}$$

for all $\lambda \in (\lambda_{\min}, \infty) \cap \{0\}^c$. Next, consider the case $\lambda = 0$. The limit of the second term in (c) is given by

$$\lim_{\lambda \to 0} \frac{\frac{2}{n} z^T (\frac{1}{n} X X^T + \lambda I_n)^{+2} X \beta_0}{\left(\frac{1}{n} \mathrm{tr}\left[(\frac{1}{n} X X^T + \lambda I_n)^+ \right] \right)^2} = \frac{\frac{2}{n} z^T (\frac{1}{n} X X^T)^{+2} X \beta_0}{\left(\frac{1}{n} \mathrm{tr}\left[(\frac{1}{n} X X^T)^+ \right] \right)^2}.$$

The similar arguments of the case $\lambda \neq 0$ for the second term in (c) show that

$$\frac{\frac{2}{n} z^T (\frac{1}{n} X X^T)^{+2} X \beta_0}{\left(\frac{1}{n} \mathrm{tr}\left[(\frac{1}{n} X X^T)^+ \right] \right)^2} \xrightarrow{a.s.} 0.$$

This completes the proof. □

Proof of Theorem 1. By Lemmas 1–4, we get

$$\begin{aligned}
|R(\lambda) - \mathrm{GCV}(\lambda)| &= \left| R_b(\lambda) + R_v(\lambda) + R_c(\lambda) - \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} - \frac{\mathrm{GCV}_v(\lambda)}{\mathrm{GCV}_d(\lambda)} - \frac{\mathrm{GCV}_c(\lambda)}{\mathrm{GCV}_d(\lambda)} \right| \\
&\leq \left| R_b(\lambda) - \frac{\mathrm{GCV}_b(\lambda)}{\mathrm{GCV}_d(\lambda)} \right| + \left| R_v(\lambda) - \frac{\mathrm{GCV}_v(\lambda)}{\mathrm{GCV}_d(\lambda)} \right| + |R_c(\lambda)| + \left| \frac{\mathrm{GCV}_c(\lambda)}{\mathrm{GCV}_d(\lambda)} \right| \\
&\xrightarrow{a.s.} 0,
\end{aligned}$$

for all $\lambda \in (\lambda_{\min}, \infty)$. Therefore, the pointwise convergence holds on $(\lambda_{\min}, \infty)$. To show the uniform convergence, it is sufficient to show that $R(\lambda)$, $\mathrm{GCV}(\lambda)$, and their derivatives are bounded on any closed interval $\Lambda \subset (\lambda_{\min}, \infty)$ for sufficiently large n. First, we evaluate $\mathrm{GCV}(\lambda)$. We can write

$$
\begin{aligned}
\frac{1}{n}\|Y\|^2 &= \frac{1}{n}\|\mu(X)\|^2 + \frac{2}{n}\mu(X)^T\epsilon + \frac{1}{n}\|\epsilon\|^2, \\
\frac{1}{n}\|\mu(X)\|^2 &= \frac{1}{n}\sum_{i=1}^n f(x_i)^2 \xrightarrow{a.s.} E[(f(x_i))^2] \leq C, \\
\frac{1}{n}\mu(X)^T\epsilon &= \frac{1}{n}\sum_{i=1}^n f(x_i)\epsilon_i \xrightarrow{a.s.} E[f(x_i)\epsilon_i] = 0.
\end{aligned}
$$

Therefore, for sufficiently large n, we get

$$
|\mathrm{GCV}(\lambda)| \leq \frac{(\lambda_{\max}(\frac{1}{n}X^TX) + \lambda)^2}{(\lambda_{\min}(\frac{1}{n}XX^T) + \lambda)^2}\frac{1}{n}\|Y\|^2 \leq C,
$$

almost surely for all $\lambda \in \Lambda$. Next, we evaluate the derivative of $\mathrm{GCV}(\lambda)$. As in Patil et al. (2021), let us define

$$
\begin{aligned}
u_n(\lambda) &= \frac{1}{n}Y^T\left(\frac{1}{n}XX^T + \lambda I_n\right)^{+2} Y, \\
v_n(\lambda) &= \left(\mathrm{tr}\left[\left(\frac{1}{n}XX^T + \lambda I_n\right)^{+}\frac{1}{n}\right]\right)^2.
\end{aligned}
$$

Then, we can write

$$
\mathrm{GCV}(\lambda) = \frac{u_n(\lambda)}{v_n(\lambda)}.
$$

The derivative with respect to λ yields

$$
\mathrm{GCV}'(\lambda) = \frac{u_n'(\lambda)v_n(\lambda) - u_n(\lambda)v_n'(\lambda)}{v_n(\lambda)^2}.
$$

Simple calculation yields

$$
\begin{aligned}
|u_n'(\lambda)| &\leq \frac{2}{n}\frac{\|Y\|^2}{\left(\lambda_{\min}\left(\frac{1}{n}XX^T\right) + \lambda\right)^3}, \\
|v_n'(\lambda)| &= \frac{2}{n^2}\left|\sum_{k=1}^n \frac{1}{\lambda_k + \lambda}\sum_{k=1}^n \frac{1}{(\lambda_k + \lambda)^2}\right| \leq \frac{2}{\left(\lambda_{\min}\left(\frac{1}{n}XX^T\right) + \lambda\right)^3}.
\end{aligned}
$$

Therefore, we get for sufficiently large n and all $\lambda \in \Lambda$,

$$\begin{aligned}
|\mathrm{GCV}(\lambda)'| &\le \frac{|u_n'(\lambda)||v_n(\lambda)| + |u_n(\lambda)||v_n(\lambda)|}{v_n'(\lambda)^2} \\
&\le \frac{\left(\lambda_{\max}(\frac{1}{n}XX^T) + \lambda\right)^2}{(\lambda_{\min}(\frac{1}{n}XX^T) + \lambda)^5} \frac{4}{n}\|Y\|^2 \\
&\le C
\end{aligned}$$

almost surely.

Now, we evaluate $R(\lambda)$ and $R(\lambda)'$. Similar arguments of above proofs show that $|R(\lambda)| \le C$ almost surely for all $\lambda \in \Lambda$. Note that $R(\lambda)'$ is given by

$$R(\lambda)' = R_b(\lambda)' + R_v(\lambda)' + R_c(\lambda)'.$$

We first evaluate $R_b(\lambda)'$ when $\lambda \neq 0$. It is given by

$$\begin{aligned}
R_b(\lambda)' =& -2\beta_0^T\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0 - \frac{2}{n}z^T X(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0 \\
&- \frac{2}{n}z^T X(\hat{\Sigma} + \lambda I_p)^{+}\Sigma\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^{+2}\beta_0 - \frac{2}{n}z^T X(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma(\hat{\Sigma} + \lambda I_p)^{+}\frac{1}{n}X^T z.
\end{aligned}$$

Note that $\frac{d}{d\lambda}(\hat{\Sigma} + \lambda I_p)^+ = -(\hat{\Sigma} + \lambda I_p)^{+2}$ for $\lambda \in \Lambda \cap \{0\}^c$. We evaluate the first term in $R_b(\lambda)'$. By the Cauchy–Scwarz inequality, we get for sufficiently large n

$$\begin{aligned}
2|\beta_0^T\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0| &\le 2\|\beta_0\|\|\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0\| \\
&\le 2\|\beta_0\|^2\|\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^{+2}\|\|\Sigma\|\|\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\| \\
&\le C.
\end{aligned}$$

almost surely. Thus, the first term is bounded. Next, for the second term,

$$\begin{aligned}
\frac{2}{n}|z^T X(\hat{\Sigma} + \lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\}\beta_0| &\le \frac{2}{\sqrt{n}}\|z\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma} + \lambda I_p)^{+2}\right\|\|\hat{\Sigma}(\hat{\Sigma} + \lambda I_p)^+ - I_p\|\|\beta_0\| \\
&\le \frac{C}{\sqrt{n}}\|z\| \\
&\le C
\end{aligned}$$

almost surely for sufficiently large n. Thus, the second term is bounded. Next, for the third term, the Cauchy–Scwarz inequality yields that for sufficiently large n

$$\begin{aligned}
\frac{2}{n}|z^T X(\hat{\Sigma}+\lambda I_p)^{+}\Sigma\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\beta_0| &\leq \frac{2}{n}\|z\|\|X(\hat{\Sigma}+\lambda I_p)^{+}\Sigma\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\beta_0\| \\
&\leq \frac{2}{\sqrt{n}}\|z\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\|\|\Sigma\|\|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\|\|\beta_0\| \\
&\leq \frac{C}{\sqrt{n}}\|z\| \\
&\leq C,
\end{aligned}$$

almost surely. Thus, the third term is bounded. Finally, for the forth term, we get

$$\begin{aligned}
\frac{1}{n^2}|z^T X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma(\hat{\Sigma}+\lambda I_p)^{+}X^T z| &\leq \frac{1}{n^2}\|z\|\|X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma(\hat{\Sigma}+\lambda I_p)^{+}X^T z\| \\
&\leq \frac{1}{n}\|z\|^2\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\|^2\|\Sigma\| \\
&\leq C
\end{aligned}$$

almost surely for sufficiently large n. Thus, the fourth term is bounded. Thus, $R_b(\lambda)'$ is bounded for all $\lambda \in \Lambda \cap \{0\}^c$. Next, we evaluate $R_b(0)'$, which we define as the limit. Namely, $R_b(0)' := \lim_{\lambda\to 0} R_b(\lambda)'$. It is given by

$$\begin{aligned}
\lim_{\lambda\to 0} R_b(\lambda)' =& -2\beta_0^T\hat{\Sigma}^{+}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+}-I_p\}\beta_0 - \frac{2}{\sqrt{n}}z^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+}-I_p\}\beta_0 \\
&-\frac{2}{\sqrt{n}}z^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\}\Sigma\hat{\Sigma}^{+}\beta_0 \\
&-\frac{2}{n}z^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\}\Sigma\left\{\lim_{\lambda\to 0}(\hat{\Sigma}+\lambda I_p)^{+}\frac{1}{\sqrt{n}}X^T\right\}z.
\end{aligned}$$

For $k=1,2$, we can write

$$\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+k} = \frac{1}{\sqrt{n}}X\hat{\Sigma}^{+k} = UQ_kV^T,$$

where $Q_k \in \mathbb{R}^{n\times p}$ is the matrix whose jth diagonal element is $\lambda_j^{(1-2k)/2}$ for $j=1,\ldots,n$, and all non-diagonal elements are zero. We can easily show that the first term is bounded almost surely because $\|\beta_0\|$, $\|\hat{\Sigma}^{+}\|$, $\|\Sigma\|$, and $\|\hat{\Sigma}\hat{\Sigma}^{+}-I_p\|$ are bounded. By the Cauchy–

Scwarz inequality,

$$\left|\frac{2}{\sqrt{n}} z^T \frac{1}{\sqrt{n}} X\hat{\Sigma}^{+2}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+} - I_p\}\beta_0\right| \leq \frac{2}{\sqrt{n}}\|z\| \left\|\frac{1}{\sqrt{n}} X\hat{\Sigma}^{+2}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+} - I_p\}\beta_0\right\|,$$

and

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}} X\hat{\Sigma}^{+2}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+} - I_p\}\beta_0\right\|^2 &= \beta_0^T\{\hat{\Sigma}\hat{\Sigma}^{+} - I_p\}\Sigma V Q_2^T Q_2 V^T \Sigma\{\hat{\Sigma}\hat{\Sigma}^{+} - I_p\}\beta_0 \\ &\leq \lambda_{\max}\left(Q_2^T Q_2\right)\|\Sigma\|^2\|\hat{\Sigma}\hat{\Sigma}^{+} - I_p\|^2\|\beta_0\|^2 \\ &\leq C,\end{aligned}$$

almost surely for sufficiently large n. Note that $\lambda_{\max}\left(Q_2^T Q_2\right) = \lambda_n^{-3} \leq (C_{\min}(1-\sqrt{\gamma})^2)^{-3}$ almost surely for sufficiently large n. This implies that the second term is bounded. Similarly,

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}} X\hat{\Sigma}^{+}\Sigma\hat{\Sigma}^{+}\beta_0\right\|^2 &= \beta_0^T\hat{\Sigma}^{+}\Sigma V Q_1^T Q_1 V^T \Sigma\hat{\Sigma}^{+}\beta_0 \\ &\leq \lambda_{\max}\left(Q_1^T Q_1\right)\|\Sigma\|^2\|\hat{\Sigma}^{+}\|^2\|\beta_0\|^2 \\ &\leq C\end{aligned}$$

almost surely for sufficiently large n. Note that $\lambda_{\max}(Q_1^T Q_1) = \lambda_n^{-1} \leq (C_{\min}(1-\sqrt{\gamma})^2)^{-1}$ for sufficiently large n. Then the Cauchy–Scwarz inequality implies that the third term is bounded. Finally, for sufficiently large n,

$$\begin{aligned}\frac{1}{n}\left\|\frac{1}{\sqrt{n}} X\hat{\Sigma}^{+2}\Sigma\hat{\Sigma}^{+}\frac{1}{\sqrt{n}} X^T z\right\|^2 &= \frac{1}{n} z^T V Q_1^T U^T \Sigma V Q_2^T Q_2 V^T \Sigma U Q_1 V^T z \\ &\leq \lambda_{\max}\left(Q_2^T Q_2\right)\|\Sigma\|^2 \lambda_{\max}\left(Q_1^T Q_1\right)\frac{1}{n}\|z\|^2 \\ &\leq C,\end{aligned}$$

almost surely. Then the Cauchy–Scwarz inequality implies that the fourth term is bounded. Now we have shown that $R_b(0)'$ is bounded. Thus, we have shown that $R_b(\lambda)'$ is bounded for all $\lambda \in \Lambda$.

Next, we evaluate $R_v(\lambda)'$ when $\lambda \neq 0$. It is given by

$$R_v(\lambda)' = -\frac{2}{n}\epsilon^T X(\hat{\Sigma} + \lambda I_p)^{+}\Sigma(\hat{\Sigma} + \lambda I_p)^{+2}\frac{1}{n} X^T \epsilon.$$

The Cauchy–Scwarz inequality shows that for sufficiently large n,

$$\begin{aligned}\frac{1}{n^2}|\epsilon^T X(\hat{\Sigma}+\lambda I_p)^+\Sigma(\hat{\Sigma}+\lambda I_p)^{+2}X^T\epsilon| &\leq \frac{1}{n^2}\|\epsilon\|\|X(\hat{\Sigma}+\lambda I_p)^+\Sigma(\hat{\Sigma}+\lambda I_p)^{+2}X^T\epsilon\| \\ &\leq \frac{1}{n}\|\epsilon\|^2\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^+\right\|^2\|\Sigma\| \\ &\leq C\end{aligned}$$

almost surely. Thus, $R_v(\lambda)'$ is bounded for all $\lambda\in\Lambda\cap\{0\}^c$. Next, we consider the case when $\lambda=0$. We define $R_v(0)'$ as the limit: $R_v(0)' := \lim_{\lambda\to 0}R_v(\lambda)'$. It is given by

$$\begin{aligned}\lim_{\lambda\to 0}R_v(\lambda)' &= -\frac{2}{n}\epsilon^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^+\right\}\Sigma\left\{\lim_{\lambda\to 0}(\hat{\Sigma}+\lambda I_p)^{+2}\frac{1}{\sqrt{n}}X^T\right\}\epsilon \\ &= -\frac{2}{n}\epsilon^T UQ_1V^T\Sigma VQ_2U^T\epsilon.\end{aligned}$$

The Cauchy–Scwarz inequality shows that

$$\frac{1}{n}|\epsilon^T UQ_1V^T\Sigma VQ_2U^T\epsilon| \leq \frac{1}{n}\|\epsilon\|\|UQ_1V^T\Sigma VQ_2U^T\epsilon\|.$$

Note that for sufficiently large n,

$$\begin{aligned}\frac{1}{n}\|UQ_1V^T\Sigma VQ_2U^T\epsilon\|^2 &\leq \lambda_{\max}\left(Q_1^TQ_1\right)\lambda_{\max}(\Sigma)^2\lambda_{\max}\left(Q_2^TQ_2\right)\frac{1}{n}\|\epsilon\|^2 \\ &\leq C,\end{aligned}$$

almost surely. This implies that for sufficiently large n,

$$\frac{1}{n}\|\epsilon\|\|UQ_1V^T\Sigma VQ_2U^T\epsilon\| \leq C,$$

almost surely. Thus, $R_v(\lambda)'$ is bounded for all $\lambda\in\Lambda$.

Next, we evaluate $R_c(\lambda)'$ when $\lambda\neq 0$. $R_c(\lambda)'$ is given by

$$\begin{aligned}R_c(\lambda)' =& -\frac{2}{n}z^TX(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma(\hat{\Sigma}+\lambda I_p)^+\frac{1}{n}X^T\epsilon - \frac{2}{n}z^TX(\hat{\Sigma}+\lambda I_p)^+\Sigma(\hat{\Sigma}+\lambda I_p)^{+2}\frac{1}{n}X^T\epsilon \\ &- \frac{2}{n}\epsilon^TX(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^+ - I_p\}\beta_0 - \frac{2}{n}\epsilon^TX(\hat{\Sigma}+\lambda I_p)^+\Sigma\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\beta_0.\end{aligned}$$

By the Cauchy–Scwarz inequality,

$$
\begin{aligned}
&\frac{2}{n^2}|z^T X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma(\hat{\Sigma}+\lambda I_p)^{+}X^T\epsilon| \\
&\leq \frac{2}{n^2}\|z\|\|X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma(\hat{\Sigma}+\lambda I_p)^{+}X^T\epsilon\| \\
&\leq \frac{2}{\sqrt{n}}\|z\|\frac{1}{\sqrt{n}}\|\epsilon\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\|\|\Sigma\| \\
&\leq C
\end{aligned}
$$

almost surely for sufficiently large n. Similar arguments also show that the second term in $R_c(\lambda)'$ is bounded. We also get for sufficiently large n,

$$
\begin{aligned}
&\frac{2}{n}|\epsilon^T X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+}-I_p\}\beta_0| \\
&\leq \frac{2}{n}\|\epsilon\|\|X(\hat{\Sigma}+\lambda I_p)^{+2}\Sigma\{\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+}-I_p\}\beta_0\| \\
&\leq \frac{2}{\sqrt{n}}\|\epsilon\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\|\|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+}-I_p\|\|\beta_0\| \\
&\leq C,
\end{aligned}
$$

almost surely. We finally get for sufficiently large n,

$$
\begin{aligned}
\frac{2}{n}|\epsilon^T X(\hat{\Sigma}+\lambda I_p)^{+}\Sigma\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\beta_0| &\leq \frac{1}{\sqrt{n}}\|\epsilon\|\left\|\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\|\|\Sigma\|\|\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\|\|\beta_0\| \\
&\leq C,
\end{aligned}
$$

almost surely. Thus, $R_c(\lambda)'$ is bounded for all $\lambda \in \Lambda \cap \{0\}^c$. Finally, we consider the case $\lambda = 0$. We define $R_c(0)'$ as the limit: $R_c(0)' := \lim_{\lambda\to 0} R_c(\lambda)$. It is given by

$$
\begin{aligned}
\lim_{\lambda\to 0} R_c(\lambda)' = &-\frac{2}{n}z^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\}\Sigma\left\{\lim_{\lambda\to 0}(\hat{\Sigma}+\lambda I_p)^{+}\frac{1}{\sqrt{n}}X^T\right\}\epsilon \\
&-\frac{2}{n}z^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\}\Sigma\left\{\lim_{\lambda\to 0}(\hat{\Sigma}+\lambda I_p)^{+2}\frac{1}{\sqrt{n}}X^T\right\}\epsilon \\
&-\frac{2}{\sqrt{n}}\epsilon^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+2}\right\}\Sigma\{\hat{\Sigma}\hat{\Sigma}^{+}-I_p\}\beta_0 \\
&-\frac{2}{\sqrt{n}}\epsilon^T\left\{\lim_{\lambda\to 0}\frac{1}{\sqrt{n}}X(\hat{\Sigma}+\lambda I_p)^{+}\right\}\Sigma\left\{\lim_{\lambda\to 0}\hat{\Sigma}(\hat{\Sigma}+\lambda I_p)^{+2}\right\}\beta_0.
\end{aligned}
$$

We get for sufficiently large n,

$$\begin{aligned}\frac{2}{n}|z^T U Q_2 V^T \Sigma V Q_1 U^T \epsilon| &\leq \frac{1}{n}\|z\|\|U Q_2 V^T \Sigma V Q_1^T U^T \epsilon\| \\ &\leq \frac{1}{n}\|z\|\|Q_2\|\|\Sigma\|\|Q_1\|\|\epsilon\| \\ &\leq C,\end{aligned}$$

almost surely. Thus the first term in $R_c(0)'$ is bounded. We also get for sufficiently large n,

$$\begin{aligned}\frac{2}{n}|z^T U Q_1 V^T \Sigma V Q_2^T U^T \epsilon| &\leq \frac{1}{n}\|z\|\|Q_1\|\|Q_2\|\|\Sigma\|\|\epsilon\| \\ &\leq C,\end{aligned}$$

almost surely. Thus the second term in $R_c(0)'$ is bounded. We also get for sufficiently large n,

$$\begin{aligned}\frac{1}{\sqrt{n}}|\epsilon^T U Q_2 V^T \Sigma \{\hat{\Sigma}\hat{\Sigma}^+ - I_p\}\beta_0| &\leq \frac{1}{\sqrt{n}}\|\epsilon\|\|Q_2\|\|\Sigma\|\|\hat{\Sigma}\hat{\Sigma}^+ - I_p\|\|\beta_0\| \\ &\leq C,\end{aligned}$$

almost surely. Thus the third term in $R_c(0)'$ is bounded. We finally get for sufficiently large n,

$$\begin{aligned}\frac{1}{\sqrt{n}}|\epsilon^T U Q_1 V^T \Sigma \hat{\Sigma}^+ \beta_0| &\leq \frac{1}{\sqrt{n}}\|\epsilon\|\|Q_1\|\|\Sigma\|\|\hat{\Sigma}^+\|\|\beta_0\| \\ &\leq C,\end{aligned}$$

almost surely. Thus the fourth term in $R_c(0)'$ is bounded. Therefore, $R_c(0)'$ is bounded. This implies that $R_c(\lambda)'$ is bounded for all $\lambda \in \Lambda$. Thus, $R(\lambda)'$ is bounded over Λ.

Now we have shown that $\mathrm{GCV}(\lambda) - R(\lambda)$ and its derivative with respect to λ are bounded over Λ. This implies that $\mathrm{GCV}(\lambda) - R(\lambda)$ is equicontinous over Λ. Then the Arzela–Ascoli theorem implies the uniform convergence on Λ. $\square$

Proof of Theorem 2. Notice that $R(\bar{\lambda}) \geq R(\lambda^*)$ by the definition of λ^*. By Theorem 1, we get

$$R(\bar{\lambda}) - R(\lambda^*) = R(\bar{\lambda}) - \mathrm{GCV}(\bar{\lambda}) + \mathrm{GCV}(\bar{\lambda}) - \mathrm{GCV}(\lambda^*) - (\mathrm{GCV}(\lambda^*) - R(\lambda^*))$$

$$\leq R(\bar{\lambda}) - \mathrm{GCV}(\bar{\lambda}) - (\mathrm{GCV}(\lambda^*) - R(\lambda^*))$$
$$\leq 2 \sup_{\lambda \in \Lambda} |R(\lambda) - \mathrm{GCV}(\lambda)| \xrightarrow{a.s} 0.$$

This completes the first part of the proof. By Lemma 5, we get

$$R(\lambda) - \mathrm{CV}(\lambda) \xrightarrow{a.s.} 0$$

uniformly on $\lambda \in \Lambda$. Therefore, the same argument of GCV also yields that

$$R(\hat{\lambda}) - R(\lambda^*) \xrightarrow{a.s.} 0.$$

This completes the second part of the proof. $\square$

References

[1] Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.

[2] Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

[3] Bai, Z.-D. and Yin, Y.-Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21:1275--1294.

[4] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117:30063–30070.

[5] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849–15854.

[6] Bellec, P. C. and Zhang, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28:713–743.

[7] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429.

[8] Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81:608–650.

[9] Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41:802–837.

[10] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732.

[11] Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25:173–187.

[12] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory, and applications*. Springer, New York.

[13] Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of statistics*, 45:615–646.

[14] Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771.

[15] Cheng, M.-Y., Feng, S., Li, G., and Lian, H. (2018). Greedy forward regression for variable screening. *Australian & New Zealand Journal of Statistics*, 60:20–42.

[16] Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data. *Annals of Statistics*, 42:1819–1849.

[17] Cheng, M.-Y., Honda, T., and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 111:1209–1221.

[18] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.

[19] Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated lasso in high dimensions. *Annals of Statistics*, 49:1300–1317.

[20] Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30:533–558.

[21] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

[22] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

[23] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

[24] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:849–911.

[25] 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.

[26] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

[27] Gilley, O. W. and Pace, R. K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31:403–405.

[28] Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223.

[29] Greene, W. H. (2012). *Econometric Analysis 7th ed.* Pearson Education, Harlow.

[30] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

[31] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50:949–986.

[32] Honda, T., Ing, C.-K., and Wu, W.-Y. (2019). Adaptively weighted group lasso for semiparametric quantile regression models. *Bernoulli*, 25:3311–3338.

[33] Honda, T. and Lin, C.-T. (2021). Forward variable selection for sparse ultra-high-dimensional generalized varying coefficient models. *Japanese Journal of Statistics and Data Science*, 4:151–179.

[34] Honda, T. and Yabe, R. (2017). Variable selection and structure identification for varying coefficient cox models. *Journal of Multivariate Analysis*, 161:103–122.

[35] Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis 2nd ed.* Cambridge University Press.

[36] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12:115–121.

[37] Jankova, J. and van de Geer, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *Annals of Statistics*, 46:2336–2359.

[38] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909.

[39] Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics*, 46:2593–2622.

[40] Josse, J. and Husson, F. (2016). missmda: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70:1–31.

[41] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

[42] Kobak, D., Lomond, J., and Sanchez, B. (2020). The optimal ridge penalty for real-world

high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21:1–16.

[43] Kozbur, D. (2020). Analysis of testing-based forward model selection. *Econometrica*, 88:2147–2173.

[44] Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109:216–229.

[45] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44:907–927.

[46] Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *Annals of Statistics*, 40:1846–1877.

[47] Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *Annals of Statistics*, 14:1101–1112.

[48] Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107:1129–1139.

[49] Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109:266–274.

[50] Liu, J., Zhong, W., and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58:1–22.

[51] Luo, S. and Chen, Z. (2014). Sequential lasso cum ebic for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, 109:1229–1240.

[52] Mai, Q. and Zou, H. (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *Annals of Statistics*, 43:1471–1497.

[53] Marchionni, L., Afsari, B., Geman, D., and Leek, J. T. (2013). A simple and reproducible breast cancer prognostic test. *BMC Genomics*, 14:1–7.

[54] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462.

[55] Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pages 3178–3186. PMLR.

[56] Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models. *Annals of Statistics*, 43:991–1026.

[57] Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59:3434–3447.

[58] Serban, N. (2011). A space–time varying coefficient model: the equity of service accessibility. *Annals of Applied Statistics*, 5:2024–2051.

[59] Song, R., Yi, F., and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24:1735–1752.

[60] Spigler, S., Geiger, M., d' Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. (2019). A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52:474001.

[61] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99:879–898.

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

[63] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

[64] Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46:1255–1287.

[65] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111:600–620.

[66] van de Geer, S. (2019). On the asymptotic variance of the debiased Lasso. *Electronic Journal of Statistics*, 13:2970–3008.

[67] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202.

[68] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

[69] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

[70] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G., editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press.

[71] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:1512–1524.

[72] Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747–757.

[73] Wang, J., He, X., and Xu, G. (2020). Debiased inference on treatment effect in a high-dimensional model. *Journal of the American Statistical Association*, 115:442–454.

[74] Wang, K. and Lin, L. (2016). Robust structure identification and variable selection in partial linear varying coefficient models. *Journal of Statistical Planning and Inference*, 174:153–168.

[75] Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556–1569.

[76] 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.

[77] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

[78] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:217–242.

[79] Zhong, W., Duan, S., and Zhu, L. (2020). Forward additive regression for ultrahigh-dimensional nonparametric additive models. *Statistica Sinica*, 30:175–192.

[80] Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106:1464–1475.