Robust estimation in high-dimensional sparse heteroscedastic linear models

Duzhe Wang dwang282@wisc.edu
Po-Ling Loh ploh@stat.wisc.edu

Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706

July 20, 2020

Abstract

We study the problem of robust estimation for the regression parameter in high-dimensional sparse heteroscedastic linear models. In our model, the variance of the random error in each observation depends on both predictor variables and regression coefficients. Our first goal is to provide theoretical guarantees of the Lasso under the assumed type of heteroscedasticity. We show in both theory and experiments that the Lasso is estimation consistent and variable selection consistent under mild sufficient conditions, even in the presence of heteroscedastic noise. The second goal of this paper is to robustify the Lasso estimates given that the parametric form of the variance function is known. We propose a new one-step estimator and derive its asymptotical properties for uncertainty quantification and valid statistical inference. Our approach creates an efficient paradigm to treat heteroscedasticity for high-dimensional sparse linear models.

1 Introduction

In standard formulations of linear regression analysis, it is often assumed that the random error is independent and identically distributed, and the variance does not depend on both predictor variables and regression parameters. However, this *homoscedasticity* assumption is not sensible for many real-world applications. For instance, it has been shown that *heteroscedasticity*, a phenomenon of unequal noise levels across different observations, is not uncommon in a variety of scientific areas ranging from genomics [11], medical imaging [14] to econometrics [9]. Moreover, heteroscedastic noise naturally exists in large and complex datasets, where the observations may be collected from multiple sources or be contaminated by outliers. While modern data collection technology brings new opportunities for regression applications, the accompanying occurrence of heteroscedasticity may cause serious problems for statistical analysis. It is a classical result, implied by the well-known Cramér-Rao lower bound, that the ordinary least squares estimator is inefficient in the presence of heteroscedastic noise [18].

Problem setup. In this paper, our primary interest is in robust estimation and inference about the regression coefficients under heteroscedasticity. More specifically, we consider the linear model

$$y = X\beta^* + \varepsilon, \tag{1}$$

where $X = (X_1, ..., X_N)^T \in \mathbb{R}^{N\times n}$ is the design matrix with $X_i \in \mathbb{R}^n$, $y = (y_1, ..., y_N)^T \in \mathbb{R}^N$ is the observed response, $\beta^* = (\beta_1^*, ..., \beta_n^*)^T \in \mathbb{R}^n$ is the unknown true regression parameter, and $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)^T \in \mathbb{R}^N$ is a vector of unobserved random errors with mean zero and diagonal conditional variance matrix $W \in \mathbb{R}^{N\times N}$. When the principal diagonal of W contains same constant elements, (1) reduces to the homoscedastic linear model, which has been studied extensively in statistics since as early as the 1800s. Otherwise, the matrix W expresses the heteroscedasticity of noise. We suspect a heteroscedastic model because the dispersion of the residuals is significantly affected by the magnitude of the fitted values [6]. In general, larger magnitude of observed value is accompanied by larger variance in practice. Therefore, it is usually common to assume that W_i, the ith diagonal component of W, is a given function of X_i and β^*. For example, Bickel [1] considered the log-linear form of W_i—namely, $W_i = c\exp\left(c' X_i^T \beta^*\right)$. On the other hand, Jia et al. [14] focused on the Poisson-like heteroscedasticity, where W_i can be written as $W_i = c\left|X_i^T\beta^*\right|$. Inspired by above settings, we are interested in the conditional heteroscedastic model, where W_i takes the following general form:

$$W_i = g\left(X_i, \beta^*\right). \tag{2}$$

Throughout the paper, we assume that the parametric function form of g is known. We will introduce more assumptions on g in Section 2 and Section 3. Furthermore, driven by a large amount of modern applications in various fields, we will work within the high-dimensional framework, which allows the number of predictors n to grow and substantially exceed the sample size N.

The bulk of statistics literature on heteroscedasticity—notably a series of influential work by Carroll and his collaborators [5, 3, 4, 6, 7, 8, 10]—dates back to the last century, when most of the theoretical work was devoted to studying the classical low-dimensional scenario that the sample size N was able to diverge with the ambient dimension n fixed. In the direction of hypothesis testing, Bickel [1], and Carroll and Ruppert [4] suggested various robust tests for heteroscedasticity. Under similar setup of the variance function to ours, Jobson and Fuller [15], and Carroll and Ruppert [6] studied different classes of generalized weighted least squares estimates for the regression parameter, and established the corresponding asymptotic properties of their estimators. On the other hand, Carroll [3] considered the case that parametric form of the variance function is unknown. He proved that if the variance function is smooth to the design or the mean response, then there still exist efficient estimates for the regression parameter. Unfortunately, none of these traditional methods are readily applied to the high-dimensional settings that are of interest in our paper.

Our contributions. Despite the substantial body of work on heteroscedastic linear regression in low dimensions as discussed earlier, however, very little research has been conducted in the high-dimensional regime. In this paper, we are interested in the following two key questions:

(a) *What are theoretical guarantees of the Lasso under heteroscedasticity?* Although heteroscedasticity is ubiquitous in real-world high-dimensional problems, practitioners usually ignore its existence and choose the standard methods in applied linear regression. For example, to estimate the sparse regression parameter in high-dimensional heteroscedastic linear models, we often use the canonical Lasso in practice, which is defined by

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}} \quad \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{3}$$

Here, $\lambda > 0$ is a tuning parameter. However, from a theoretical perspective, the performance of the Lasso is only well understood under assumptions of the standard sparse linear model

with homoscedastic noise. Therefore, as an initial step to study heteroscedasticity in high dimensions, we are interested in performance of the Lasso for the conditional heteroscedastic model defined in (1).

(b) *How can we robustify the Lasso estimates for the regression parameter when some side information on the variance function is available?* In particular, we focus on the scenario that the parametric form of the variance function g is known. In such a setting, we are interested in designing estimators with small variances and quantifying their uncertainties.

The current paper is well motivated by above questions. We now summarize the main contributions of our paper as follows:

(a) Our first major contribution is to provide a general set of mild sufficient conditions under which the Lasso is statistically consistent, even in the presence of heteroscedastic noise. More specifically, in Section 2, we consider the sub-Gaussian random design, and derive upper bounds for the ℓ_2-estimation error and the ℓ_1-estimation error of the Lasso under the conditional heteroscedastic model (1). In Theorem 1, we show that for an appropriate choice of the tuning parameter λ, the Lasso is ℓ_2-consistent if $\frac{sL_2 \log n}{N} = o(1)$, where s is the number of nonzero elements of β^* and L_2 is the largest conditional error variance. Furthermore, in Theorem 2, we prove that the Lasso is variable selection consistent under similar mutual incoherence condition and beta-min condition as in the standard case.

(b) Our second major contribution is to study robustification of the Lasso under heteroscedasticity. In Section 3, we develop a new type of one-step estimator for the regression parameter and derive its limiting distribution in Theorem 3. Asymptotical properties of the one-step estimator not only helps us to quantify its uncertainty, but also provides us theoretical foundations to make valid confidence intervals and statistical tests for the regression parameter.

Related work. Next, we highlight several recent papers on related topics. Jia et al. [14] examined performance of the Lasso under Poisson-like heteroscedasticity. They provided necessary and sufficient conditions for the sign consistency of the Lasso under a sparse Poisson-like model in the high-dimension regime. But the authors did not develop new estimation approaches to utilize the useful information given by Poisson-like heteroscedasticity. Another related work is that of Wagener and Dette [21]. In their paper, the authors proposed a weighted adaptive Lasso to estimate the regression parameter. They showed that the weighted adaptive Lasso performed consistent model selection and was asymptotically normal. Wagener and Dette [21] dealt with the deterministic design, while our paper focuses on the random design. Compared with theirs, the randomness in the design matrix requires a more involved analysis.

Finally, we discuss another two work by Daye et al. [11] and Sharpnack and Kolar [19]. Slightly different with our setup, both these two papers considered a log-linear heteroscedastic model with variances parameterized by an independent set of parameters rather than the regression vector. Daye et al. [11] focused exclusively on estimating the regression parameter. The authors discussed a likelihood-based approach, which optimized the regularized negative log-likelihood with ℓ_1-penalties on both the regression and variance parameters simultaneously. They devised an efficient coordinate descent algorithm to minimize the non-convex objective function. On the other hand, the goal of Sharpnack and Kolar [19] was to estimate both the regression and variance parameters. They proposed a heteroscedastic iterative penalized pseudolikelihood optimizer, which employed non-convex penalties.

Notation. For functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ to mean that $f(n) \leqslant cg(n)$ for some universal constant $c \in (0, \infty)$. Similarly, we write $f(n) \gtrsim g(n)$ when $f(n) \geqslant c'g(n)$ for some universal constant $c' \in (0, \infty)$, and write $f(n) \asymp g(n)$ to mean that $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold simultaneously. We write $f(n) = o(g(n))$ to mean that $\lim_{n\to\infty} f(n)/g(n) = 0$. We write $X_n = o_{\mathbb{P}}(a_n)$ to mean X_n/a_n converges to 0 in probability, and write $X_n = \mathcal{O}_{\mathbb{P}}(a_n)$ to mean X_n/a_n is stochastically bounded. We use I_n to denote the identity matrix of size $n \times n$. For a matrix $M \in \mathbb{R}^{m\times n}$, we write $\|M\|_{op}$ to denote the ℓ_2 operator norm, $\|M\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |M_{ij}|$ to denote the ℓ_1 operator norm, $\|M\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |M_{ij}|$ to denote the ℓ_∞ operator norm, $\|M\|_F = \left(\sum_j \sum_i M_{ij}^2\right)^{1/2}$ to denote the Frobenius norm, and write $\|M\|_{1,1} = \sum_{i=1}^m \sum_{j=1}^n |M_{ij}|$ to denote the element-wise ℓ_1 norm. We write $\|\cdot\|_\infty$ to denote the element-wise infinity norm for both vectors and matrices. We write $\|\cdot\|_0$ to denote the number of non-zero elements in a vector. For squared matrices, we write $\lambda_{\min}$ and $\lambda_{\max}$ to denote the minimum and maximum eigenvalues respectively. For $q, r > 0$, we write $\mathbb{B}_q(r)$ to denote the ℓ_q-ball of radius r centered around 0. We use c, c', c'', C', and C'' to denote positive constants, where we may use the same notation to refer to different constants as we move between results. For a constant $\sigma > 0$, a random variable $X \in \mathbb{R}$ is said to be σ-sub-Gaussian if $\mathbb{E}(X) = 0$ and its moment generating function satisfies $\mathbb{E}[\exp(tX)] \leqslant \exp\left(\sigma^2 t^2/2\right)$ for all $t \in \mathbb{R}$. Similarly, a random vector $X \in \mathbb{R}^n$ is said to be σ-sub-Gaussian if $\mathbb{E}(X) = 0$ and $u^T X$ is σ-sub-Gaussian for any unit vector $u \in \mathbb{S}^{n-1}$. Furthermore, if a random matrix $X \in \mathbb{R}^{N\times n}$ is formed by drawing each row $X_i \in \mathbb{R}^n$ in an i.i.d. manner from a σ-sub-Gaussian distribution with covariance matrix Σ, then we say X is a row-wise (σ, Σ)-sub-Gaussian random matrix.

2 The Lasso under heteroscedasticity

We start with studying theoretical properties of the Lasso for the conditional heteroscedastic linear model (1). In this section, we make use of the following assumptions:

Assumption 1. The true regression parameter β^* is s-sparse. That is, $\|\beta^*\|_0 \leqslant s$, where $s < N \ll n$. In what follows, let S denote the support set of β^*. Furthermore, the triple (s, N, n) is allowed to increase to ∞.

Assumption 2. (a) The design matrix X is a row-wise (σ_x, Σ_x)-sub-Gaussian random matrix. Furthermore, eigenvalues of Σ_x are bounded by dimension-free constants. That is,

$$c_{\min} \leqslant \lambda_{\min}(\Sigma_x) \leqslant \lambda_{\max}(\Sigma_x) \leqslant c_{\max},$$

where $c_{\min} > 0$ and $c_{\max} > 0$ are constants.

(b) The random errors ε_i, $i = 1, ..., N$, are independent, conditionally normal random variables with

$$\mathbb{E}(\varepsilon_i \mid X_i) = 0, \quad \mathbb{E}\left(\varepsilon_i^2 \mid X_i\right) = g(X_i, \beta^*),$$

where $0 < L_1 \leqslant g(X_i, \beta^*) \leqslant L_2 < \infty$.

Assumption 3. For covariance matrix Σ_x, there exists a constant $\alpha \in (0, 1)$, such that

$$\left\|(\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant \frac{\alpha}{2}.$$

Assumption 1 and Assumption 2 contain a standard set of conditions for elements of β^*, X, and ε. Note that in Part (b) of Assumption 2, we assume that the variance function is lower and upper bounded, which is often imposed when analyzing heteroscedasticity [15, 21]. Assumption 3, namely the mutual incoherence condition, is usually assumed for variable selection of the Lasso in the standard high-dimensional sparse linear models [22]. With above assumptions at hand, we have the first main result, which concerns the estimation error of the Lasso under heteroscedasticity.

Theorem 1. Assume that Assumption 1 and Assumption 2 hold. If $N \gtrsim s \log n$ and

$$\lambda \geqslant 4(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2)^{1/2}\sqrt{\frac{L_2 \log n}{N}},$$

then with probability at least $1 - c\exp(-c' \log n)$, we have

$$\|\hat{\beta} - \beta^*\|_2 \leqslant 6\lambda_{\min}^{-1}(\Sigma_x)\,\lambda\sqrt{s}, \quad \|\hat{\beta} - \beta^*\|_1 \leqslant 24\lambda_{\min}^{-1}(\Sigma_x)\,\lambda s.$$

The proof of Theorem 1 is contained in Section 6.1.

Remark 1. We provide several comments on the upper bounds derived in Theorem 1. First, if $\lambda \asymp \sqrt{\frac{L_2 \log n}{N}}$, then we have

$$\left\|\hat{\beta} - \beta^*\right\|_2 \leqslant \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{sL_2 \log n}{N}}\right) \quad \text{and} \quad \left\|\hat{\beta} - \beta^*\right\|_1 \leqslant \mathcal{O}_{\mathbb{P}}\left(s\sqrt{\frac{L_2 \log n}{N}}\right). \tag{4}$$

Secondly, (4) implies that the Lasso is ℓ_2-consistent if $\frac{sL_2 \log n}{N} = o(1)$. Lastly, if we consider the homoscedastic case—namely, $W_i = \sigma_\varepsilon^2$ for $i = 1, ..., N$, then inequalities in (4) reduce to

$$\left\|\hat{\beta} - \beta^*\right\|_2 \leqslant \mathcal{O}_{\mathbb{P}}\left(\sigma_\varepsilon\sqrt{\frac{s \log n}{N}}\right) \quad \text{and} \quad \left\|\hat{\beta} - \beta^*\right\|_1 \leqslant \mathcal{O}_{\mathbb{P}}\left(\sigma_\varepsilon s\sqrt{\frac{\log n}{N}}\right),$$

which are standard results of the Lasso in statistics literature [17].

Next, we shift our attention to variable selection performance of the Lasso under heteroscedasticity. We say that the Lasso is variable selection consistent if there exists an appropriate choice of λ, such that $\mathbb{P}\left(S = \widehat{S}\right) \to 1$ as the triple (s, N, n) goes to ∞, where $\widehat{S}$ is the support set of $\hat{\beta}$. The following theorem, proved in Section 6.2, claims that the Lasso performs consistent variable selection, even in the presence of heteroscedasticity. The proof of Theorem 2 is based on the primal-dual witness (PDW) type arguments. See Appendix B for a detailed introduction to the PDW construction.

Theorem 2. Assume that Assumption 1, Assumption 2, and Assumption 3 hold. Let

$$\lambda \geqslant \frac{4}{1-\alpha}\left(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2\right)^{1/2}\sqrt{\frac{L_2 \log(n-s)}{N}}.$$

(a) Under the sample size scaling

$$N^2 \gtrsim \frac{s^5 \log[s(n-s)]}{\alpha^2}, \quad N \gtrsim \frac{s^4 \log[s(n-s)]}{\alpha^2}, \quad \text{and} \quad N \gtrsim \frac{s^5}{\alpha^2},$$

with probability at least $1 - c\exp(-c' \log s)$, we have $\widehat{S} \subseteq S$.

(b) In addition, if

$$\min_{i\in S}|\beta_i^*| > \sqrt{\frac{8}{\lambda_{\min}(\Sigma_x)}}\sqrt{\frac{L_2\log s}{N}} + \lambda\left(\left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty + cs\sqrt{\frac{s}{N}}\right),$$

then with probability at least $1-c'\exp\left(-c''\log s\right)$, we have $S=\widehat{S}$.

Finally, results in Theorem 1 and Theorem 2 together lead to the following corollary, which is useful for the statistical analysis in the next section.

Corollary 1. Assume that Assumption 1, Assumption 2, and Assumption 3 hold. Let

$$\lambda = \frac{4}{1-\alpha}\left(\lambda_{\max}\left(\Sigma_x\right)+4\sigma_x^2\right)^{1/2}\sqrt{\frac{L_2\log n}{N}}. \tag{5}$$

Under the sample size scaling in Part (a) of Theorem 2, with probability at least $1-c\exp\left(-c'\log s\right)$, $\hat{\beta}$ has the following properties:

(a) No false inclusion: $\widehat{S}\subseteq S$.

(b) ℓ_2-estimation error:

$$\left\|\hat{\beta}-\beta^*\right\|_2 \leqslant \frac{24}{1-\alpha}\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)^{1/2}\lambda_{\min}^{-1}\left(\Sigma_x\right)\sqrt{\frac{sL_2\log n}{N}}.$$

(c) ℓ_1-estimation error:

$$\left\|\hat{\beta}-\beta^*\right\|_1 \leqslant \frac{96}{1-\alpha}\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)^{1/2}\lambda_{\min}^{-1}\left(\Sigma_x\right)s\sqrt{\frac{L_2\log n}{N}}.$$

Theorem 1, Theorem 2, and Corollary 1 provide theoretical guarantees for the estimation and variable selection performance of the Lasso under heteroscedasticity. But as a starting point in the statistical analysis routine for high-dimensional heteroscedastic linear models, the Lasso has two drawbacks: (a) it is difficult to quantify uncertainty of the Lasso performance. In other words, the Lasso does not have a uniform tractable limiting distribution [16], which can be seen from Figure 5 in Section 4.2. (b) The standard procedure in (3) does not efficiently employ the given information of the variance function. We will discuss a mature treatment paradigm to heteroscedasticity in the next section.

3 Robustifying the Lasso under heteroscedasticity

We now address the problem of robustifying the Lasso estimates when the parametric form of the variance function g is known. As mentioned in previous sections, we have two main goals: (a) make use of the variance information to devise new estimators, and (b) quantify uncertainties of the proposed estimators.

Following recent advances in statistical inference for high-dimensional homoscedastic linear models [13, 20, 24], we adopt the general framework of one-step estimators: start with the Lasso estimates, and then take a single update towards a certain direction. In this paper, we carefully choose the desired direction in the heteroscedastic context. The whole procedure of our approach is described in Algorithm 1 below:

Algorithm 1 One-step estimator for β^* in high-dimensional heteroscedastic linear models

Input: N pairs of (X_i, y_i), variance function g, and tuning parameters λ and μ.
Output: One-step estimator $\tilde{\beta}$.

1. Solve the following optimization problem to obtain a preliminary estimator $\hat{\beta}$:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \quad \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

2. Set $\widehat{W}_i = g(X_i, \hat{\beta})$, $\widehat{W} = \operatorname{diag}(\widehat{W}_1, ..., \widehat{W}_N) \in \mathbb{R}^{N\times N}$, and $\hat{\Sigma}_N = \frac{1}{N} X^T \widehat{W}^{-1} X \in \mathbb{R}^{n\times n}$.

3. Solve the following optimization problem to obtain $\widehat{\Theta}$:

$$\begin{aligned} \widehat{\Theta} = \underset{\Theta \in \mathbb{R}^{n\times n}}{\operatorname{argmin}} \quad & \|\Theta\|_{1,1} \\ \text{subject to} \quad & \|\Theta\hat{\Sigma}_N - I_n\|_\infty \leqslant \mu. \end{aligned} \tag{6}$$

   The above optimization consists of n subproblems: for $i = 1, 2, ..., n$,

$$\begin{aligned} \hat{\theta}_i = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \quad & \|\theta\|_1 \\ \text{subject to} \quad & \left\|\hat{\Sigma}_N\theta - e_i\right\|_\infty \leqslant \mu. \end{aligned} \tag{7}$$

   Then $\widehat{\Theta} = \left(\hat{\theta}_1, ..., \hat{\theta}_n\right)^T$. (6) and (7) can be efficiently solved via the ADMM algorithm [2].

4. Construct the one-step estimator as

$$\tilde{\beta} = \hat{\beta} + \frac{1}{N}\widehat{\Theta} X^T \widehat{W}^{-1}(y - X\hat{\beta}).$$

Next, we present theoretical guarantees of the proposed one-step estimator. Our theory requires another set of assumptions:

Assumption 4. For all β and X_i, the first two partial derivatives of $g(X_i, \beta)$ with respect to β are uniformly bounded functions of β. That is, there exists a constant $L_g > 0$, such that

$$\left\|\frac{\partial g(X_i, \beta)}{\partial \beta}\right\|_\infty \leqslant L_g \quad \text{and} \quad \left\|\frac{\partial^2 g(X_i, \beta)}{\partial \beta^2}\right\|_\infty \leqslant L_g.$$

Assumption 5. Expectation of $\frac{X_i X_i^T}{W_i}$ exists. That is, $\mathbb{E}\left(\frac{X_i X_i^T}{W_i}\right) = \Sigma^*$, where $\Sigma^* \in \mathbb{R}^{n\times n}$ is positive definite. In what follows, let $\Theta^* \in \mathbb{R}^{n\times n}$ denote the inverse of Σ^*. Furthermore, we assume the following conditions for Θ^*:

(a) Eigenvalues of Θ^* are bounded. That is,

$$c_{\min} \leqslant \lambda_{\min}(\Theta^*) \leqslant \lambda_{\max}(\Theta^*) \leqslant c_{\max},$$

where $c_{\min} > 0$ and $c_{\max} > 0$ are constants.

(b) For $1 \leqslant i \leqslant n$, $\|\Theta_i^*\|_0 \leqslant k$, where Θ_i^* is the ith row of Θ^* and $k \ll n$.

(c) There exists a number M, such that $\|\Theta^*\|_\infty \leqslant M$.

Assumption 4 imposes similar conditions on the variance function g as in [15] and [21]. As mentioned in [15], this assumption is usually required in nonlinear least squares estimation. In Assumption 5, we define an appropriate inverse variance matrix Θ^* in order to adapt the statistical analysis to the heteroscedastic setting. Note that $\widehat{\Theta}$ in step 3 of Algorithm 1 can be viewed as an estimator of Θ^*. We now derive the limiting distribution of the one-step estimator in the following:

Theorem 3. Assume that Assumption 1 to Assumption 5 hold. In Algorithm 1, let

$$\lambda \asymp \sqrt{\frac{L_2 \log n}{N}} \quad \text{and} \quad \mu \asymp \frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}.$$

If N satisfies the sample size scaling in Part (a) of Theorem 2 and $L_2 M s^2 (\log n)^2/\sqrt{N} \to 0$, then we have for $1 \leqslant j \leqslant n$,

$$\sqrt{N}\left(\tilde{\beta}_j - \beta_j^*\right) \xrightarrow{d} N\left(0, e_j^T \Theta^* e_j\right).$$

The proof of Theorem 3 is contained in Section 6.3.

Remark 2. Theorem 3 shows that $\tilde{\beta}_j$ is asymptotically normal with mean β_j^* and variance $\frac{e_j^T \Theta^* e_j}{N}$. Furthermore, it has several consequences for statistical inference of regression parameter under heteroscedasticity. For example, similar arguments yield

$$\frac{\sqrt{N}\left(\tilde{\beta}_j - \beta_j^*\right)}{\sqrt{e_j^T \widehat{\Theta}\widehat{\Sigma}_N \widehat{\Theta}^T e_j}} \to N(0,1). \tag{8}$$

Let $\Phi(x)$ be the cumulative distribution function of $N(0,1)$. Then

$$\left[\tilde{\beta}_j - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{e_j^T \widehat{\Theta}\widehat{\Sigma}_N \widehat{\Theta}^T e_j}{N}}, \quad \tilde{\beta}_j + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{e_j^T \widehat{\Theta}\widehat{\Sigma}_N \widehat{\Theta}^T e_j}{N}}\right]$$

provides an asymptotically valid $(1-\alpha)$-confidence interval for β_j. On the other hand, the pivot in (8) is a test statistic for testing whether β_j is equal to 0.

4 Simulations

In this section, we perform simulation studies to validate our theoretical results of this paper. Our focus of this section is two-fold. First, we study the finite sample performance of the Lasso under different kinds of heteroscedasticity. Second, we investigate the asymptotic behavior of the Lasso and one-step estimator. The code for replicating our experiments is available on GitHub https://github.com/dzwang91/HHLM.

4.1 Performance of the Lasso under heteroscedasticity

In the first set of simulations, our goal is to verify statistical consistency (estimation consistency, prediction consistency, and variable selection consistency) of the Lasso when the underlying model is heteroscedastic. In all experiments, we set

$$\beta^* = (3, 4, 3, 1.5, 2, 1.5, 0, ..., 0)^T .$$

We first generate predictor variables X_i from the normal distribution $N(0, \Sigma_x)$, where $(\Sigma_x)_{S^cS} = 0$, and $(\Sigma_x)_{ij} = 0.5^{|i-j|}$ if both i and j in are S or both i and j are in S^c. Then the random error ε_i is drawn from the conditional distribution $\varepsilon_i \mid X_i \sim N(0, W_i)$. Finally, we generate the response y_i via the linear model $y_i = X_i^T\beta^* + \varepsilon_i$. We consider the following four different scenarios on W_i:

(a) Type 1:

$$W_i = \min\left(\frac{1}{25}\exp\left(\frac{1}{2}\left|X_i^T\beta^*\right|\right), 5\right).$$

(b) Type 2:

$$W_i = \min\left(\frac{1}{50}\exp\left(\frac{1}{20}\left(X_i^T\beta^*\right)^2\right), 5\right).$$

(c) Type 3:

$$W_i = \begin{cases} \frac{1}{20} & \frac{1}{4}|X_i^T\beta^*| \leqslant \frac{1}{20} \\ \frac{1}{4}|X_i^T\beta^*| & \frac{1}{20} < \frac{1}{4}|X_i^T\beta^*| \leqslant 5 \\ 5 & \frac{1}{4}|X_i^T\beta^*| > 5 \end{cases}$$

(d) Type 4:

$$W_i = \begin{cases} \frac{1}{20} & \frac{1}{16}\left(X_i^T\beta^*\right)^2 \leqslant \frac{1}{20} \\ \frac{1}{16}\left(X_i^T\beta^*\right)^2 & \frac{1}{20} < \frac{1}{16}\left(X_i^T\beta^*\right)^2 \leqslant 5 \\ 5 & \frac{1}{16}\left(X_i^T\beta^*\right)^2 > 5 \end{cases}$$

These four types are similar to those in [21]. For each type of heteroscedastic models, we consider three ambient dimensions: $n = 150, 300$, and 500. For each n, we run simulations for eight different sample sizes (see Figures 1, 2, 3, and 4). Furthermore, following our theory, we set the tuning parameter $\lambda = 0.75\sqrt{\lambda_{\max}(\Sigma_x)\frac{L_2 \log n}{N}}$ in each simulation.

Figure 1, Figure 2, Figure 3, and Figure 4 show the simulation results for those four heteroscedastic linear models listed above. Each point in these figures represents an average over 200 trials. Panel (a), Panel (b), and Panel (c) in each figure confirm that the ℓ_2-error, the ℓ_1-error, and the mean squared prediction error decrease to 0 as the sample size increases, showing the estimation and prediction consistency of the Lasso under different kinds of heteroscedasticity. In Panel (d), we see that the empirical probability of recovering the correct support set transitions sharply from 0 to 1, which implies that the Lasso is variable selection consistent, as stated in Theorem 2.

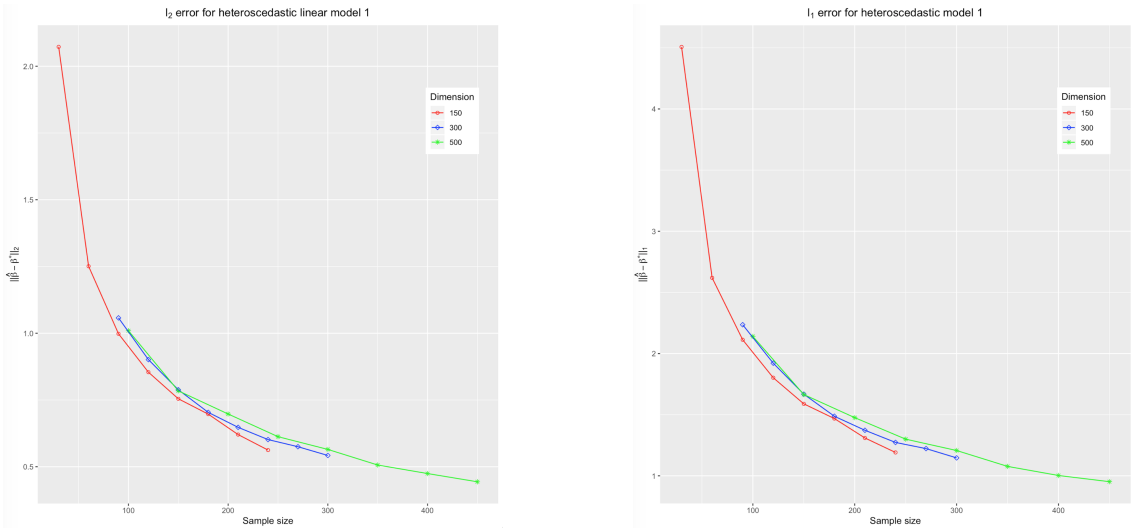


Figure 1: Plots showing simulation results of the Lasso for high-dimensional heteroscedastic linear model 1 with three problem sizes: $n = 150$ (red), $n = 300$ (blue), and $n = 500$ (green). Each point represents an average over 200 trials. (a) Plot showing consistency of the ℓ_2-estimation error $\left\|\hat{\beta} - \beta^*\right\|_2$. (b) Plot showing consistency of the ℓ_1-estimation error $\left\|\hat{\beta} - \beta^*\right\|_1$. (c) Plot showing consistency of the mean squared prediction error $\frac{1}{N}\left\|y - X\hat{\beta}\right\|_2^2$. (d) Plot showing variable selection consistency, measured by the empirical success probability in recovering the support set of β^*.

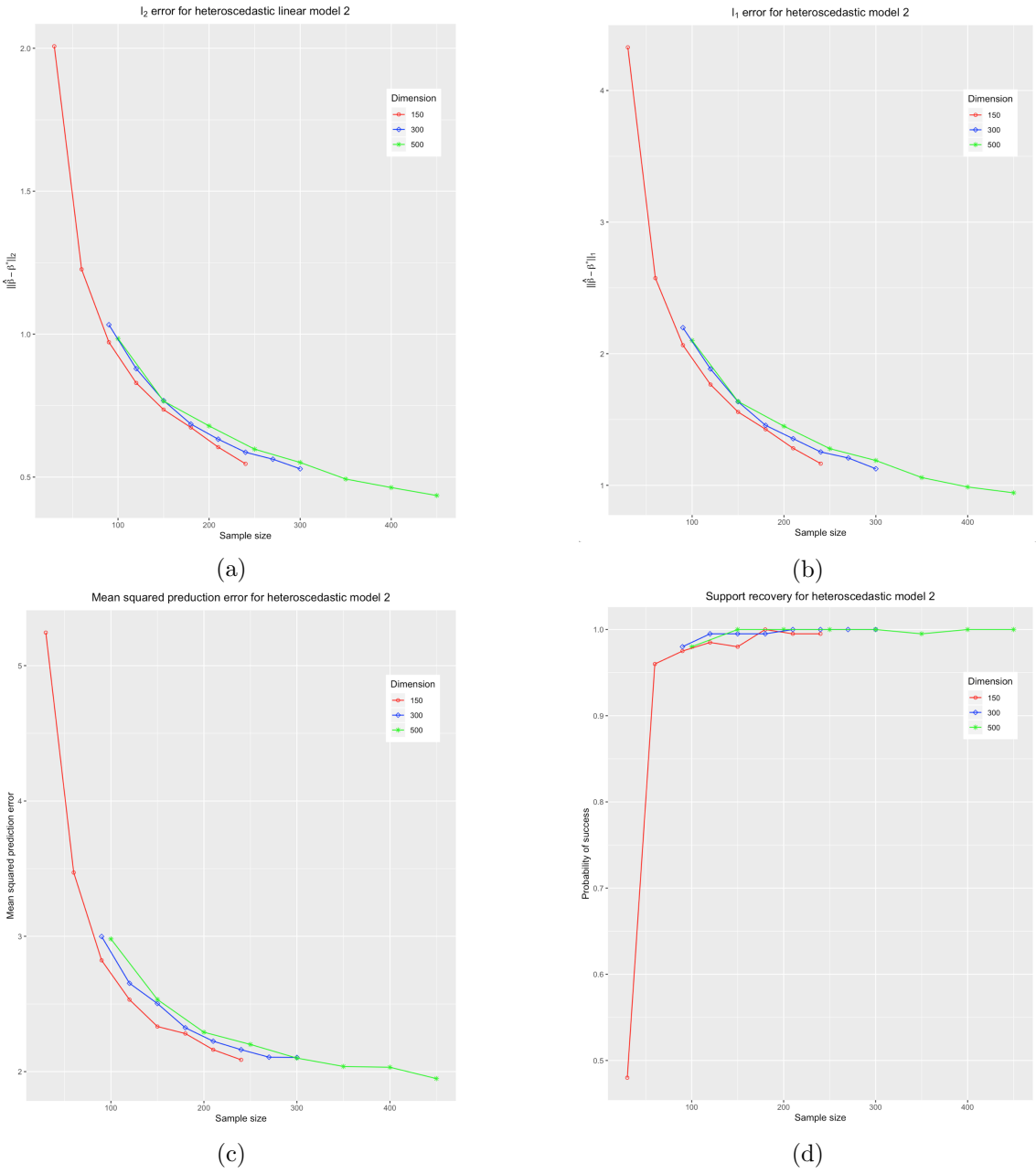


Figure 2: Plots showing simulation results of the Lasso for high-dimensional heteroscedastic linear model 2 with three problem sizes: $n = 150$ (red), $n = 300$ (blue), and $n = 500$ (green). Each point represents an average over 200 trials. (a) Plot showing consistency of the ℓ_2-estimation error $\left\|\hat{\beta} - \beta^*\right\|_2$. (b) Plot showing consistency of the ℓ_1-estimation error $\left\|\hat{\beta} - \beta^*\right\|_1$. (c) Plot showing consistency of the mean squared prediction error $\frac{1}{N}\left\|y - X\hat{\beta}\right\|_2^2$. (d) Plot showing variable selection consistency, measured by the empirical success probability in recovering the support set of β^*.

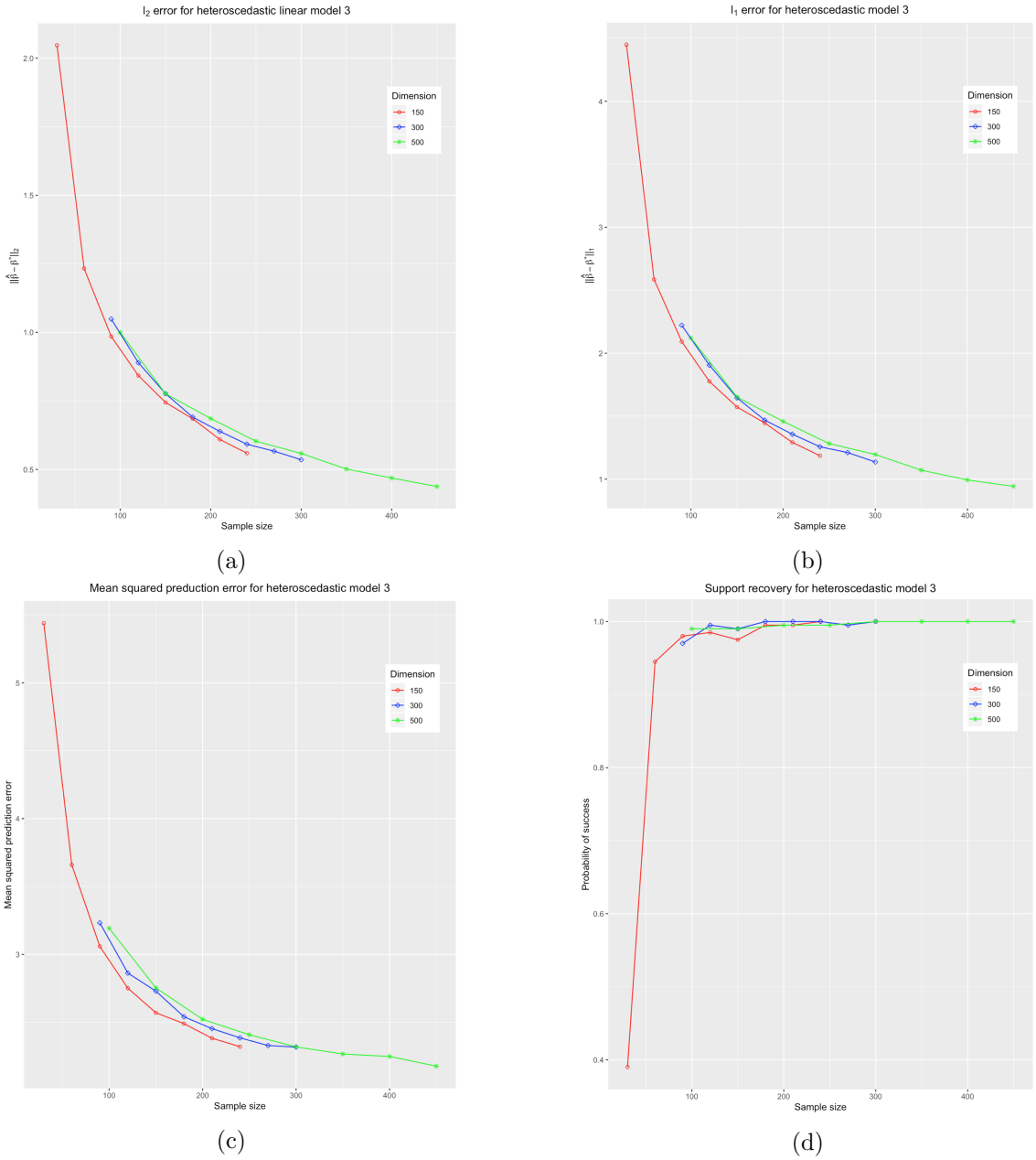


Figure 3: Plots showing simulation results of the Lasso for high-dimensional heteroscedastic linear model 3 with three problem sizes: $n = 150$ (red), $n = 300$ (blue), and $n = 500$ (green). Each point represents an average over 200 trials. (a) Plot showing consistency of the ℓ_2-estimation error $\left\|\hat{\beta} - \beta^*\right\|_2$. (b) Plot showing consistency of the ℓ_1-estimation error $\left\|\hat{\beta} - \beta^*\right\|_1$. (c) Plot showing consistency of the mean squared prediction error $\frac{1}{N}\left\|y - X\hat{\beta}\right\|_2^2$. (d) Plot showing variable selection consistency, measured by the empirical success probability in recovering the support set of β^*.

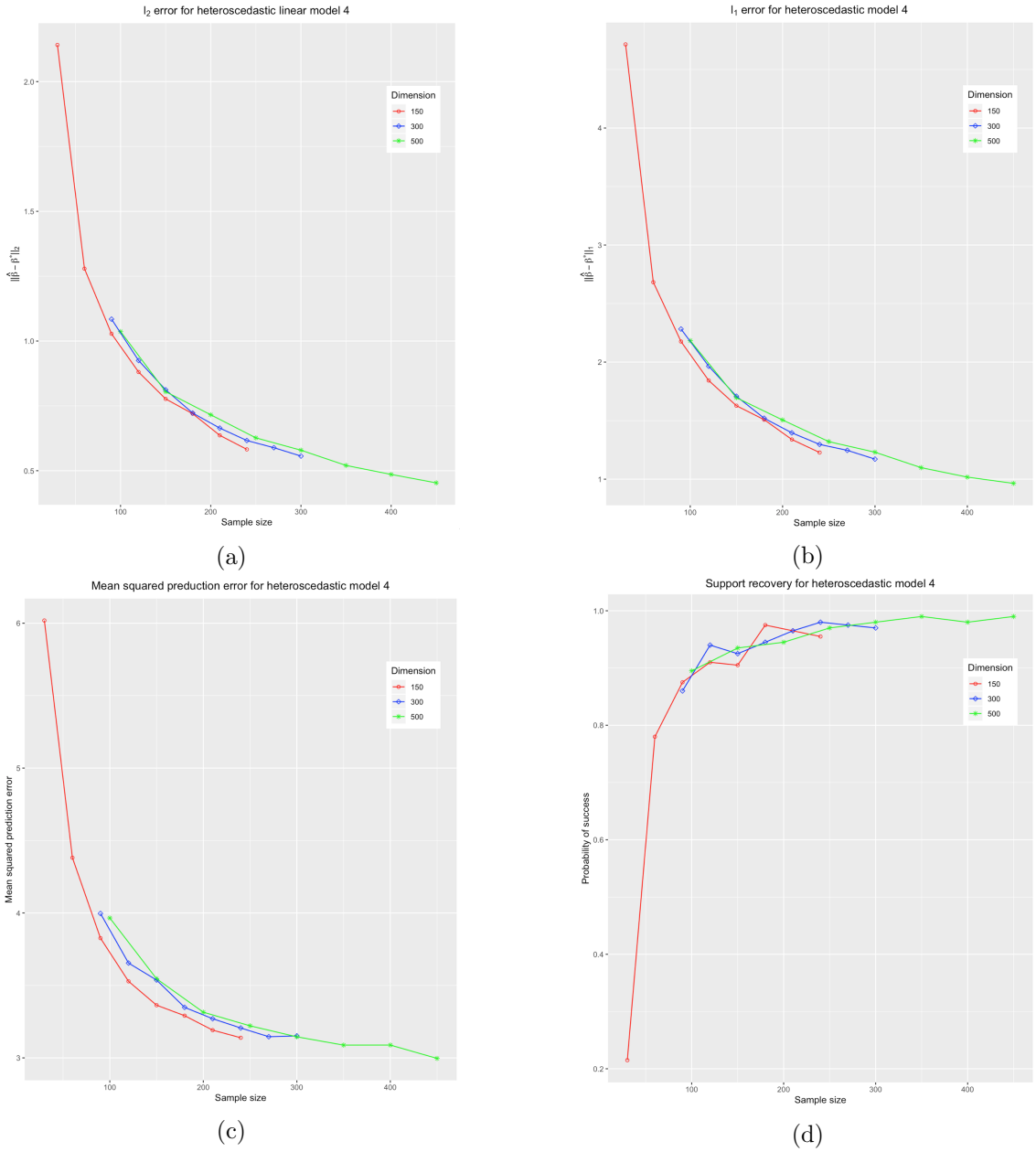


Figure 4: Plots showing simulation results of the Lasso for high-dimensional heteroscedastic linear model 4 with three problem sizes: $n = 150$ (red), $n = 300$ (blue), and $n = 500$ (green). Each point represents an average over 200 trials. (a) Plot showing consistency of the ℓ_2-estimation error $\left\|\hat{\beta} - \beta^*\right\|_2$. (b) Plot showing consistency of the ℓ_1-estimation error $\left\|\hat{\beta} - \beta^*\right\|_1$. (c) Plot showing consistency of the mean squared prediction error $\frac{1}{N}\left\|y - X\hat{\beta}\right\|_2^2$. (d) Plot showing variable selection consistency, measured by the empirical success probability in recovering the support set of β^*.

4.2 Performance of one-step estimators

Our second set of simulations investigates asymptotics of the proposed one-step estimator in Algorithm 1. We take the heteroscedastic linear model 1 with $(N, n) = (120, 150)$ as an example. The data generating process in this set of simulations is the same as the previous set. Based on our theoretical results, in Algorithm 1, we set

$$\lambda = 0.75\sqrt{\lambda_{\max}\left(\Sigma_x\right)\frac{L_2 \log n}{N}} \quad \text{and} \quad \mu = 0.0001\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}\right).$$

Finally, we repeat the whole procedure of Algorithm 1 200 times.

In order to confirm the limiting distribution of one-step estimators stated in Theorem 3, we report simulation results in two cases: (a) the sampling distribution of $\sqrt{N}\left(\tilde{\beta}_5 - \beta_5^*\right)$ where $\beta_5^* \neq 0$, and (b) the sampling distribution of $\sqrt{N}\left(\tilde{\beta}_7 - \beta_7^*\right)$ where $\beta_7^* = 0$. In Figure 5, we present Q-Q plots of $\sqrt{N}\left(\tilde{\beta}_5 - \beta_5^*\right)$ and $\sqrt{N}\left(\tilde{\beta}_7 - \beta_7^*\right)$ in Panel (a) and Panel (b), respectively. It is very clear that both sampling distributions are normal with mean 0. We also perform the Shapiro-Wilk normality test for samples from $\sqrt{N}\left(\tilde{\beta}_5 - \beta_5^*\right)$ and $\sqrt{N}\left(\tilde{\beta}_7 - \beta_7^*\right)$. The corresponding p-values are 0.30 and 0.95, which also certifies our conclusions. Furthermore, in Panel (c) and Panel (d) of Figure 5, we display Q-Q plots of $\sqrt{N}\left(\hat{\beta}_5 - \beta_5^*\right)$ and $\sqrt{N}\left(\hat{\beta}_7 - \beta_7^*\right)$, where $\hat{\beta}_5$ and $\hat{\beta}_7$ are obtained in the first step of Algorithm 1. Obviously, we are not able to claim normality from these two Q-Q plots, compared with those in Panel (a) and Panel (b).

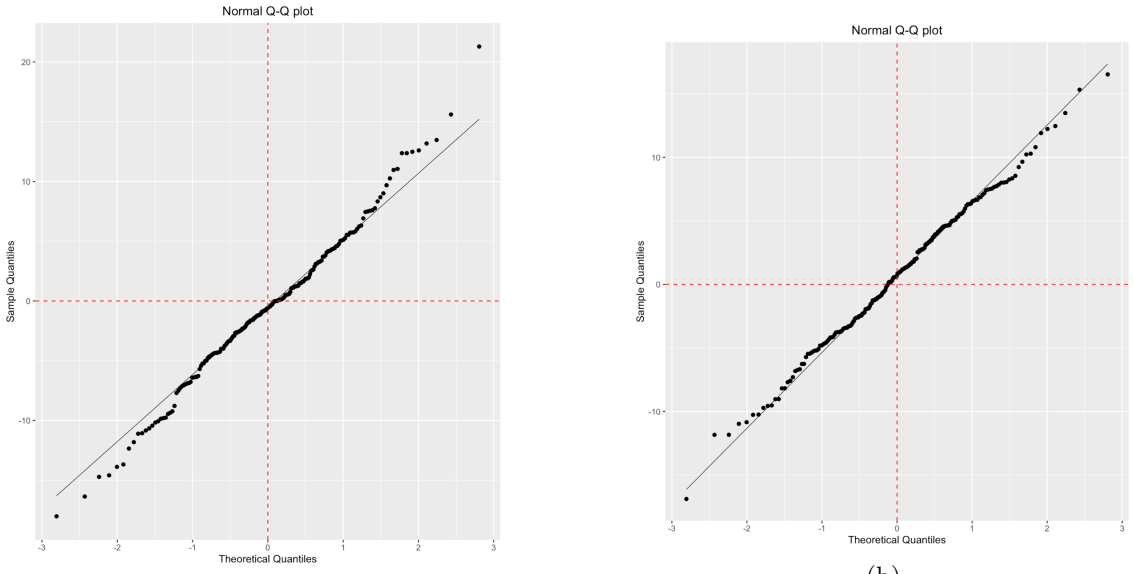
5 Discussion

Our work recommends a number of directions for future research. First, we have only studied the case that the variance is a function of regression parameter. We can extend the current model to a broader one, where variances also depend on parameters other than β^* as in [15]. Secondly, in this paper, we have access to partial information of the variance function—the parametric form of g, through domain knowledge. Another scenario of interest involves completely unknown variance function. As mentioned earlier in Section 1, when only assuming the variance function was smooth in the classical low-dimensional regime, Carroll [3] constructed an estimator of the regression parameter and showed that it was asymptotically equivalent to the weighted least squares estimator with known variances. Likewise, for the high-dimensional regime, it is worthwhile to explore whether there exists some nonparametric approach to recover the results in our paper.

Finally, another line of future work is motivated by some interesting results of Jobson and Fuller [15], which are derived for the fixed design in the low-dimensional setting. Under a similar model to ours, the authors developed a weighted joint least squares estimator $\hat{\beta}^{JF}$ which is asymptotically equivalent to the maximum likelihood estimator. More specifically, they proved

$$\sqrt{N}\left(\hat{\beta}^{JF} - \beta^*\right) \to N\left(0, \left(\lim_{N\to\infty}\frac{1}{N}H^T F^{-1} H + \lim_{N\to\infty} X^T W^{-1} X\right)^{-1}\right), \tag{9}$$

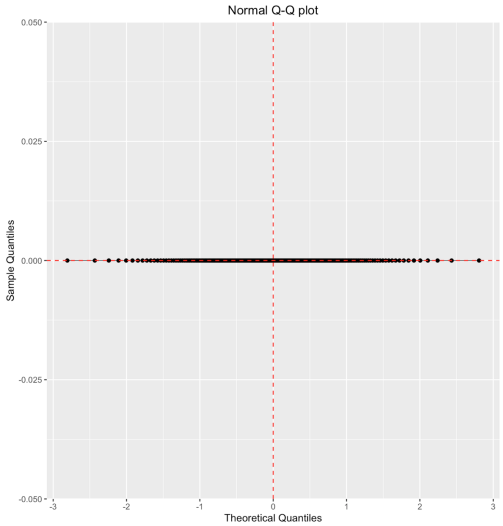
where $F \in \mathbb{R}^{N\times N}$ is a diagonal matrix with ith diagonal element $2g^2\left(X_i, \beta^*\right)$, and $H \in \mathbb{R}^{N\times n}$ with (i, j)th element $H_{ij} = \frac{\partial g(X_i, \beta^*)}{\partial \beta_j}$. It can be seen that the asymptotic variance in (9) is smaller than



(a)

(b)

(c)



(d)

Figure 5: Plots showing simulation results for high-dimensional heteroscedastic linear model 1 with $(N, n) = (120, 150)$. (a) Q-Q plot of $\sqrt{N}\left(\tilde{\beta}_5 - \beta_5^*\right)$. (b) Q-Q plot of $\sqrt{N}\left(\tilde{\beta}_7 - \beta_7^*\right)$. (c) Q-Q plot of $\sqrt{N}\left(\hat{\beta}_5 - \beta_5^*\right)$. (d) Q-Q plot of $\sqrt{N}\left(\hat{\beta}_7 - \beta_7^*\right)$. Panel (a) and Panel (b) verify the asymptotic normality of one-step estimators, while Panel (c) and Panel (d) illustrate that the Lasso does not have a tractable limiting distribution.

$\left(\lim_{N\to\infty} \frac{1}{N} X^T W^{-1} X\right)^{-1}$, the counterpart of which in the random design is Θ^* defined in Assumption 5. Therefore, an important but challenging open question is whether our proposed one-step estimator is statistically efficient in some sense for the high-dimensional conditional heteroscedastic linear model (1). Of course, this question is somewhat vague in that the notion of efficiency has not been well established in the high-dimensional setting. From a semiparametric point of view, Janková and van de Geer [12] studied efficiency lower bounds on variance for high-dimensional homoscedastic linear models. It should be interesting to investigate the similar problem for high-dimensional heteroscedastic linear models to understand limits of the proposed one-step estimator.

6 Proofs of main results

In this section, we prove Theorem 1, Theorem 2, and Theorem 3. Proofs of the supporting lemmas are contained in Appendix A.

6.1 Proof of Theorem 1

First, we assume two conditions:

- The tuning parameter λ satisfies
$$\lambda > \frac{2}{N}\|\varepsilon^T X\|_\infty. \tag{10}$$
- There exists $\eta > 0$, such that
$$\frac{1}{N}\|Xv\|_2^2 \geqslant \eta\|v\|_2^2 \tag{11}$$
for $v \in \mathbb{R}^n$ and $\|v_{S^c}\|_1 \leqslant 3\|v_S\|_1$.

By optimality of $\hat{\beta}$, we have the following basic inequality:

$$\frac{1}{2N}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leqslant \frac{1}{2N}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1.$$

Rearranging terms and letting $\Delta = \hat{\beta} - \beta^*$, we have

$$\begin{aligned}
\frac{1}{2N}\|X\Delta\|_2^2 &\leqslant \frac{1}{N}\left(y - X\beta^*\right)^T X\Delta + \lambda\left(\|\beta^*\|_1 - \|\hat{\beta}\|_1\right) \\
&= \frac{1}{N}\varepsilon^T X\Delta + \lambda\left(\|\beta_S^*\|_1 - \|\hat{\beta}_S\|_1 - \|\hat{\beta}_{S^c}\|_1\right) \\
&\leqslant \frac{1}{N}\varepsilon^T X\Delta + \lambda\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right) \\
&\leqslant \frac{1}{N}\|\varepsilon^T X\|_\infty\|\Delta\|_1 + \lambda\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right).
\end{aligned}$$

When condition (10) holds, we have

$$0 \leqslant \frac{1}{2N}\|X\Delta\|_2^2 \leqslant \frac{\lambda}{2}\left(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1\right) + \lambda\left(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\right).$$

Therefore, we have

$$\|\Delta_{S^c}\|_1 \leqslant 3\|\Delta_S\|_1.$$

Combining condition (11), we have

$$\eta\|\Delta\|_2^2 \leqslant \frac{1}{N}\|X\Delta\|_2^2.$$

Therefore, we have

$$\eta\|\Delta\|_2^2 \leqslant \frac{1}{N}\|X\Delta\|_2^2 \leqslant 3\lambda\|\Delta_S\|_1 \leqslant 3\lambda\sqrt{s}\|\Delta\|_2,$$

implying that

$$\|\Delta\|_2 \leqslant \frac{3\lambda\sqrt{s}}{\eta}. \tag{12}$$

Furthermore, we have

$$\|\Delta\|_1 \leqslant 4\|\Delta_S\|_1 \leqslant 4\sqrt{s}\|\Delta\|_2 \leqslant \frac{12\lambda s}{\eta}. \tag{13}$$

Next, to establish the results in the theorem, we use the following two lemmas, which are proved in Appendix A.1 and Appendix A.2, respectively.

Lemma 1. Assume that Assumption 2 holds. Let $\lambda \geqslant 4(\lambda_{\max}(\Sigma_x)+4\sigma_x^2)^{1/2}L_2^{1/2}(\log n/N)^{1/2}$. Then we have

$$\mathbb{P}\left(\frac{2}{N}\|\varepsilon^T X\|_\infty > \lambda\right) \leqslant 2\exp\left(-\log n\right) + \exp\left(\log n - cN\right),$$

where $c > 0$ is a constant.

Lemma 2. Assume that Part (a) of Assumption 2 holds. Then we have

$$\frac{1}{N}\|Xv\|_2^2 \geqslant \frac{1}{2}\lambda_{\min}(\Sigma_x)\|v\|_2^2$$

for $v \in \mathbb{R}^n$ and $\|v_{S^c}\|_1 \leqslant 3\|v_S\|_1$ with probability at least $1 - 2\exp\left(cs\log n - c'N\right)$, where s and S are defined in Assumption 1, and $c > 0$ and $c' > 0$ are constants.

Therefore, letting $\lambda \geqslant 4(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2)^{1/2}L_2^{1/2}(\log n/N)^{1/2}$ in (10) and $\eta = \frac{1}{2}\lambda_{\min}(\Sigma_x)$ in (11), and combining Lemma 1, Lemma 2, (12), and (13), we have

$$\|\Delta\|_2 \leqslant 6\lambda_{\min}^{-1}\left(\Sigma_x\right)\lambda\sqrt{s}, \quad \|\Delta\|_1 \leqslant 24\lambda_{\min}^{-1}\left(\Sigma_x\right)\lambda s,$$

with probability at least $1 - 2\exp\left(-\log n\right) - c\exp\left(c's\log n - c''N\right)$. Finally, letting $N \gtrsim s\log n$ yields the desired probability in the theorem. Hence the proof is complete.

6.2 Proof of Theorem 2

Results in Theorem 2 build upon the following lemma:

Lemma 3. Let S and $\widehat{S}$ be support sets of β^* and $\hat{\beta}$, respectively.

(a) Assume that the following conditions hold:

(C1) The smallest eigenvalue of $\frac{1}{N}X_S^T X_S$ is bounded below. That is,

$$\lambda_{\min}\left(\frac{1}{N}X_S^T X_S\right) \geqslant c_{\min} > 0. \tag{14}$$

(C2) There exists a constant $\alpha \in (0,1)$ such that

$$\max_{j \in S^c} \left\| (X_S^T X_S)^{-1} X_S^T X_j \right\|_1 \leqslant \alpha. \tag{15}$$

(C3) The tuning parameter λ satisfies

$$\lambda \geqslant \frac{2}{1-\alpha} \left\| X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T \right) \frac{\varepsilon}{N} \right\|_\infty. \tag{16}$$

Then we have $\widehat{S} \subseteq S$.

(b) In addition, assume that the following condition holds:

(C4) $\min_{i \in S} |\beta_i^*|$ satisfies

$$\min_{i \in S} |\beta_i^*| > B(\lambda, X_S, \varepsilon) = \left\| \left(\frac{X_S^T X_S}{N} \right)^{-1} X_S^T \frac{\varepsilon}{N} \right\|_\infty + \lambda \left\| \left(\frac{X_S^T X_S}{N} \right)^{-1} \right\|_\infty. \tag{17}$$

Then we have $S = \widehat{S}$.

The proof of Lemma 3 is deferred to Appendix A.3. Therefore, it suffices to show that for the given assumptions in Theorem 2, conditions (14), (15), (16), and (17) hold with high probability. These probabilistic results are established in the following lemmas:

Lemma 4. Assume that Part (a) of Assumption 2 holds. Under the sample size scaling $N \gtrsim s$, we have

$$\lambda_{\min} \left(\frac{1}{N} X_S^T X_S \right) \geqslant \frac{1}{2} \lambda_{\min}(\Sigma_x),$$

with probability at least $1 - 2\exp(-cN)$.

Lemma 5. Assume that Part (a) of Assumption 2 and Assumption 3 hold. When

$$N^2 \gtrsim \frac{s^5 \log[s(n-s)]}{\alpha^2}, \quad N \gtrsim \frac{s^4 \log[s(n-s)]}{\alpha^2}, \quad \text{and} \quad N \gtrsim \frac{s^5}{\alpha^2},$$

we have (15) holds with probability at least $1 - c' \exp(-c''s) - C' \exp(-C'' \log[s(n-s)])$.

Lemma 6. Assume that Assumption 2 holds. Let

$$\lambda \geqslant \frac{4}{1-\alpha} \left(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2 \right)^{1/2} \sqrt{\frac{L_2 \log(n-s)}{N}}.$$

Under the sample size scaling $N \gtrsim \log(n-s)$, we have

$$\lambda \geqslant \frac{2}{1-\alpha} \left\| X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T \right) \frac{\varepsilon}{N} \right\|_\infty,$$

with probability at least $1 - c\exp(-c' \log(n-s))$.

Lemma 7. Assume that Assumption 2 holds. Under the sample size scaling $N \gtrsim s$, we have

$$B(\lambda, X_S, \varepsilon) \leqslant \sqrt{\frac{8}{\lambda_{\min}(\Sigma_x)}} \sqrt{\frac{L_2 \log s}{N}} + \lambda \left(\left\| (\Sigma_x)_{SS}^{-1} \right\|_\infty + cs\sqrt{\frac{s}{N}} \right),$$

with probability at least $1 - c' \exp(-c'' \log s)$.

Proofs of Lemma 4, Lemma 5, Lemma 6, and Lemma 7 are deferred to Appendix A.4, Appendix A.5, Appendix A.6, and Appendix A.7, respectively. Therefore, the proof of Theorem 2 is complete.

6.3 Proof of Theorem 3

Let $\check{\beta} = \hat{\beta} + \frac{1}{N}\Theta^* X^T W^{-1}\left(y - X\hat{\beta}\right)$ and $\Sigma_N = \frac{1}{N}X^T W^{-1} X$. We divide the proof of Theorem 3 into three main steps: first, we show that

$$\max_j \frac{\left|\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)\right|}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} = o_{\mathbb{P}}(1) \tag{18}$$

under given assumptions in Section 6.3.1. Next, in Section 6.3.2, we show that for $1 \leqslant j \leqslant n$,

$$\frac{\sqrt{N}\left(\check{\beta}_j - \beta_j^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} \xrightarrow{d} N(0,1). \tag{19}$$

Finally, we combine these previous results in Section 6.3.3 to obtain the desired result in the theorem.

6.3.1 Step 1 of proof

For $1 \leqslant j \leqslant n$, we have

$$\begin{aligned}\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right) = \sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right) - \sqrt{N} e_j^T \left(\widehat{\Theta}\widehat{\Sigma}_N - I_n\right)\left(\hat{\beta} - \beta^*\right) \\ + \frac{1}{\sqrt{N}} e_j^T \left(\widehat{\Theta} - \Theta^*\right) X^T \widehat{W}^{-1}\varepsilon + \frac{1}{\sqrt{N}} e_j^T \Theta^* X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\max_j \left|\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)\right| &\leqslant \max_j \left|\sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right)\right| + \max_j \left|\sqrt{N} e_j^T \left(\widehat{\Theta}\widehat{\Sigma}_N - I_n\right)\left(\hat{\beta} - \beta^*\right)\right| \\ &\quad + \max_j \left|\frac{1}{\sqrt{N}} e_j^T \left(\widehat{\Theta} - \Theta^*\right) X^T \widehat{W}^{-1}\varepsilon\right| + \max_j \left|\frac{1}{\sqrt{N}} e_j^T \Theta^* X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right| \\ &\leqslant \sqrt{N}\left(\left\|\Theta^*\Sigma_N - I_n\right\|_\infty + \mu\right)\left\|\hat{\beta} - \beta^*\right\|_1 + \left\|\widehat{\Theta} - \Theta^*\right\|_\infty \left\|\frac{1}{\sqrt{N}} X^T \widehat{W}^{-1}\varepsilon\right\|_\infty + \left\|\Theta^*\right\|_\infty \left\|\frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right\|_\infty.\end{aligned}$$

We've bounded $\left\|\hat{\beta} - \beta^*\right\|_1$ in Corollary 1. Therefore, it remains to bound $\left\|\Theta^*\Sigma_N - I_n\right\|_\infty$, $\left\|\widehat{\Theta} - \Theta^*\right\|_\infty$, $\left\|\frac{1}{\sqrt{N}} X^T \widehat{W}^{-1}\varepsilon\right\|_\infty$, and $\left\|\frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right\|_\infty$. We establish these bounds in the following lemmas and their proofs are contained in Appendix A.8, Appendix A.9, and Appendix A.10, respectively.

Lemma 8. Assume that Part (a) of Assumption 2 holds. We have

$$\left\|\Theta^*\Sigma_N - I_n\right\|_\infty \leqslant \frac{c}{L_1}\sqrt{\frac{\log n}{N}},$$

with probability at least $1 - \frac{2}{n}$.

Lemma 9. Assume that same conditions as Corollary 1 and Assumption 4 hold. Let

$$\mu \asymp \frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}$$

in Algorithm 1 and $N \geqslant s^2 L_2 \log n$. Then we have

$$\left\|\widehat{\Theta} - \Theta^*\right\|_\infty \leqslant ckM\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}\right),$$

with probability at least $1 - c' \exp(-c'' \log s)$.

Lemma 10. Assume that same conditions as Corollary 1 and Assumption 4 hold.

(a) We have

$$\left\|\frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right\|_\infty \leqslant c\frac{L_2 s^2 (\log n)^2}{\sqrt{N}},$$

with probability at least $1 - c' \exp(-c'' \log s)$.

(b) We have

$$\left\|\frac{1}{\sqrt{N}} X^T \widehat{W}^{-1}\varepsilon\right\|_\infty \leqslant c\left(\frac{L_2 s^2 (\log n)^2}{\sqrt{N}} + \sqrt{\log n}\right),$$

with probability at least $1 - c' \exp(-c'' \log s)$.

Therefore, by above lemmas, we have

$$\max_j \left|\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)\right| \leqslant c\left[\frac{s \log n}{L_1}\sqrt{\frac{L_2}{N}} + \frac{L_2 s^2 \log n}{\sqrt{N}} + kM\left(\frac{\log n}{L_1\sqrt{N}} + s \log n\sqrt{\frac{L_2}{N}}\right) + \frac{L_2 M s^2 (\log n)^2}{\sqrt{N}}\right],$$

with probability at least $1 - c' \exp(-c'' \log s)$. Hence, under the given assumption that

$$\frac{L_2 M s^2 (\log n)^2}{\sqrt{N}} \to 0, \tag{20}$$

we have

$$\max_j \left|\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)\right| = o_{\mathbb{P}}(1). \tag{21}$$

On the other hand, we have

$$\max_j \left|e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j - e_j^T \Theta^* e_j\right| = \max_j \left|e_j^T \left(\Theta^* \Sigma_N - I_n\right)\Theta^* e_j\right| \leqslant \left\|\Theta^* \Sigma_N - I_n\right\|_\infty \left\|\Theta^*\right\|_\infty.$$

Therefore, by Lemma 8 and Assumption 5, we have

$$\mathbb{P}\left(\max_j \left|e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j - e_j^T \Theta^* e_j\right| \leqslant c\frac{M}{L_1}\sqrt{\frac{\log n}{N}}\right) \geqslant 1 - \frac{2}{n}.$$

Since (20) implies that

$$\frac{M}{L_1}\sqrt{\frac{\log n}{N}} \to 0,$$

so we have

$$\max_j \left|e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j - e_j^T \Theta^* e_j\right| = o_{\mathbb{P}}(1). \tag{22}$$

Hence, combining (21) and (22) yields the desired result in (18).

6.3.2 Step 2 of proof

By definition of $\check{\beta}$, we have

$$\frac{\sqrt{N}\left(\check{\beta}_j - \beta_j^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} = \frac{\frac{1}{\sqrt{N}} e_j^T \Theta^* X^T W^{-1} \varepsilon}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} - \frac{\sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}.$$

First, let

$$Z = \frac{\frac{1}{\sqrt{N}} e_j^T \Theta^* X^T W^{-1} \varepsilon}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}.$$

We show that Z is a standard normal random variable. Note that the characteristic function of Z is

$$\begin{aligned}
\mathbb{E}\left(e^{itZ}\right) &= \mathbb{E}\left[\exp\left(it \frac{\frac{1}{\sqrt{N}} e_j^T \Theta^* X^T W^{-1} \varepsilon}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}\right)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\exp\left(it \frac{\frac{1}{\sqrt{N}} e_j^T \Theta^* X^T W^{-1} \varepsilon}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}\right) \middle| X\right]\right] \\
&= \mathbb{E}_X\left[\exp\left(-\frac{t^2}{2}\right) \middle| X\right] = \exp\left(-\frac{t^2}{2}\right).
\end{aligned}$$

Therefore, we obtain the desired result. Next, let

$$\Psi_j = \frac{\sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}.$$

Our goal is to show that $\max_j |\Psi_j| = o_{\mathbb{P}}(1)$ under given assumptions. We have

$$\max_j \left|\sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right)\right| \leqslant \sqrt{N} \left\|\Theta^* \Sigma_N - I_n\right\|_\infty \left\|\hat{\beta} - \beta^*\right\|_1.$$

Therefore, by Lemma 8 and Corollary 1, we have

$$\max_j \left|\sqrt{N} e_j^T \left(\Theta^* \Sigma_N - I_n\right)\left(\hat{\beta} - \beta^*\right)\right| \leqslant c \frac{s \log n}{L_1} \sqrt{\frac{L_2}{N}}, \tag{23}$$

with probability at least $1 - c' \exp\left(-c'' \log s\right)$. Hence, combining (22) and (23), we have $\max_j |\Psi_j| = o_{\mathbb{P}}(1)$ under the given high-dimensional regime. Finally, applying Slutsky's theorem yields the desired result in (19).

6.3.3 Conclusion of proof

We have

$$\sqrt{N}\left(\tilde{\beta}_j - \beta_j^*\right) = \sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j} \left[\frac{\sqrt{N}\left(\check{\beta}_j - \beta_j^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} + \frac{\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}}\right].$$

By (18), (19), and Slutsky's theorem, under given assumptions, we have

$$\frac{\sqrt{N}\left(\check{\beta}_j - \beta_j^*\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} + \frac{\sqrt{N}\left(\tilde{\beta}_j - \check{\beta}_j\right)}{\sqrt{e_j^T \Theta^* \Sigma_N (\Theta^*)^T e_j}} \xrightarrow{d} N(0,1). \tag{24}$$

Then combining (22) and (24), and applying Slutsky's theorem again, we have

$$\sqrt{N}\left(\tilde{\beta}_j - \beta_j^*\right) \xrightarrow{d} N\left(0, e_j^T \Theta^* e_j\right).$$

Therefore, the proof is complete.

References

[1] P. Bickel. Using residuals robustly I: tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, 6(2):266–291, 03 1978.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

[3] R. Carroll. Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10(4):1224–1233, 12 1982.

[4] R. Carroll and D. Ruppert. On robust tests for heteroscedasticity. *Annals of Statistics*, 9(1):206–210, 01 1981.

[5] R. Carroll and D. Ruppert. A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, 77(380):878–882, 1982.

[6] R. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, 10(2):429–441, 06 1982.

[7] R. Carroll and D. Ruppert. *Transformation and Weighting in Regression*, volume 30. CRC Press, 1988.

[8] R. Carroll, J. Wu, and D. Ruppert. The effect of estimating weights in weighted least squares. *Journal of the American Statistical Association*, 83(404):1045–1054, 1988.

[9] M. Cattaneo, M. Jansson, and W. Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018.

[10] M. Davidian and R. Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82(400):1079–1091, 1987.

[11] Z. Daye, J. Chen, and H. Li. High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, 68(1):316–326, 2012.

[12] J. Janková and S. van de Geer. Semiparametric efficiency bounds for high-dimensional models. *Annals of Statistics*, 46(5):2336–2359, 10 2018.

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

[14] J. Jia, K. Rohe, and B. Yu. The Lasso under Poisson-like heteroscedasticity. *Statistica Sinica*, pages 99–118, 2013.

[15] J. Jobson and W. Fuller. Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association*, 75(369):176–181, 1980.

[16] K. Knight and W. Fu. Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 10 2000.

[17] G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

[18] G. Seber and A. Lee. *Linear Regression Analysis*, volume 329. John Wiley & Sons, 2012.

[19] J. Sharpnack and M. Kolar. Mean and variance estimation in high-dimensional heteroscedastic models with non-convex penalties. *arXiv preprint arXiv:*1410.7874, 2014.

[20] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 06 2014.

[21] J. Wagener and H. Dette. The adaptive Lasso in high-dimensional sparse heteroscedastic models. *Mathematical Methods of Statistics*, 22(2):137–154, 2013.

[22] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[23] D. Wang and P. Loh. Adaptive estimation and statistical inference for high-dimensional graph-based linear models. *arXiv preprint arXiv:*2001.10679, 2020.

[24] C. Zhang and S. Zhang. Confidence intervals for low-dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

A Proofs

In this appendix, we provide proofs of various lemmas stated in Section 6 of the paper. Furthermore, see Figure A.1 for a flow diagram which illustrates the dependence structure among lemmas, theorems, and corollaries in the paper.

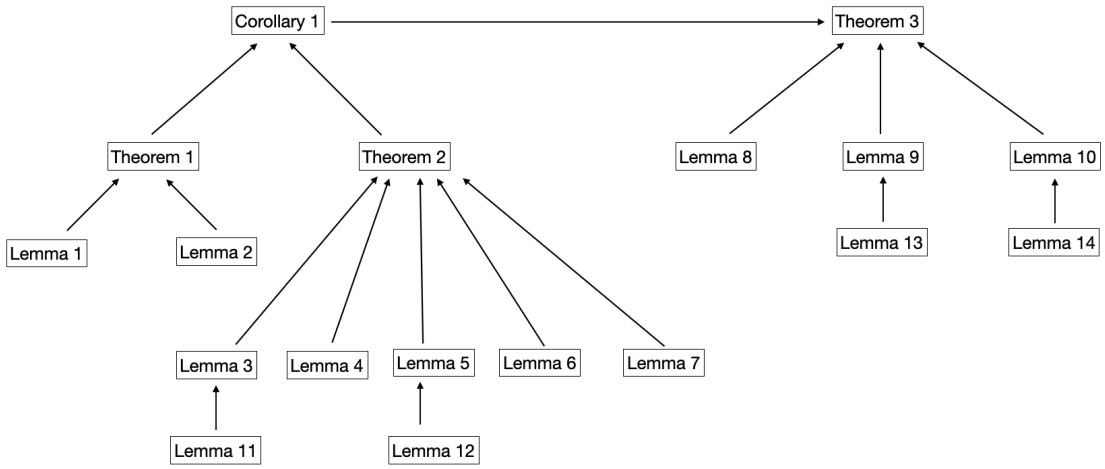


Figure A.1: Flowchart proofs. $A \to B$ reads as "result B is partially dependent on result A".

A.1 Proof of Lemma 1

First, we write

$$\|\varepsilon^T X\|_\infty = \max_{i=1,\dots,n} \left|\varepsilon^T X e_i\right|.$$

Note that the conditional distribution of $\varepsilon^T X e_i$ given X is

$$\varepsilon^T X e_i \mid X \sim N\left(0, e_i^T X^T W X e_i\right).$$

Furthermore, by our assumption on W, we have

$$e_i^T X^T W X e_i \leqslant L_2 \|X e_i\|_2^2. \tag{A.1}$$

Applying Lemma 13 in [23], we have

$$\mathbb{P}\left(\|X e_i\|_2^2 \leqslant N\left(\lambda_{\max}\left(\Sigma_x\right) + 4\sigma_x^2\right)\right) \geqslant 1 - \exp\left(-Nc\right), \tag{A.2}$$

where $c > 0$ is a constant. Combining (A.1) and (A.2), we have

$$\mathbb{P}\left(e_i^T X^T W X e_i \leqslant L_2 N\left(\lambda_{\max}\left(\Sigma_x\right) + 4\sigma_x^2\right)\right) \geqslant 1 - \exp\left(-Nc\right). \tag{A.3}$$

Next, for $i = 1, ..., n$, we define the following event

$$\mathcal{E}_i = \left\{X \in \mathbb{R}^{N\times n} : e_i^T X^T W X e_i \leqslant L_2 N\left(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2\right)\right\}.$$

Conditioning on $\mathcal{E}_i$ and applying the standard tail bound of Gaussian random variable, for $t > 0$, we have

$$\mathbb{P}\left(|\varepsilon^T X e_i| > t \mid \mathcal{E}_i\right) \leqslant 2\exp\left\{-\frac{t^2}{2L_2 N\left(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2\right)}\right\}. \tag{A.4}$$

Therefore, combining (A.3) and (A.4), we have

$$\mathbb{P}\left(|\varepsilon^T X e_i| > t\right) \leqslant 2\exp\left\{-\frac{t^2}{2L_2 N\left(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2\right)}\right\} + \exp\left(-Nc\right).$$

Finally, applying a union bound and letting $t \geqslant 2(\lambda_{\max}(\Sigma_x) + 4\sigma_x^2)^{1/2} L_2^{1/2} (N \log n)^{1/2}$ yield

$$\mathbb{P}\left(\|\varepsilon^T X\|_\infty > t\right) \leqslant 2\exp\left(-\log n\right) + \exp\left(\log n - cN\right),$$

implying the desired result in the lemma.

A.2 Proof of Lemma 2

First, we assume the following condition: Let $\mathbb{L}_0(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$. For $v \in \mathbb{L}_0(2s)$, we have

$$\left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| \leqslant \frac{\lambda_{\min}(\Sigma_x)}{150}. \tag{A.5}$$

Note that for $v \in \mathbb{R}^n$, we have $v^T \Sigma_x v \geqslant \lambda_{\min}(\Sigma_x)\|v\|_2^2$. Therefore, when condition (A.5) holds, by Lemma 11 in [23], we have

$$\frac{1}{N}\|Xv\|_2^2 \geqslant \frac{\lambda_{\min}(\Sigma_x)}{2}\|v\|_2^2 \tag{A.6}$$

for $v \in \mathbb{R}^n$ and $\|v_{S^c}\|_1 \leqslant 3\|v_S\|_1$.

Next, we use a discretization argument to verify that condition (A.5) holds with high probability. For an index subset $I \subset [n]$ with $|I| \leqslant 2s$, we define

$$S_I = \{v \in \mathbb{R}^n; \|v\|_2 \leqslant 1, \text{support}(v) \subset I\}.$$

Then we have $\mathbb{L}_0(2s) = \cup_{|I| \leqslant 2s} S_I$. For a fixed S_I, let $\mathcal{A}_I$ be a $\frac{1}{3}$-cover of S_I with $|\mathcal{A}_I| \leqslant 9^{2s}$. Therefore, for $v \in S_I$, there exists $a_v \in \mathcal{A}_I$, such that $\|\Delta_v\|_2 = \|v - a_v\|_2 \leqslant \frac{1}{3}$. Hence,

$$\begin{aligned}
\left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| &= \left| (\Delta_v + a_v)^T \left(\frac{X^T X}{N} - \Sigma_x \right) (\Delta_v + a_v) \right| \\
&\leqslant \left| \Delta_v^T \left(\frac{X^T X}{N} - \Sigma_x \right) \Delta_v \right| + 2\left| a_v^T \left(\frac{X^T X}{N} - \Sigma_x \right) \Delta_v \right| + \left| a_v^T \left(\frac{X^T X}{N} - \Sigma_x \right) a_v \right| \\
&\leqslant \frac{1}{9} \sup_{v \in S_I} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| + \frac{2}{3} \sup_{v \in S_I} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| + \sup_{v \in \mathcal{A}_I} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right|.
\end{aligned}$$

Rearranging terms, we have

$$\sup_{v \in S_I} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| \leqslant \frac{9}{2} \sup_{v \in \mathcal{A}_I} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right|.$$

Applying Lemma 13 in [23] and taking union bounds, we have

$$\mathbb{P}\left(\sup_{v \in \mathbb{L}_0(2s)} \left| v^T \left(\frac{X^T X}{N} - \Sigma_x \right) v \right| \geqslant \frac{\lambda_{\min}(\Sigma_x)}{150} \right) \leqslant 2\exp\left(cs \log n - c' N\right),$$

where $c > 0$ and $c' > 0$ are constants.

Therefore, we obtain that inequality (A.6) holds at least with the same probability.

A.3 Proof of Lemma 3

First, we show Part (a). By Lemma 11, it suffices to show that $\|\hat{z}_{S^c}\|_\infty < 1$. Note that we have

$$\hat{z}_{S^c} = X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \hat{z}_S + \frac{1}{\lambda N} X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T\right) \varepsilon,$$

which implies

$$\|\hat{z}_{S^c}\|_\infty \leqslant \left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \hat{z}_S \right\|_\infty + \left\| \frac{1}{\lambda N} X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T\right) \varepsilon \right\|_\infty.$$

Therefore, it suffices to bound $\left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \hat{z}_S \right\|_\infty$ and $\left\| \frac{1}{\lambda N} X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T\right) \varepsilon \right\|_\infty$. For the first term, we have

$$\begin{aligned} \left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} \hat{z}_S \right\|_\infty &= \max_j \left| e_j^T X_{S^c}^T X_S (X_S^T X_S)^{-1} \hat{z}_S \right| \\ &\leqslant \max_j \left\| e_j^T X_{S^c}^T X_S (X_S^T X_S)^{-1} \right\|_1 \|\hat{z}_S\|_\infty \leqslant \alpha, \end{aligned}$$

where the last inequality follows from (15) and $\|\hat{z}_S\|_\infty \leqslant 1$. For the second term, we have

$$\left\| \frac{1}{\lambda N} X_{S^c}^T \left(I_N - X_S (X_S^T X_S)^{-1} X_S^T\right) \varepsilon \right\|_\infty \leqslant \frac{1-\alpha}{2},$$

where the inequality follows from (16). Therefore, we have

$$\|\hat{z}_{S^c}\|_\infty \leqslant \alpha + \frac{1-\alpha}{2} = \frac{1+\alpha}{2} < 1.$$

Hence, Lemma 11 implies that there is a unique solution to (3) and its support $\widehat{S}$ is contained in S.

Next, we show Part (b). It remains to show that the true support S is contained in the set $\widehat{S}$. Note that we have

$$\hat{\beta}_S - \beta_S^* = (X_S^T X_S)^{-1} X_S^T \varepsilon - \lambda N (X_S^T X_S)^{-1} \hat{z}_S.$$

Therefore, we have

$$\begin{aligned} \left\| \hat{\beta}_S - \beta_S^* \right\|_\infty &\leqslant \left\| (X_S^T X_S)^{-1} X_S^T \varepsilon \right\|_\infty + \|\lambda N (X_S^T X_S)^{-1} \hat{z}_S\|_\infty \\ &\leqslant \left\| \left(\frac{X_S^T X_S}{N}\right)^{-1} X_S^T \frac{\varepsilon}{N} \right\|_\infty + \lambda \left\| \left(\frac{X_S^T X_S}{N}\right)^{-1} \right\|_\infty = B(\lambda, X_S, \varepsilon). \end{aligned}$$

By (17), we have $\hat{\beta}_i \neq 0$ for $i \in S$, which implies $S \subseteq \widehat{S}$. Putting together the pieces, we conclude that $S = \widehat{S}$.

Therefore, the proof is complete.

A.4 Proof of Lemma 4

Since $X \in \mathbb{R}^{N\times n}$ is a row-wise (σ_x, Σ_x)-sub-Gaussian random matrix, so $X_S \in \mathbb{R}^{N\times s}$ is also a row-wise $(\sigma_x, (\Sigma_x)_{SS})$-sub-Gaussian random matrix, where $(\Sigma_x)_{SS} \in \mathbb{R}^{s\times s}$ is a submatrix of Σ_x with rows and columns are restricted to S. By Lemma 13 in [23], we have

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{N}X_S^T X_S\right) \geqslant \frac{1}{2}\lambda_{\min}\left((\Sigma_x)_{SS}\right)\right) \geqslant 1 - 2\exp\left(cs - c'N\right).$$

Furthermore, we have $\lambda_{\min}\left((\Sigma_x)_{SS}\right) \geqslant \lambda_{\min}(\Sigma_x)$. Hence, we have

$$\lambda_{\min}\left(\frac{1}{N}X_S^T X_S\right) \geqslant \frac{1}{2}\lambda_{\min}(\Sigma_x),$$

with probability at least $1 - 2\exp\left(-cN\right)$ assuming the scaling $N \gtrsim s$. Therefore, the proof is complete.

A.5 Proof of Lemma 5

Let $\widehat{\Gamma} = \frac{1}{N}X^T X$. Then we have

$$\begin{aligned}
\left\|\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty &\leqslant \left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty \left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty \\
&+ \left\|\left[\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right](\Sigma_x)_{SS}^{-1}\right\|_\infty + \left\|(\Sigma_x)_{SS}\left[\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right]\right\|_\infty \\
&\leqslant \left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty \left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty \\
&+ \left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty \left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty + \left\|(\Sigma_x)_{SS}\right\|_\infty \left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty.
\end{aligned}$$

Therefore, it suffices to bound $\left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty$, $\left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty$, $\left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty$, and $\left\|(\Sigma_x)_{SS}\right\|_\infty$.

First, we bound the term $\left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty$. We have

$$\begin{aligned}
\left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty &= \max_{j\in S^c}\left\|\widehat{\Gamma}_{jS} - (\Sigma_x)_{jS}\right\|_1 \\
\leqslant s \times \max_{j\in S^c}\left\|\widehat{\Gamma}_{jS} - (\Sigma_x)_{jS}\right\|_\infty &= s \times \max_{j\in S^c, k\in S}\left|\widehat{\Gamma}_{jk} - (\Sigma_x)_{jk}\right|.
\end{aligned}$$

By Part (a) of Lemma 12, we have

$$\mathbb{P}\left(\left\|\widehat{\Gamma}_{S^cS} - (\Sigma_x)_{S^cS}\right\|_\infty \leqslant cs\sqrt{\frac{\log[s(n-s)]}{N}}\right) \geqslant 1 - c'\exp\left(-c''\log\left[s(n-s)\right]\right).$$

Next, we bound the second term $\left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty$. We have

$$\begin{aligned}
\left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty &= \max_{j\in S}\left\|\left(\widehat{\Gamma}_{SS}^{-1}\right)_{j\cdot} - \left((\Sigma_x)_{SS}^{-1}\right)_{j\cdot}\right\|_1 \\
\leqslant s \times \max_{j\in S}\left\|\left(\widehat{\Gamma}_{SS}^{-1}\right)_{j\cdot} - \left((\Sigma_x)_{SS}^{-1}\right)_{j\cdot}\right\|_\infty &= s \times \left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty,
\end{aligned}$$

where $\left(\widehat{\Gamma}_{SS}^{-1}\right)_{j\cdot}$ and $\left((\Sigma_x)_{SS}^{-1}\right)_{j\cdot}$ are jth row of $\widehat{\Gamma}_{SS}^{-1}$ and $(\Sigma_x)_{SS}^{-1}$, respectively. Therefore, by Part (b) of Lemma 12, we have

$$\mathbb{P}\left(\left\|\widehat{\Gamma}_{SS}^{-1}-(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant cs\sqrt{\frac{s}{N}}\right) \geqslant 1-c'\exp\left(-c''s\right).$$

Furthermore, for the last two terms, we have

$$\left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant s\left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant s\left\|(\Sigma_x)_{SS}^{-1}\right\|_{op} \leqslant \frac{s}{\lambda_{\min}(\Sigma_x)},$$

and

$$\left\|(\Sigma_x)_{SS}\right\|_\infty \leqslant s\left\|(\Sigma_x)_{SS}\right\|_\infty \leqslant s\left\|(\Sigma_x)_{SS}\right\|_{op} \leqslant \lambda_{\max}(\Sigma_x)s.$$

Putting together the pieces, we conclude that

$$\left\|\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}-(\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant c\left(\sqrt{\frac{s^5\log[s(n-s)]}{N^2}}+s^2\sqrt{\frac{\log[s(n-s)]}{N}}+s^2\sqrt{\frac{s}{N}}\right),$$

with probability at least $1-c'\exp\left(-c''s\right)-C'\exp\left(-C''\log\left[s(n-s)\right]\right)$. Therefore, when

$$N^2 \gtrsim \frac{s^5\log[s(n-s)]}{\alpha^2}, \quad N \gtrsim \frac{s^4\log[s(n-s)]}{\alpha^2}, \quad \text{and} \quad N \gtrsim \frac{s^5}{\alpha^2},$$

we have

$$\left\|\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}-(\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant \frac{\alpha}{2},$$

with at least the same probability. Hence, we have

$$\left\|\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}\right\|_\infty \leqslant \left\|(\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty + \left\|\widehat{\Gamma}_{S^cS}\widehat{\Gamma}_{SS}^{-1}-(\Sigma_x)_{S^cS}(\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant \frac{\alpha}{2}+\frac{\alpha}{2}=\alpha,$$

with at least the same probability. Therefore, the proof is complete.

A.6 Proof of Lemma 6

First, we write

$$\left\|X_{S^c}^T\left(I_N-X_S(X_S^TX_S)^{-1}X_S^T\right)\frac{\varepsilon}{N}\right\|_\infty = \max_{1\leqslant j\leqslant n-s}\left|e_j^TX_{S^c}^T\left(I_N-X_S(X_S^TX_S)^{-1}X_S^T\right)\frac{\varepsilon}{N}\right|.$$

Let $\Pi_{S^\perp}(X)=I_N-X_S(X_S^TX_S)^{-1}X_S^T$. Then we have

$$e_j^TX_{S^c}^T\Pi_{S^\perp}(X)\frac{\varepsilon}{N} \mid X \sim N\left(0,\frac{1}{N^2}e_j^TX_{S^c}^T\Pi_{S^\perp}(X)W\Pi_{S^\perp}(X)X_{S^c}e_j\right).$$

By our assumption on W and the fact that $\Pi_{S^\perp}(X)$ is an orthogonal projection matrix, we have

$$\frac{1}{N^2}e_j^TX_{S^c}^T\Pi_{S^\perp}(X)W\Pi_{S^\perp}(X)X_{S^c}e_j \leqslant \frac{1}{N^2}\left\|W\right\|_{op}\left\|\Pi_{S^\perp}(X)\right\|_{op}^2\left\|X_{S^c}e_j\right\|_2^2 \leqslant \frac{1}{N^2}L_2\left\|X_{S^c}e_j\right\|_2^2.$$

Applying Lemma 13 in [23], we have

$$\mathbb{P}\left(\frac{L_2}{N^2}\|X_{S^c}e_j\|_2^2 \leqslant \frac{L_2}{N}\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)\right) \geqslant 1-\exp\left(-Nc\right),$$

where $c>0$ is a constant.

Next, for $j=1,...,n-s$, we define the following event

$$\mathcal{E}_j = \left\{X\in\mathbb{R}^{N\times n} : \frac{1}{N^2}e_j^T X_{S^c}^T \Pi_{S^\perp}(X) W \Pi_{S^\perp}(X) X_{S^c} e_j \leqslant \frac{L_2}{N}\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)\right\}.$$

Conditioning on $\mathcal{E}_j$ and applying the standard tail bound of Gaussian random variable, for $t>0$, we have

$$\mathbb{P}\left(\left|e_j^T X_{S^c}^T \Pi_{S^\perp}(X)\frac{\varepsilon}{N}\right| > t \mid \mathcal{E}_j\right) \leqslant 2\exp\left\{-\frac{Nt^2}{2L_2\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)}\right\}.$$

Therefore, we have

$$\mathbb{P}\left(\left|e_j^T X_{S^c}^T \Pi_{S^\perp}(X)\frac{\varepsilon}{N}\right| > t\right) \leqslant 2\exp\left\{-\frac{Nt^2}{2L_2\left(\lambda_{\max}(\Sigma_x)+4\sigma_x^2\right)}\right\} + \exp\left(-Nc\right).$$

Finally, applying a union bound and letting $t \geqslant 2\sqrt{\lambda_{\max}(\Sigma_x)+4\sigma_x^2}\sqrt{\frac{L_2\log(n-s)}{N}}$ yields the desired result in the lemma. Therefore, the proof is complete.

A.7 Proof of Lemma 7

It suffices to bound $\left\|\left(\frac{X_S^T X_S}{N}\right)^{-1} X_S^T \frac{\varepsilon}{N}\right\|_\infty$ and $\lambda\left\|\left(\frac{X_S^T X_S}{N}\right)^{-1}\right\|_\infty$. We use a similar argument with Lemma 6 to bound the first term, and conclude that

$$\mathbb{P}\left(\left\|\widehat{\Gamma}_{SS}^{-1} X_S^T \frac{\varepsilon}{N}\right\|_\infty > \sqrt{\frac{8}{\lambda_{\min}(\Sigma_x)}}\sqrt{\frac{L_2\log s}{N}}\right) \leqslant 2\exp\left(-\log s\right) + 2\exp\left(\log s + cs - c'N\right).$$

For the second term, we use a similar argument with Lemma 5, and conclude that

$$\lambda\left\|\left(\frac{X_S^T X_S}{N}\right)^{-1}\right\|_\infty \leqslant \lambda\left(\left\|(\Sigma_x)_{SS}^{-1}\right\|_\infty + cs\sqrt{\frac{s}{N}}\right),$$

with probability greater than $1-2\exp\left(-c's\right)-2\exp\left(c''s-C'N\right)$. Therefore, the proof is complete.

A.8 Proof of Lemma 8

Let $\Gamma = \Theta^*\Sigma_N - I_n$. Then the (j,k)-th entry of Γ is

$$\Gamma^{jk} = \frac{1}{N}\sum_{i=1}^N e_j^T\left(\Theta^*\frac{X_i X_i^T}{W_i} - I_n\right)e_k = \frac{1}{N}\sum_{i=1}^N \Gamma_i^{jk}.$$

Note that $\mathbb{E}(\Gamma_i^{jk})=0$, and

$$\|\Gamma_i^{jk}\|_{\psi_1} \leqslant 2\left\|e_j^T\Theta^*\frac{X_i X_i^T}{W_i}e_k\right\|_{\psi_1} \leqslant \frac{4}{L_1}\left\|e_j^T\Theta^* X_i\right\|_{\psi_2}\left\|X_i^T e_k\right\|_{\psi_2} \leqslant \frac{4c}{L_1}\lambda_{\max}(\Theta^*)\sigma_x^2 = K.$$

Applying Lemma 12 in [23], we have for $t > 0$,

$$\mathbb{P}\left(\left|\Gamma^{jk}\right| \geqslant t\right) = \mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\Gamma_i^{jk}\right| \geqslant t\right) \leqslant 2\exp\left[-C_b N \min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right].$$

Taking a union bound, we have

$$\|\Gamma\|_\infty \leqslant c\frac{\lambda_{\max}\left(\Theta^*\right)\sigma_x^2}{L_1}\sqrt{\frac{\log n}{N}}, \tag{A.7}$$

with probability at least $1 - \frac{2}{n}$. Therefore, the proof is complete.

A.9 Proof of Lemma 9

Let $\widehat{\Theta}_j$ and Θ_j^* be the jth row of $\widehat{\Theta}$ and Θ^*, respectively. First, we show that on the event

$$\mathcal{E} = \left\{\left\|\widehat{\Theta}_j\right\|_1 \leqslant \|\Theta_j^*\|_1, 1 \leqslant j \leqslant n\right\},$$

we have

$$\left\|\widehat{\Theta} - \Theta^*\right\|_\infty \leqslant 8k\|\widehat{\Theta} - \Theta^*\|_\infty, \tag{A.8}$$

where k is defined in Assumption 5. Let $\xi = \left\|\widehat{\Theta} - \Theta^*\right\|_\infty$ and $r_j = \widehat{\Theta}_j - \Theta_j^*$. Then we have

$$r_j = \left[\left(\widehat{\Theta}_{ji}1_{\{|\widehat{\Theta}_{ji}|\geqslant 2\xi\}}; 1 \leqslant i \leqslant n\right) - \Theta_j^*\right] + \left(\widehat{\Theta}_{ji}1_{\{|\widehat{\Theta}_{ji}|<2\xi\}}; 1 \leqslant i \leqslant n\right) = r_j^1 + r_j^2.$$

Furthermore, on the event $\mathcal{E}$, we have

$$\left\|\Theta_j^*\right\|_1 - \left\|r_j^1\right\|_1 + \left\|r_j^2\right\|_1 \leqslant \left\|\Theta_j^* + r_j^1\right\|_1 + \left\|r_j^2\right\|_1 = \|\widehat{\Theta}_j\|_1 \leqslant \|\Theta_j^*\|_1,$$

where the first inequality follows from the triangle inequality. Therefore we have

$$\|r_j\|_1 \leqslant \left\|r_j^1\right\|_1 + \left\|r_j^2\right\|_1 \leqslant 2\left\|r_j^1\right\|_1. \tag{A.9}$$

Furthermore, we have

$$\begin{aligned}\left\|r_j^1\right\|_1 &\leqslant \sum_{i=1}^{n}|\widehat{\Theta}_{ji} - \Theta_{ji}^*|1_{\{|\widehat{\Theta}_{ji}|\geqslant 2\xi\}} + \sum_{i=1}^{n}|\Theta_{ji}^*|1_{\{|\widehat{\Theta}_{ji}|<2\xi\}} \\ &\leqslant \sum_{i=1}^{n}\xi 1_{\{|\Theta_{ji}^*|\geqslant\xi\}} + \sum_{i=1}^{n}|\Theta_{ji}^*|1_{\{|\Theta_{ji}^*|<3\xi\}} \leqslant \xi k + 3\xi k = 4\xi k,\end{aligned} \tag{A.10}$$

where the second inequality follows from the triangle inequality. Therefore, combining (A.9) and (A.9) yields the result in (A.8).

Next, we have

$$\begin{aligned}\|\widehat{\Theta} - \Theta^*\|_\infty &= \|\widehat{\Theta}(I_n - \widehat{\Sigma}_N\Theta^*) + (\widehat{\Theta}\widehat{\Sigma}_N - I_n)\Theta^*\|_\infty \\ &\leqslant \left\|\widehat{\Theta}\right\|_\infty\left\|I_n - \widehat{\Sigma}_N\Theta^*\right\|_\infty + \|\widehat{\Theta}\Sigma_N - I_n\|_\infty\left\|\Theta^*\right\|_1 \\ &\leqslant \left\|\widehat{\Theta}\right\|_\infty\left\|I_n - \widehat{\Sigma}_N\Theta^*\right\|_\infty + \mu M.\end{aligned}$$

We define the event

$$\mathcal{E}' = \left\{ X : \left\| \Theta^* \hat{\Sigma}_N - I_n \right\|_\infty \leqslant \mu \right\}.$$

Then on the event $\mathcal{E}'$, we have $\|\widehat{\Theta}_j\|_1 \leqslant \|\Theta_j^*\|_1$ for $1 \leqslant j \leqslant n$, which implies $\left\|\widehat{\Theta}\right\|_\infty \leqslant \|\Theta^*\|_\infty$. Therefore on the event $\mathcal{E}'$, we have

$$\|\widehat{\Theta} - \Theta^*\|_\infty \leqslant 2\mu M.$$

Furthermore, since event $\mathcal{E}'$ implies event $\mathcal{E}$, so on the event $\mathcal{E}'$, we have

$$\left\|\widehat{\Theta} - \Theta^*\right\|_\infty \leqslant 8k \left\|\widehat{\Theta} - \Theta^*\right\|_\infty \leqslant 16k\mu M.$$

Finally, letting

$$\mu = c\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}\right)$$

and applying Lemma 13, we have

$$\left\|\widehat{\Theta} - \Theta^*\right\|_\infty \leqslant ckM\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2 \log n}{N}}\right),$$

with probability at least $1 - c' \exp(-c'' \log s)$. Therefore, the proof is complete.

A.10 Proof of Lemma 10

First, we show Part (a). We have

$$\left\|\frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right\|_\infty = \max_{1\leqslant j\leqslant n}\left|\frac{1}{\sqrt{N}} e_j^T X^T \left(\widehat{W}^{-1} - W^{-1}\right)\varepsilon\right| = \max_{1\leqslant j\leqslant n}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\widehat{W}_i^{-1} - W_i^{-1}\right)\right|.$$

For a fixed j, using Taylor's theorem for $\widehat{W}_i^{-1} - W_i^{-1}$, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\widehat{W}_i^{-1} - W_i^{-1}\right) &= \frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T d_i(\beta^*) \\ &+ \frac{1}{2\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T H_i\left(\beta^* + c_i\left(\hat{\beta} - \beta^*\right)\right)\left(\hat{\beta} - \beta^*\right),\end{aligned}$$

where $d_i(\beta)$, $H_i(\beta)$, and c_i are defined in the proof of Lemma 13. Therefore, we obtain

$$\begin{aligned}\max_j\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\widehat{W}_i^{-1} - W_i^{-1}\right)\right| &\leqslant \max_j\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T d_i(\beta^*)\right| \\ &+ \max_j\left|\frac{1}{2\sqrt{N}}\sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T H_i\left(\beta^* + c_i\left(\hat{\beta} - \beta^*\right)\right)\left(\hat{\beta} - \beta^*\right)\right|\end{aligned}$$

Let $I_j = \frac{1}{\sqrt{N}} \sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T d_i(\beta^*)$. We have

$$\max_j |I_j| \leqslant \left\|\hat{\beta} - \beta^*\right\|_1 \max_j \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N e_j^T X_i \varepsilon_i d_i(\beta^*) \right\|_\infty .$$

By Corollary 1 and Part (a) of Lemma 14, we have

$$\mathbb{P}\left(\max_j |I_j| \leqslant cs \log n \sqrt{\frac{L_2}{N}} \right) \geqslant 1 - c' \exp\left(-c'' \log s\right). \tag{A.11}$$

Next, let $II_j = \frac{1}{2\sqrt{N}} \sum_{i=1}^N e_j^T X_i \varepsilon_i \left(\hat{\beta} - \beta^*\right)^T H_i \left(\beta^* + c_i \left(\hat{\beta} - \beta^*\right)\right) \left(\hat{\beta} - \beta^*\right)$. We have

$$\max_j |II_j| \leqslant \frac{1}{2} \left(\max_j \frac{1}{\sqrt{N}} \sum_{i=1}^N |X_{ij} \varepsilon_i| \right) \max_i \left| \left(\hat{\beta} - \beta^*\right)^T H_i \left(\beta^* + c_i \left(\hat{\beta} - \beta^*\right)\right) \left(\hat{\beta} - \beta^*\right) \right|.$$

If

$$\widehat{S} \subseteq S \quad \text{and} \quad \|\hat{\beta} - \beta^*\|_2 \leqslant c \sqrt{\frac{sL_2 \log n}{N}},$$

then we have

$$\begin{aligned}
\max_j |II_j| &\leqslant \frac{1}{2} \left(\max_j \frac{1}{\sqrt{N}} \sum_{i=1}^N |X_{ij} \varepsilon_i| \right) \left\|\hat{\beta} - \beta^*\right\|_2^2 \left\| H_i \left(\beta^* + c_i \left(\hat{\beta} - \beta^*\right)\right)_{SS} \right\|_{op} \\
&\leqslant \frac{1}{2} \left(\max_j \frac{1}{\sqrt{N}} \sum_{i=1}^N |X_{ij} \varepsilon_i| \right) \left\|\hat{\beta} - \beta^*\right\|_2^2 \left\| H_i \left(\beta^* + c_i \left(\hat{\beta} - \beta^*\right)\right)_{SS} \right\|_{F} \\
&\leqslant \frac{c}{2} \left(\max_j \frac{1}{\sqrt{N}} \sum_{i=1}^N |X_{ij} \varepsilon_i| \right) \frac{s^2 L_2 \log n}{N},
\end{aligned}$$

where the last inequality follows from Part (b) of Assumption 2 and Assumption 4. Therefore, by Part (b) of Lemma 14 and Corollary 1, we have

$$\mathbb{P}\left(\max_j |II_j| \leqslant c \frac{L_2 s^2 (\log n)^2}{\sqrt{N}} \right) \geqslant 1 - c' \exp\left(-c'' \log s\right). \tag{A.12}$$

Finally, combining (A.11) and (A.12), we have

$$\left\| \frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right) \varepsilon \right\|_\infty \leqslant c \frac{L_2 s^2 (\log n)^2}{\sqrt{N}},$$

with probability at least $1 - c' \exp\left(-c'' \log s\right)$.

Next, we show Part (b). We have

$$\frac{1}{\sqrt{N}} X^T \widehat{W}^{-1} \varepsilon = \frac{1}{\sqrt{N}} X^T \left(\widehat{W}^{-1} - W^{-1}\right) \varepsilon + \frac{1}{\sqrt{N}} X^T W^{-1} \varepsilon,$$

which implies that

$$\left\|\frac{1}{\sqrt{N}}X^T\widehat{W}^{-1}\varepsilon\right\|_\infty \leqslant \left\|\frac{1}{\sqrt{N}}X^T\left(\widehat{W}^{-1}-W^{-1}\right)\varepsilon\right\|_\infty + \left\|\frac{1}{\sqrt{N}}X^TW^{-1}\varepsilon\right\|_\infty.$$

Therefore, by Part (a) of this lemma and Part (c) of Lemma 14, we have

$$\left\|\frac{1}{\sqrt{N}}X^T\widehat{W}^{-1}\varepsilon\right\|_\infty \leqslant c\left(\frac{L_2s^2(\log n)^2}{\sqrt{N}}+\sqrt{\log n}\right),$$

with probability at least $1-c'\exp\left(-c''\log s\right)$. Therefore, the proof is complete.

B Primal-dual witness construction

In this appendix, we introduce the primal-dual witness construction [22], which is used for establishing the variable selection consistency. PDW includes the following three steps:

(1) Set $\hat{\beta}_{S^c}=0$.

(2) Determine $\hat{\beta}_S\in\mathbb{R}^s$ by solving

$$\hat{\beta}_S\in\underset{\beta_S\in\mathbb{R}^s}{\operatorname{argmin}}\quad\frac{1}{2N}\|y-X_S\beta_S\|_2^2+\lambda\|\beta_S\|_1. \tag{B.1}$$

Then choose $\hat{z}_S\in\partial\|\hat{\beta}_S\|_1$ such that

$$\frac{1}{N}X_S^T\left(X_S\hat{\beta}_S-y\right)+\lambda\hat{z}_S=0.$$

(3) Solve for $\hat{z}_{S^c}\in\mathbb{R}^{n-s}$ via the equation

$$\frac{1}{N}X^T\left(X\hat{\beta}-y\right)+\lambda\hat{z}=0,$$

and check whether or not the condition $\|\hat{z}_{S^c}\|_\infty<1$ holds.

The PDW construction has the following important consequence:

Lemma 11. If there exists a constant $c_{\min}>0$ such that $\lambda_{\min}(\frac{1}{N}X_S^TX_S)\geqslant c_{\min}$, then the success of the PDW construction implies that $\left(\hat{\beta}_S^T,0\right)^T\in\mathbb{R}^n$ is the unique optimal solution to (3).

Proof. Note if the PDW construction succeeds, then by zero subgradient condition, $\hat{\beta}=(\hat{\beta}_S,0)$ is an optimal solution to (3) with associated subgradient vector $\hat{z}$ satisfying $\|\hat{z}_{S^c}\|_\infty<1$ and $\hat{z}^T\hat{\beta}=\|\hat{\beta}\|_1$.

Let $\bar{\beta}$ be any other optimal solution to (3). Then we have

$$\frac{1}{2N}\left\|y-X\hat{\beta}\right\|_2^2+\lambda\hat{z}^T\hat{\beta}=\frac{1}{2N}\|y-X\bar{\beta}\|_2^2+\lambda\|\bar{\beta}\|_1.$$

Therefore, we have

$$\frac{1}{2N}\left\|y-X\hat{\beta}\right\|_2^2-\lambda\hat{z}^T\left(\bar{\beta}-\hat{\beta}\right)=\frac{1}{2N}\|y-X\bar{\beta}\|_2^2+\lambda\left(\|\bar{\beta}\|_1-\hat{z}^T\bar{\beta}\right).$$

Furthermore, by zero subgradient condition, we have

$$\lambda \hat{z} = -\frac{1}{N} X^T \left(X\hat{\beta} - y \right),$$

which implies that

$$\frac{1}{2N} \left\| y - X\hat{\beta} \right\|_2^2 + \frac{1}{N} \left(X\hat{\beta} - y \right)^T X \left(\bar{\beta} - \hat{\beta} \right) - \frac{1}{2N} \left\| y - X\bar{\beta} \right\|_2^2 = \lambda \left(\|\bar{\beta}\|_1 - \hat{z}^T \bar{\beta} \right) \leqslant 0.$$

Therefore, we have $\|\bar{\beta}\|_1 \leqslant \hat{z}^T \bar{\beta}$. On the other hand, we have

$$\hat{z}^T \bar{\beta} \leqslant \|\hat{z}\|_\infty \|\bar{\beta}\|_1 \leqslant \|\bar{\beta}\|_1.$$

Combining above arguments, we have $\|\bar{\beta}\|_1 = \hat{z}^T \bar{\beta}$. Since $\|\hat{z}_{S^c}\|_\infty < 1$, so we have $\bar{\beta}_j = 0$ for $j \in S^c$. Therefore, all optimal solution are supported only on the subset S, and hence can be obtained by solving the problem (B.1). Given the condition stated in the lemma, (B.1) is strictly convex, and so has a unique minimizer. Therefore, the proof is complete. □

C Additional supporting lemmas

Lemma 12. Assume that Part (a) of Assumption 2 holds. Let $\widehat{\Gamma} = \frac{1}{N} X^T X$.

(a) We have

$$\left\| \widehat{\Gamma}_{S^c S} - (\Sigma_x)_{S^c S} \right\|_\infty \leqslant c \sqrt{\frac{\log[s(n-s)]}{N}},$$

with probability at least $1 - c' \exp\left(-c'' \log\left[s(n-s)\right]\right)$.

(b) Under the sample size scaling $N \gtrsim s$, we have

$$\left\| \widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1} \right\|_\infty \leqslant \frac{2c}{\lambda_{\min}^2 (\Sigma_x)} \sqrt{\frac{s}{N}},$$

with probability at least $1 - c' \exp\left(-c'' s\right)$.

Proof. First, we show Part (a). For a fixed $j \in S^c$ and a fixed $k \in S$, we have

$$\left| \widehat{\Gamma}_{jk} - (\Sigma_x)_{jk} \right| = \left| \frac{1}{N} \sum_{i=1}^N \left[X_{ij} X_{ik} - (\Sigma_x)_{jk} \right] \right|,$$

and

$$\|X_{ij} X_{ik} - (\Sigma_x)_{jk}\|_{\psi_1} \leqslant 2 \|X_{ij} X_{ik}\|_{\psi_1} \leqslant 4 \|X_{ij}\|_{\psi_2} \|X_{ik}\|_{\psi_2} \leqslant 4c\sigma_x^2.$$

Then by Lemma 12 in [23], for $0 < t \leqslant K = 4c\sigma_x^2$, we have

$$\mathbb{P}\left(\left| \widehat{\Gamma}_{jk} - (\Sigma_x)_{jk} \right| \geqslant t \right) \leqslant 2 \exp\left(-C_b N \frac{t^2}{K^2} \right).$$

Applying a union bound, we further have

$$\mathbb{P}\left(\left\| \widehat{\Gamma}_{S^c S} - (\Sigma_x)_{S^c S} \right\|_\infty \geqslant t \right) \leqslant 2s(n-s) \exp\left(-C_b N \frac{t^2}{K^2} \right).$$

Therefore, letting $t = c\sqrt{\frac{\log[s(n-s)]}{N}}$ yields the desired result in Part (a).

Next, we show Part (b). Since

$$\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1} = \widehat{\Gamma}_{SS}^{-1}\left[(\Sigma_x)_{SS} - \widehat{\Gamma}_{SS}\right](\Sigma_x)_{SS}^{-1},$$

so we have

$$\left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty = \max_{j,k}\left|e_j^T\widehat{\Gamma}_{SS}^{-1}\left((\Sigma_x)_{SS} - \widehat{\Gamma}_{SS}\right)(\Sigma_x)_{SS}^{-1}e_k\right|$$
$$\leqslant \left\|\widehat{\Gamma}_{SS}^{-1}\right\|_{op}\left\|(\Sigma_x)_{SS} - \widehat{\Gamma}_{SS}\right\|_{op}\left\|(\Sigma_x)_{SS}^{-1}\right\|_{op} = \frac{1}{\lambda_{\min}\left(\widehat{\Gamma}_{SS}\right)}\left\|(\Sigma_x)_{SS} - \widehat{\Gamma}_{SS}\right\|_{op}\frac{1}{\lambda_{\min}\left((\Sigma_x)_{SS}\right)}.$$

By Lemma 13 in [23], we have

$$\mathbb{P}\left(\left\|(\Sigma_x)_{SS} - \widehat{\Gamma}_{SS}\right\|_{op} \geqslant c\sqrt{\frac{s}{N}}\right) \leqslant 2\exp\left(-c's\right).$$

Then combining Lemma 4 and the fact that $\lambda_{\min}\left((\Sigma_x)_{SS}\right) \geqslant \lambda_{\min}(\Sigma_x)$, we have

$$\mathbb{P}\left(\left\|\widehat{\Gamma}_{SS}^{-1} - (\Sigma_x)_{SS}^{-1}\right\|_\infty \leqslant \frac{2c}{\lambda_{\min}^2(\Sigma_x)}\sqrt{\frac{s}{N}}\right) \geqslant 1 - c'\exp\left(-c''s\right).$$

Hence, we obtain the result in Part (b).

□

Lemma 13. Assume that same conditions as Corollary 1 and Assumption 4 hold. Under the sample size scaling $N \geqslant s^2 L_2 \log n$, we have

$$\left\|\Theta^*\widehat{\Sigma}_N - I_n\right\|_\infty \leqslant c\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}} + s\sqrt{\frac{L_2\log n}{N}}\right),$$

with probability at least $1 - c'\exp\left(-c''\log s\right)$.

Proof. We have

$$\left\|\Theta^*\widehat{\Sigma}_N - I_n\right\|_\infty = \left\|(\Theta^*\Sigma_N - I_n) + \Theta^*\left(\widehat{\Sigma}_N - \Sigma_N\right)\right\|_\infty \leqslant \left\|\Theta^*\Sigma_N - I_n\right\|_\infty + \left\|\Theta^*\left(\widehat{\Sigma}_N - \Sigma_N\right)\right\|_\infty$$
$$\leqslant \left\|\Theta^*\Sigma_N - I_n\right\|_\infty + \left\|\frac{1}{N}\sum_{i=1}^N\left(\widehat{W}_i^{-1} - W_i^{-1}\right)\Theta^*\left(X_iX_i^T - \Sigma_x\right)\right\|_\infty + \left\|\frac{1}{N}\sum_{i=1}^N\left(\widehat{W}_i^{-1} - W_i^{-1}\right)\Theta^*\Sigma_x\right\|_\infty$$
$$\leqslant \left\|\Theta^*\Sigma_N - I_n\right\|_\infty + \left\|\widehat{W}^{-1} - W^{-1}\right\|_\infty\left[\max_{j,k}\left(\frac{1}{N}\sum_{i=1}^N\left|e_j^T\Theta^*\left(X_iX_i^T - \Sigma_x\right)e_k\right|\right) + \lambda_{\max}(\Theta^*)\lambda_{\max}(\Sigma_x)\right].$$

We've bounded $\left\|\Theta^*\Sigma_N - I_n\right\|_\infty$ in Lemma 8. Furthermore, using the Bernstein-type inequality and taking a union bound, we have

$$\mathbb{P}\left(\max_{j,k}\left(\frac{1}{N}\sum_{i=1}^N\left|e_j^T\Theta^*\left(X_iX_i^T - \Sigma_x\right)e_k\right|\right) > 4\lambda_{\max}(\Theta^*)\sigma_x^2 + c\sqrt{\frac{\log n}{N}}\right) \leqslant \frac{1}{n}. \tag{C.1}$$

Therefore, it remains to bound $\left\|\widehat{W}^{-1}-W^{-1}\right\|_{\infty}$. First, by Taylor's theorem, we have

$$\frac{1}{g(X_i,\hat{\beta})}-\frac{1}{g(X_i,\beta^*)}=\left(\hat{\beta}-\beta^*\right)^T d_i\left(\beta^*\right)+\frac{1}{2}\left(\hat{\beta}-\beta^*\right)^T H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)\left(\hat{\beta}-\beta^*\right),$$

where

$$c_i\in(0,1),\quad d_i(\beta)=\frac{-1}{g^2(X_i,\beta)}\frac{\partial g(X_i,\beta)}{\partial\beta},$$

and

$$H_i(\beta)=\frac{2}{g^3(X_i,\beta)}\frac{\partial g(X_i,\beta)}{\partial\beta}\left(\frac{\partial g(X_i,\beta)}{\partial\beta}\right)^T-\frac{1}{g^2(X_i,\beta)}\frac{\partial^2 g(X_i,\beta)}{\partial\beta^2}.$$

Next, if

$$\widehat{S}\subseteq S\quad\text{and}\quad\left\|\hat{\beta}-\beta^*\right\|_2\leqslant c\sqrt{\frac{sL_2\log n}{N}},$$

then we have

$$\left\|\widehat{W}^{-1}-W^{-1}\right\|_{\infty}=\max_i\left|\frac{1}{g(X_i,\hat{\beta})}-\frac{1}{g(X_i,\beta^*)}\right|\leqslant\left\|\hat{\beta}-\beta^*\right\|_2\max_i\left\|d_i(\beta^*)_S\right\|_2$$
$$+\frac{1}{2}\left\|\hat{\beta}-\beta^*\right\|_2^2\max_i\left\|H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)_{SS}\right\|_{op}\leqslant c\left(s\sqrt{\frac{L_2\log n}{N}}+\frac{s^2L_2\log n}{N}\right),$$

where $d_i(\beta^*)_S$ is subvector of $d_i(\beta^*)$ with elements restricted to S, $H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)_{SS}$ is submatrix of $H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)$ with rows and columns restricted to S, and the last inequality follows from

$$\max_i\left\|d_i\left(\beta^*\right)_S\right\|_2\leqslant c\sqrt{s},$$

and

$$\max_i\left\|H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)_{SS}\right\|_{op}\leqslant\left\|H_i\left(\beta^*+c_i\left(\hat{\beta}-\beta^*\right)\right)_{SS}\right\|_F\leqslant c's.$$

Applying Corollary 1 and assuming $N\geqslant s^2L_2\log n$, we have

$$\mathbb{P}\left(\left\|\widehat{W}^{-1}-W^{-1}\right\|_{\infty}\leqslant cs\sqrt{\frac{L_2\log n}{N}}\right)\geqslant 1-c'\exp\left(-c''\log s\right).\tag{C.2}$$

Finally, combining Lemma 8, (C.1), and (C.2), we have

$$\mathbb{P}\left(\left\|\Theta^*\hat{\Sigma}_N-I_n\right\|_{\infty}\leqslant c\left(\frac{1}{L_1}\sqrt{\frac{\log n}{N}}+s\sqrt{\frac{L_2\log n}{N}}\right)\right)\geqslant 1-c'\exp\left(-c''\log s\right).$$

Therefore, the proof is complete.

□

Lemma 14. (a) Assume that Assumption 2 and Assumption 4 hold. Under the scaling $N\gtrsim\log n$, we have

$$\max_j\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^N e_j^T X_i\varepsilon_i d_i(\beta^*)\right\|_{\infty}\leqslant c\sqrt{\log n},$$

with probability at least $1-c'\exp\left(-c''\log n\right)$.

(b) Assume that Assumption 2 holds. We have

$$\max_j \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |X_{ij}\varepsilon_i| \leqslant c\sqrt{N}\log n,$$

with probability at least $1 - c'\exp(-c''\log n)$.

(c) Assume that Assumption 2 holds. Under the scaling $N \gtrsim \log n$, we have

$$\left\| \frac{1}{\sqrt{N}} X^T W^{-1} \varepsilon \right\|_\infty \leqslant c\sqrt{\log n},$$

with probability at least $1 - c'\exp(-c''\log n)$.

Proof. First, we show Part (a). We have

$$\max_j \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_j^T X_i \varepsilon_i d_i(\beta^*) \right\|_\infty = \max_{1\leqslant j\leqslant n, 1\leqslant k\leqslant n} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_j^T X_i \varepsilon_i d_{ik}(\beta^*) \right|,$$

where $d_{ik}(\beta^*)$ is the kth element of $d_i(\beta^*)$. Note that

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_j^T X_i \varepsilon_i d_{ik}(\beta^*) \mid X \sim N\left(0, \frac{1}{N}\sum_{i=1}^{N} X_{ij}^2 d_{ik}^2 W_i\right),$$

where

$$\frac{1}{N}\sum_{i=1}^{N} X_{ij}^2 d_{ik}^2 W_i \leqslant \frac{c}{N}\sum_{i=1}^{N} X_{ij}^2.$$

By Lemma 13 in [23], we have

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} X_{ij}^2 \leqslant (\Sigma_x)_{jj} + c\right) \geqslant 1 - \exp(-c'N).$$

Therefore, using a similar argument with Lemma 1 and assuming $N \gtrsim \log n$, we have

$$\max_j \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} e_j^T X_i \varepsilon_i d_i(\beta^*) \right\|_\infty \leqslant c\sqrt{\log n},$$

with probability at least $1 - c'\exp(-c''\log n)$.

Next, using similar arguments with Part (a) yields results in Part (b) and Part (c). Therefore, the proof is complete.

□