To appear on the Annals of Statistics

SUPPLEMENTARY MATERIALS FOR "INNOVATED INTERACTION SCREENING FOR HIGH-DIMENSIONAL NONLINEAR CLASSIFICATION"

By Yingying Fan, Yinfei Kong, Daoji Li and Zemin Zheng

University of Southern California

APPENDIX A: PROOFS FOR PROPOSITION 1 AND MAIN LEMMAS

In this section, we prove all key Lemmas in the order they appear in the main text. Additional Lemmas and their proofs are provided in Appendix B.

Deviation of sub-Gaussian distribution. Recall that a random vector $\mathbf{w} = (W_1, \cdots, W_p)^T \in \mathbb{R}^p$ is sub-Gaussian if there exist some positive constants a and b such that

$$P(|\mathbf{v}^T\mathbf{w}| > t) \leq a\exp(-bt^2)$$

for any $t > 0$ and any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$.

Suppose $\mathbf{w} = (W_1, \cdots, W_p)$ is sub-Gaussian with constants a, b, mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{w}_1, \cdots, \mathbf{w}_n$ be n independent copies of $\mathbf{w}$. Since $\mathbf{w}$ is sub-Gaussian, by Lemma 10, we have $\mathbf{w} - \boldsymbol{\mu}$ is also sub-Gaussian. Then there exists some constants $\tilde{a}$ and $\tilde{b}$ such that

$$P(|\mathbf{v}^T(\mathbf{w}_i - \boldsymbol{\mu})(\mathbf{w}_i - \boldsymbol{\mu})^T\mathbf{v}| > x) \leq \tilde{a}\exp(-\tilde{b}x)$$

for all $x > 0$ and $\|\mathbf{v}\|_2 = 1$, which implies that

$$E\{\exp[t\mathbf{v}^T(\mathbf{w}_i - \boldsymbol{\mu})(\mathbf{w}_i - \boldsymbol{\mu})^T\mathbf{v}]\} < \infty$$

for all $0 < t < \tilde{b}$ and $\|\mathbf{v}\|_2 = 1$. Similar as in the proof of Lemma 3 in [1], we know that there exist some constants $C > 0$ and $\rho > 0$ depending on $\tilde{a}$ and $\tilde{b}$, such that for all $0 < x < \rho$ and any unit vector $\|\mathbf{v}\| = 1$,

$$P\{|(1/n)\sum_{i=1}^{n}\mathbf{v}^T(\mathbf{w}_i - \boldsymbol{\mu})(\mathbf{w}_i - \boldsymbol{\mu})^T\mathbf{v} - \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}| > x\} \leq Ce^{-nx^2\rho/2}. \quad \text{(A.1)}$$

A.1. Proof of Proposition 1. The main idea is to prove that our test statistic $\widetilde{D}_j$ converges to its population counterpart D_j uniformly over j. Since Condition 3 ensures that D_j is bounded from below for $j \in \mathcal{A}_1$ and $D_j = 0$ for $j \in \mathcal{A}_1^c$, the uniform convergence of $\widetilde{D}_j$ will imply the results in Proposition 1.

We proceed to prove the uniform convergence of $\widetilde{D}_j$ to D_j. To this end, we decompose the difference between them into the following five terms:

(A.2)
$$\begin{aligned}|\widetilde{D}_j - D_j| \le{}& \left|\log(\tilde{\sigma}_j^2/\sigma_j^2)\right| + \pi\left|\log\left[(\tilde{\sigma}_j^{(1)})^2/(\sigma_j^{(1)})^2\right]\right| + (1-\pi)\left|\log\left[(\tilde{\sigma}_j^{(2)})^2/(\sigma_j^{(2)})^2\right]\right| \\ &+ |n_1/n - \pi| \cdot \left|\log\left[(\tilde{\sigma}_j^{(1)})^2\right]\right| + |n_2/n - (1-\pi)| \cdot \left|\log\left[(\tilde{\sigma}_j^{(2)})^2\right]\right|.\end{aligned}$$

We will establish successively the deviation bounds of the terms on the right hand side above. The same notation C will be used to denote a generic constant without loss of generality.

By Lemma 7, the estimators $\tilde{\sigma}_j^2$ converge to σ_j^2 uniformly over all $j = 1, \cdots, p$, with probability at least $1 - p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$. Define this event as $\mathcal{E}$. We will condition on the event $\mathcal{E}$ hereafter. Since $x_n^{-1}\log(1+x_n) \to 1$ as $x_n \to 0$, it follows that

$$\log(\tilde{\sigma}_j^2/\sigma_j^2)/(\tilde{\sigma}_j^2/\sigma_j^2 - 1) \to 1$$

uniformly for all j as $n \to \infty$. Thus, uniformly over all $j = 1, \cdots, p$, with sufficiently large n, we have the following bound for the first term $|\log(\tilde{\sigma}_j^2/\sigma_j^2)|$ on the right hand side of (A.2)

$$P(|\log(\tilde{\sigma}_j^2/\sigma_j^2)| > 4^{-1}cn^{-\kappa}|\mathcal{E}) \le P(|\tilde{\sigma}_j^2/\sigma_j^2 - 1| > 8^{-1}cn^{-\kappa}|\mathcal{E}), \tag{A.3}$$

where constants c and κ are defined in Condition 3. Then (A.3) together with (A.7) in the proof of Lemma 7 entails that

$$P(|\log(\tilde{\sigma}_j^2/\sigma_j^2)| > 4^{-1}cn^{-\kappa}|\mathcal{E}) \le \exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}). \tag{A.4}$$

Using the same arguments, for either $k = 1$ or 2, we can prove that

$$P(|\log[(\tilde{\sigma}_j^{(k)})^2/(\sigma_j^{(k)})^2]| > 4^{-1}cn^{-\kappa}|\mathcal{E}) \le \exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}). \tag{A.5}$$

By the proof of Lemma 6, we know that

$$\tilde{\tau}_{2,p} \ge \pi\lambda_{\max}(\mathbf{\Omega}_1) + (1-\pi)\lambda_{\max}(\mathbf{\Omega}_1\mathbf{\Sigma}_2\mathbf{\Omega}_1).$$

Since $(\tilde{\sigma}_j^{(1)})^2$ and $(\tilde{\sigma}_j^{(2)})^2$ can be bounded from above by $\lambda_{\max}(\mathbf{\Omega}_1)$ and $\lambda_{\max}(\mathbf{\Omega}_1\mathbf{\Sigma}_2\mathbf{\Omega}_1)$, respectively, we know that $(\tilde{\sigma}_j^{(k)})^2$ can be bounded from

above by $\pi^{-1}(1-\pi)^{-1}\tilde{\tau}_{2,p}$ for $k=1,2$. As $n_1=\sum_{i=1}^n \Delta_i$, by Hoeffding's inequality we get

$$
\begin{aligned}
(\text{A.6}) \qquad & P(|n_1/n-\pi|\cdot|\log[(\tilde{\sigma}_j^{(1)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E}) \\
&\le P(|\frac{1}{n}\sum_{i=1}^n \Delta_i-\pi|>\frac{cn^{-\kappa}}{4|\log[\pi^{-1}(1-\pi)^{-1}\tilde{\tau}_{2,p}]|}|\mathcal{E}) \\
&\le 2\exp(-\frac{2nc^2n^{-2\kappa}}{16\log^2[\pi^{-1}(1-\pi)^{-1}\tilde{\tau}_{2,p}]})\le \exp\{-Cn^{1-2\kappa}/\log^2(\tilde{\tau}_{2,p})\}.
\end{aligned}
$$

Similarly, we have

(A.7)
$$
P(|n_2/n-(1-\pi)|\cdot|\log[(\tilde{\sigma}_j^{(2)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E})\le \exp\{-Cn^{1-2\kappa}/\log^2(\tilde{\tau}_{2,p})\}.
$$

In view of (A.2), we have

(A.8)
$$
\begin{aligned}
& P(|\widetilde{D}_j-D_j|>cn^{-\kappa}|\mathcal{E}) \\
&\le P(|\log(\tilde{\sigma}_j^2/\sigma_j^2)|>4^{-1}cn^{-\kappa}|\mathcal{E})+P(|\log[(\tilde{\sigma}_j^{(1)})^2/(\sigma_j^{(1)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E}) \\
&+ P(|\log[(\tilde{\sigma}_j^{(1)})^2/(\sigma_j^{(1)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E})+P(|\frac{n_1}{n}-\pi|\cdot|\log[(\tilde{\sigma}_j^{(1)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E}) \\
&+ P(|\frac{n_2}{n}-(1-\pi)|\cdot|\log[(\tilde{\sigma}_j^{(2)})^2]|>4^{-1}cn^{-\kappa}|\mathcal{E}).
\end{aligned}
$$

Combining the probability bounds in (A.4), (A.5), (A.6) and (A.7) gives

$$
\begin{aligned}
P(|\widetilde{D}_j-D_j|>cn^{-\kappa}|\mathcal{E}) &\le 3\exp(-C\tilde{\tau}_{1,p}^2n^{1-2\kappa})+2\exp\{-Cn^{1-2\kappa}/\log^2(\tilde{\tau}_{2,p})\} \\
&\le \exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
P(\max_{1\le j\le p}|\widetilde{D}_j-D_j|>cn^{-\kappa}|\mathcal{E}) &\le \sum_{j=1}^p P(|\widetilde{D}_j-D_j|>cn^{-\kappa}|\mathcal{E}) \\
&\le p\exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\},
\end{aligned}
$$

which is the deviation of the statistic $\widetilde{D}_j$ from its population counterpart D_j. By Lemma 7, $P(\mathcal{E}^c)\le p\exp(-C\tilde{\tau}_{1,p}^2n^{1-2\kappa})$. Thus we have

$$
\begin{aligned}
& P(\max_{1\le j\le p}|\widetilde{D}_j-D_j|>cn^{-\kappa})\le P(\max_{1\le j\le p}|\widetilde{D}_j-D_j|>cn^{-\kappa}|\mathcal{E})+P(\mathcal{E}^c) \\
& \le p\exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\}.
\end{aligned}
$$

Therefore, for any p satisfying $\log p = O(n^\gamma)$ with $\gamma > 0$, $\gamma + 2\kappa < 1$ and $\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p}) = o(n^{1-2\kappa-\gamma})$, it follows that

$$P(\max_{1\le j\le p} |\widetilde{D}_j - D_j| > cn^{-\kappa}) \le \exp\{n^\gamma - Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p})]\}$$
$$\le \exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2} + \log^2(\tilde{\tau}_{2,p})]\}.$$

By Condition 3 and its discussion, we know that $D_j \ge 3cn^{-\kappa}$ when $j \in \mathcal{A}_1$, and $D_j = 0$ otherwise. It follows that

$$\{\min_{j\in\mathcal{A}_1} \widetilde{D}_j < 2cn^{-\kappa}\} \cup \{\max_{j\in\mathcal{A}_1^c} \widetilde{D}_j > cn^{-\kappa}\} \subset \{\max_{j\in\{1,\cdots,p\}} |\widetilde{D}_j - D_j| > cn^{-\kappa}\},$$

which shows that with probability at least $1-\exp\{-Cn^{1-2\kappa}/[\tilde{\tau}_{1,p}^{-2}+\log^2(\tilde{\tau}_{2,p})]\}$,

$$\min_{j\in\mathcal{A}_1} \widetilde{D}_j \ge 2cn^{-\kappa} \text{ and } \max_{j\in\mathcal{A}_1^c} \widetilde{D}_j \le cn^{-\kappa}$$

for sufficiently large n.

As the same conditions hold for the covariance matrix $\mathbf{\Sigma}_2$ and the data after the second transformation, the results above also apply to the covariates in $\mathcal{A}_2$ with the test statistics calculated based on data transformed by $\mathbf{\Omega}_2$. This completes the proof of Proposition 1.

A.2. Lemma 1 and its proof.

LEMMA 1. Under model setting (2) and conditions in Theorem 1, for sufficiently large n, with probability at least $1-p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$, it holds that

$$\max_{1\le j\le p} |\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1| \le T_{n,p}/6$$

for some positive constant C.

PROOF. Let $\overline{\mathbf{Z}} = \mathbf{1}_n(\bar{z}_1, \cdots, \bar{z}_p)$ with $\bar{z}_j = \sum_{i=1}^n z_{ij}/n$ for the original data matrix without transformation. We will first bound the term $\hat{\sigma}_j^2 - \tilde{\sigma}_j^2$ by writing it as

$$\hat{\sigma}_j^2 - \tilde{\sigma}_j^2 = \mathbf{e}_j^T[\widehat{\mathbf{\Omega}}_1(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\mathbf{\Omega}_1]\mathbf{e}_j/n.$$

Through a further decomposition, it gives

$$\hat{\sigma}_j^2 - \tilde{\sigma}_j^2 = \mathbf{e}_j^T(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)\mathbf{e}_j/n$$
$$+2\mathbf{e}_j^T(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\mathbf{\Omega}_1\mathbf{e}_j/n. \quad \text{(A.9)}$$

We will then bound the two terms on the right hand side above separately.

Recall that $\|\cdot\|_{\max}$ denotes the componentwise infinity norm for a matrix. For the first term, we have

$$\begin{aligned}
&|\mathbf{e}_j^T(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)\mathbf{e}_j/n| \\
&\le \|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}\|(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)\mathbf{e}_j\|_1^2/n \\
&\le \|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}[(K_p + K_p') \cdot C_1 K_p^2\sqrt{(\log p)/n}]^2/n \\
&= (\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n) \cdot [(K_p + K_p')^2 C_1^2 K_p^4(\log p)/n],
\end{aligned}$$

where the second inequality follows from the definition of acceptable estimator and the fact that $\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1$ is $(K_p + K_p')$-sparse.

For the second term, similarly we get

$$\begin{aligned}
&|\mathbf{e}_j^T(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\mathbf{\Omega}_1\mathbf{e}_j/n| \\
&\le \|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}\|(\widehat{\mathbf{\Omega}}_1 - \mathbf{\Omega}_1)\mathbf{e}_j\|_1\|\mathbf{\Omega}_1\mathbf{e}_j\|_1/n \\
&\le (\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n) \cdot [(K_p + K_p')C_1 K_p^2\sqrt{(\log p)/n}] \cdot K_p\|\mathbf{\Omega}_1\|_{\max}.
\end{aligned}$$

Since $\|\mathbf{\Omega}_1\|_{\max}$ is assumed to be upper bounded in Condition 4, and the above two bounds are independent of the index j, in view of (A.9), we know that there exists some constant $\tilde{C}$ such that

(A.10)
$$\max_{1\le j\le p}|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| \le \tilde{C}(\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n)[(K_p + K_p')K_p^3\sqrt{(\log p)/n}] \cdot \max\{(K_p + K_p')K_p\sqrt{(\log p)/n}, 1\}.$$

This together with the definition of $T_{n,p}$ before Theorem 1 ensures that

$$\max_{1\le j\le p}|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| \le \tilde{C}(\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n) \cdot T_{n,p}\tilde{\tau}_{1,p}/(\tilde{C}_1\tau_{2,p}).$$

By Lemma 6, σ_j^2 are uniformly bounded from below by $\tilde{\tau}_{1,p}$. By Lemma 7, $\max_{1\le j\le p}|\tilde{\sigma}_j^2/\sigma_j^2 - 1| \le cn^{-\kappa}/8$. Combining these two results entails that for n large enough,

$$\tilde{\sigma}_j^2 \ge (1 - cn^{-\kappa}/8)\sigma_j^2 > \tilde{\tau}_{1,p}/2,$$

uniformly for all j, with probability at least $1 - p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$. Denote by $\mathcal{E}$ the event that the results in Lemma 7 hold. By Lemma 8, when $\tilde{C}_1 \ge 12c_3\tilde{C}$, we have

(A.11)
$$\begin{aligned}
&P(\max_{1\le j\le p}|\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1| > T_{n,p}/6|\mathcal{E}) \\
&\le P(\max_{1\le j\le p}|\hat{\sigma}_j^2 - \tilde{\sigma}_j^2| > T_{n,p}\tilde{\tau}_{1,p}/12|\mathcal{E}) \\
&\le P(\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n > c_3\tau_{2,p}|\mathcal{E}) \le p^2\exp(-\tilde{C}_2 n).
\end{aligned}$$

Under the conditions in Theorem 1, we have $\log p = o(n^\gamma)$ with $\gamma > 0$ and $\gamma + 2\kappa < 1$. It follows that

$$P(\max_{1\le j\le p} |\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1| > T_{n,p}/6) \le P(\max_{1\le j\le p} |\hat{\sigma}_j^2/\tilde{\sigma}_j^2 - 1| > T_{n,p}/6|\mathcal{E}) + P(\mathcal{E}^c)$$
$$\le p^2 \exp(-\tilde{C}_2 n) + p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}) = p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}),$$

where we use the same notation C here to denote a generic constant without lost of generality. This completes the proof of Lemma 1. □

A.3. Lemma 2 and its proof.

LEMMA 2. Under Condition 5, we have

$$\widetilde{C} n^{-1}\|\mathbf{X}\boldsymbol{\delta}\|_2^2 + \text{pen}(\widehat{\boldsymbol{\theta}}) \le \|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty \|\boldsymbol{\delta}\|_1 + \text{pen}(\boldsymbol{\theta}_0), \tag{A.12}$$

where $\widetilde{C}$ is some positive constant depending on the positive constant $\pi_{\min}$ in Condition 6, and $\boldsymbol{\delta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is the estimation error for the regularized estimator $\widehat{\boldsymbol{\theta}}$ defined in (18), and $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})$ with $\mathbf{y} = (\Delta_1, \cdots, \Delta_n)^T$.

PROOF. Define $\ell_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \ell_n(\mathbf{x}_i^T\boldsymbol{\theta}, \Delta_i)$ where $\ell_n(\mathbf{x}^T\boldsymbol{\theta}, \Delta) = -\Delta\mathbf{x}^T\boldsymbol{\theta} + \log[1 + \exp(\mathbf{x}^T\boldsymbol{\theta})]$. Then, in matrix form, $\ell_n(\boldsymbol{\theta})$ can be rewritten as

$$\ell_n(\boldsymbol{\theta}) = -n^{-1}\{\mathbf{y}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{1}^T\mathbf{b}(\mathbf{X}\boldsymbol{\theta})\},$$

where $\mathbf{y} = (\Delta_1, \cdots, \Delta_n)^T$ is an n-dimensional response vector with $\Delta_i \in \{0, 1\}$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{\tilde{p}})$ is an $n \times \tilde{p}$ augmented design matrix, $\mathbf{1}$ is an n-dimensional vector with each component being one, $\mathbf{b}(\boldsymbol{\beta}) = (b(\beta_1), \cdots, b(\beta_n))^T$ is a vector-valued function with $\beta_i = \mathbf{x}_i^T\boldsymbol{\theta}$ and $b(u) = \log[1 + \exp(u)]$.

By the definition of $\widehat{\boldsymbol{\theta}}$, we have $\ell_n(\boldsymbol{\theta}) + \text{pen}(\boldsymbol{\theta}) \le \ell_n(\boldsymbol{\theta}_0) + \text{pen}(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is the true regression coefficient vector of $\boldsymbol{\theta}$. Rearranging terms yields

$$n^{-1}\mathbf{1}^T[\mathbf{b}(\mathbf{X}\widehat{\boldsymbol{\theta}}) - \mathbf{b}(\mathbf{X}\boldsymbol{\theta}_0)] + \text{pen}(\widehat{\boldsymbol{\theta}}) \le n^{-1}\mathbf{y}^T\mathbf{X}\boldsymbol{\delta} + \text{pen}(\boldsymbol{\theta}_0), \tag{A.13}$$

where $\boldsymbol{\delta} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is the estimation error. Applying Taylor expansion to the function of $\mathbf{1}^T\mathbf{b}(\mathbf{X}\widehat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}_0$ gives

$$\mathbf{1}^T[\mathbf{b}(\mathbf{X}\widehat{\boldsymbol{\theta}}) - \mathbf{b}(\mathbf{X}\boldsymbol{\theta}_0)] = [\mathbf{b}'(\mathbf{X}\boldsymbol{\theta}_0)]^T\mathbf{X}\boldsymbol{\delta} + 2^{-1}\boldsymbol{\delta}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\boldsymbol{\delta}, \tag{A.14}$$

where $\mathbf{H} = \mathbf{H}(\mathbf{X}, \tilde{\boldsymbol{\theta}}_1, \cdots, \tilde{\boldsymbol{\theta}}_n) = \text{diag}\{b''(\mathbf{x}_1^T\tilde{\boldsymbol{\theta}}_1), \cdots, b''(\mathbf{x}_n^T\tilde{\boldsymbol{\theta}}_n)\}$ is an $n \times n$ diagonal matrix with $\tilde{\boldsymbol{\theta}}_i \in \mathbb{R}^{\tilde{p}}$ lying on the line segment adjoining $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, $i = 1, \cdots, n$. Combining (A.13) and (A.14) and rearranging terms yield

$$(2n)^{-1}\boldsymbol{\delta}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\boldsymbol{\delta} + \text{pen}(\widehat{\boldsymbol{\theta}}) \le n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\boldsymbol{\delta} + \text{pen}(\boldsymbol{\theta}_0), \tag{A.15}$$

where $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X}) = \mathbf{y} - \mathbf{b}'(\mathbf{X}\boldsymbol{\theta}_0)$. The right hand side of the above inequality can be bounded as

$$n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\boldsymbol{\delta} + \text{pen}(\boldsymbol{\theta}_0) \le \|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty\|\boldsymbol{\delta}\|_1 + \text{pen}(\boldsymbol{\theta}_0). \tag{A.16}$$

By Condition 5, we have

$$\boldsymbol{\delta}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\boldsymbol{\delta} \ge 2\widetilde{C}\|\mathbf{X}\boldsymbol{\delta}\|_2^2$$

for some positive constant $\widetilde{C}$, which depends on the constant $\pi_{\min}$ in Condition 5. This inequality, together with (A.15) and (A.16), completes the proof. □

A.4. Lemma 3 and its proof.

LEMMA 3. Assume that Condition 1 holds. If $\log(p) = o(n)$, then with probability $1 - O(p^{-\tilde{c}_1})$, we have $\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty \le 2^{-1}c_0\sqrt{\log(p)/n}$, where c_0 is some positive constant and $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y}|\mathbf{X})$ with $\mathbf{y} = (\Delta_1, \cdots, \Delta_n)^T$.

PROOF. Recall that $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_{\tilde{p}})$. An application of the Bonferroni inequality gives that

$$P(\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty > \lambda_0) \le \sum_{j=1}^{\tilde{p}} P(|n^{-1}\boldsymbol{\varepsilon}^T\tilde{\mathbf{x}}_j| > \lambda_0) \tag{A.17}$$

for any λ_0. The key idea is to bound $P(|n^{-1}\boldsymbol{\varepsilon}^T\tilde{\mathbf{x}}_j| > \lambda_0)$. To this end, consider the following three cases.

Case 1: $j = 1$. In this case, $\tilde{\mathbf{x}}_j = \mathbf{1}$, where $\mathbf{1}$ is a n-dimensional vector with each component being one. Recall that $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ with $\varepsilon_i = \Delta_i - E(\Delta_i|\mathbf{x}_i)$ and $\Delta_i \in \{0, 1\}$. So $-1 \le \varepsilon_i \le 1$ for $i = 1, \cdots, n$. Thus, by Hoeffding's inequality [3], we have

$$P(|n^{-1}\boldsymbol{\varepsilon}^T\tilde{\mathbf{x}}_j| > \lambda_0) = P(|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{1}| > \lambda_0) \le 2\exp\left(-n\lambda_0^2/2\right).$$

Case 2: $2 \le j \le p + 1$. In this case, $\tilde{\mathbf{x}}_j = (Z_{1,j-1}, \cdots, Z_{n,j-1})^T$. Thus, $n^{-1}\boldsymbol{\varepsilon}^T\tilde{\mathbf{x}}_j = n^{-1}\sum_{i=1}^n \varepsilon_i Z_{i,j-1}$. From Lemma 9, under Condition 1, we have that $\mathbf{z} = (Z_1, \cdots, Z_p)^T$ is sub-Gaussian, that is, there exist some positive constants a_1 and b_1 such that

$$P(|\mathbf{v}^T\mathbf{z}| > t) \le a_1\exp(-b_1t^2)$$

for any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$. Therefore, for any $2 \le j \le p+1$, taking $\mathbf{v} = e_{j-1}$ being a unit vector with the $(j-1)$ component being one and zero elsewhere in the inequality above gives

$$P(|Z_{j-1}| > t) \le a_1 \exp(-b_1 t^2)$$

holds uniformly for all $2 \le j \le p+1$. By Lemma 11, we have $E(e^{b_1 Z_{j-1}^2/2}) \le 1 + a_1$. This, together with the inequality $ab \le (a^2+b^2)/2$ for any $a, b > 0$, gives

$$\begin{aligned} e^{(b_1/2)|\varepsilon_i Z_{i,j-1}|} &\le e^{b_1(\varepsilon_i^2 + Z_{i,j-1}^2)/4} = e^{b_1\varepsilon_i^2/4} e^{b_1 Z_{i,j-1}^2/4} \\ &\le (e^{b_1\varepsilon_i^2/2} + e^{b_1 Z_{i,j-1}^2/2})/2 \le (e^{b_1/2} + 1 + a_1)/2 \end{aligned} \tag{A.18}$$

for all $1 \le i \le n$ and $2 \le j \le p+1$, where we have used the fact that $-1 \le \varepsilon_i \le 1$ in the last inequality. Thus, it follows from Lemma 12 that for any $0 < \lambda_0 \le 1$, there exist some positive constants a_2 and b_2 such that

$$P(|n^{-1}\boldsymbol{\varepsilon}^T \tilde{\mathbf{x}}_j| \ge \lambda_0) = P(|n^{-1}\sum_{i=1}^n \varepsilon_i Z_{i,j-1}| \ge \lambda_0) \le a_2 \exp(-b_2 n \lambda_0^2)$$

for all $2 \le j \le p+1$.

Case 3: $p+2 \le j \le \tilde{p}$. In this case, $\tilde{\mathbf{x}}_j = (Z_{1,k}Z_{1,\ell}, \cdots, Z_{n,k}Z_{n,\ell})^T$ for some $1 \le k \le \ell \le p$. We can use similar arguments for Case 2 to bound $P(|n^{-1}\boldsymbol{\varepsilon}^T \tilde{\mathbf{x}}_j| \ge \lambda_0)$. Similarly to (A.18), we have

$$\begin{aligned} e^{(b_1/2)|\varepsilon_i Z_{ik} Z_{i\ell}|} &\le e^{(b_1/2)|Z_{ik}Z_{i\ell}|} \le e^{b_1(Z_{ik}^2 + Z_{i\ell}^2)/4} = e^{b_1 Z_{ik}^2/4} e^{b_1 Z_{i\ell}^2/4} \\ &\le (e^{b_1 Z_{ik}^2/2} + e^{b_1 Z_{i\ell}^2/2})/2 \le 1 + a_1, \end{aligned}$$

for all $1 \le i \le n$ and all $1 \le k \le \ell \le p$. This together with Lemma 12 gives that for any $0 < \lambda_0 \le 1$, there exist some positive constants a_3 and b_3 such that

$$P(|n^{-1}\boldsymbol{\varepsilon}^T \tilde{\mathbf{x}}_j| \ge \lambda_0) = P(|n^{-1}\sum_{i=1}^n \varepsilon_i Z_{ik} Z_{i\ell}| \ge \lambda_0) \le a_3 \exp(-b_3 n \lambda_0^2)$$

for all $p+2 \le j \le \tilde{p}$.

Combining Cases 1-3 above yields

$$P(|n^{-1}\boldsymbol{\varepsilon}^T \tilde{\mathbf{x}}_j| \ge \lambda_0) \le a_4 \exp(-b_4 n \lambda_0^2) \tag{A.19}$$

for any $0 \le \lambda_0 \le 1$, where $a_4 = \max\{2, a_2, a_3\}$ and $b_4 = \min\{1/2, b_2, b_3\}$. Let $\lambda_0 = 2^{-1} c_0 \sqrt{\log(p)/n}$ with some positive constant c_0. Since $\log(p) = o(n)$,

we have $0 < \lambda_0 \leq 1$ for all sufficiently large n. In view of (A.17) and (A.19), we have

$$\begin{aligned} P(\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty > 2^{-1}c_0\sqrt{\log(p)/n}) \leq & \tilde{p}a_4 \exp\{-4^{-1}b_4c_0^2\log(p)\} \\ \leq & 3a_4p^{-(4^{-1}b_4c_0^2-2)}, \end{aligned}$$

where $c_0 > \sqrt{8/b_4}$ and we have used the fact that $\tilde{p} = 1+p+p(p+1)/2 \leq 3p^2$. Thus, we conclude that $\|n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X}\|_\infty \leq c_0\sqrt{\log(p)/n}$ holds with probability at least $1 - O(p^{-\tilde{c}_1})$ with $\tilde{c}_1 = 4^{-1}b_4c_0^2 - 2 > 0$.

□

A.5. Lemma 4 and its proof.

LEMMA 4. Assume that there exists some constant $\phi > 0$ such that

$$\boldsymbol{\delta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\delta} \geq \phi^2\boldsymbol{\delta}_S^T\boldsymbol{\delta}_S \tag{A.20}$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$, where $\tilde{\boldsymbol{\Sigma}} = E(\mathbf{x}^T\mathbf{x})$. If both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are sub-Gaussian, $5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(n^{\xi/2})$, and $\log(p) = o(n^{1/2-2\xi})$ with constant $0 \leq \xi < 1/4$, then with probability at least $1 - O(p^{-\tilde{c}_2})$,

$$n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \geq (\phi/2)\|\boldsymbol{\delta}_S\|_2$$

holds for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$ when n is sufficiently large.

PROOF. The idea is to show that the desired inequality holds conditioning on an event and the probability of this event occurring is at most $O(p^{-\tilde{c}_2})$ with some positive constant $\tilde{c}_2$.

Conditioning the event $\mathcal{E}_4 = \left\{\|n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}}\|_\infty < C_1n^{-\xi}\right\}$ where positive constant C_1 will be specified later, we have

$$\begin{aligned} |\boldsymbol{\delta}^T(n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}})\boldsymbol{\delta}| < C_1n^{-\xi}\|\boldsymbol{\delta}\|_1^2 &= C_1n^{-\xi}(\|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1)^2 \\ &\leq C_1n^{-\xi}(s^{1/2}\|\boldsymbol{\delta}_S\|_2 + \|\boldsymbol{\delta}_{S^c}\|_1)^2, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. The following arguments are all conditioning on the event $\mathcal{E}_4$. Thus, for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$, we obtain

$$|\boldsymbol{\delta}^T(n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}})\boldsymbol{\delta}| \leq C_1n^{-\xi}(5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)^2\|\boldsymbol{\delta}_S\|_2^2.$$

Since $5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 = O(n^{\xi/2})$, there exists some positive constant C_2 such that $5s^{1/2} + 4\lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2 \le C_2 n^{\xi/2}$. Thus,

$$|\boldsymbol{\delta}^T(n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}})\boldsymbol{\delta}| \le C_1 C_2^2 \|\boldsymbol{\delta}_S\|_2^2$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$. Note that $n^{-1}\boldsymbol{\delta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\delta} = \boldsymbol{\delta}^T(n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}})\boldsymbol{\delta} + \boldsymbol{\delta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\delta}$. This, together with the above inequality and the assumption (A.20), yields

$$n^{-1}\boldsymbol{\delta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\delta} \ge -C_1C_2^2\|\boldsymbol{\delta}_S\|_2^2 + \boldsymbol{\delta}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{\delta} \ge (\phi^2 - C_1C_2^2)\|\boldsymbol{\delta}_S\|_2^2.$$

Choose $C_1 = 3\phi^2/(4C_2^2)$. Thus, with probability $1 - P(\mathcal{E}_4^c)$, we have

$$n^{-1/2}\|\mathbf{X}\boldsymbol{\delta}\|_2 \ge (\phi/2)\|\boldsymbol{\delta}_S\|_2$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{\tilde{p}}$ satisfying $\|\boldsymbol{\delta}_{S^c}\|_1 \le 4(s^{1/2} + \lambda_1^{-1}\lambda_2\|\boldsymbol{\theta}_0\|_2)\|\boldsymbol{\delta}_S\|_2$.

It remains to show that $P(\mathcal{E}_4^c) \le O(p^{-\tilde{c}_2})$ with some positive constant $\tilde{c}_2$. For any matrix A, denote by $\|A\|_\infty$ the entrywise matrix infinity norm of A and $(A)_{k\ell}$ the (k,ℓ) entry of A. Since both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are sub-Gaussian, by Lemmas 9, 10, 11, and 13, we have

$$\begin{aligned}
&P(\mathcal{E}_4^c) = P\left\{\|n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}}\|_\infty \ge K_2/(2C_2^2)n^{-\xi}\right\} \\
\le &\sum_{k=1}^{\tilde{p}}\sum_{\ell=1}^{\tilde{p}} P\{|(n^{-1}\mathbf{X}^T\mathbf{X} - \tilde{\boldsymbol{\Sigma}})_{k\ell}| \ge K_2/(2C_2^2)n^{-\xi}\} \\
\le &\tilde{p}^2 a\exp(-4^{-1}bC_2^{-2}K_2^2 n^{1/2-2\xi}) \le O(p^{-\tilde{c}_2})
\end{aligned}$$

for all n sufficiently large, where $a, b, \tilde{c}_2$ are positive constants and the last inequality holds since $\tilde{p} = 1 + p + p(p+1)/2$ and $\log(p) = o(n^{1/2-2\xi})$. This completes the proof of Lemma 4. □

A.6. Lemma 5 and its proof.

LEMMA 5. Assume that $\mathbf{w} = (W_1, \cdots, W_p)^T \in \mathbb{R}^p$ is sub-Gaussian. Then for any positive constant c_1, there exists some positive constant C_2 such that

$$P\left\{\max_{1\le j\le p}|W_j| > C_2\sqrt{\log(p)}\right\} = O(p^{-c_1}).$$

PROOF. Since $\mathbf{w} = (W_1, \cdots, W_p)^T$ is sub-Gaussian, there exist some positive constants $\tilde{c}_1$ and $\tilde{c}_2$ such that $P(|\mathbf{v}^T\mathbf{w}| > t) \le \tilde{c}_1\exp(-\tilde{c}_2 t^2)$ for any

$\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$. Taking $\mathbf{v}$ be a unit vector with jth component 1 and all other components 0 yields

$$P(|W_j| > t) \leq \tilde{c}_1 \exp(-\tilde{c}_2 t^2)$$

for all $1 \leq j \leq p$. An application of the Bonferroni inequality gives

$$P(\max_{1\leq j\leq p} |W_j| > t) \leq \sum_{j=1}^{p} P(|W_j| > t) \leq p\tilde{c}_1 \exp(-\tilde{c}_2 t^2).$$

If we choose $t = C_2\sqrt{\log(p)}$ with some positive constant $C_2 = \sqrt{1 + c_1/\tilde{c}_2}$, then we have $P\left\{\max_{1\leq j\leq p} |W_j| > C_2\sqrt{\log(p)}\right\} = O(p^{-c_1})$. This completes the proof of Lemma 5.

□

APPENDIX B: PROOFS FOR SECONARY LEMMAS

B.1. Lemma 6 and its proof.

LEMMA 6. Under Condition 2, it holds that

$$\tilde{\tau}_{1,p} \leq \lambda_{\min}[\mathrm{cov}(\tilde{\mathbf{z}})] \leq \lambda_{\max}[\mathrm{cov}(\tilde{\mathbf{z}})] \leq \tilde{\tau}_{2,p},$$

where $\tilde{\tau}_{1,p} = \{\pi\tau_{2,p}^{-1} + (1-\pi)\tau_1\tau_{2,p}^{-2}\} \wedge 1$ and $\tilde{\tau}_{2,p} = \{\pi\tau_1^{-1} + (1-\pi)\tau_1^{-2}\tau_{2,p} + \pi(1-\pi)\tau_1^{-2}\|\boldsymbol{\mu}_1\|_2^2\} \vee \exp(1)$.

PROOF. In order to prove Lemma 6, we will first calculate the covariance matrix of $\mathbf{z} = \Delta\mathbf{z}^{(1)} + (1-\Delta)\mathbf{z}^{(2)}$. Recall that $\boldsymbol{\mu}_2 = E(\mathbf{z}^{(2)}) = \mathbf{0}$ and Δ is a Bernoulli variable taking value 1 with probability π. We will apply the formula $\mathrm{cov}(\mathbf{z}) = \mathrm{cov}[E(\mathbf{z}|\Delta)] + E[\mathrm{cov}(\mathbf{z}|\Delta)]$ to calculate the covariance matrix of $\mathbf{z}$.

For the first term $\mathrm{cov}[E(\mathbf{z}|\Delta)]$, we can calculate it as

$$E(\mathbf{z}|\Delta) = \Delta E(\mathbf{z}^{(1)}) + (1-\Delta)E(\mathbf{z}^{(2)}) = \Delta\boldsymbol{\mu}_1,$$

which gives $\mathrm{cov}[E(\mathbf{z}|\Delta)] = \mathrm{cov}(\Delta)\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T = \pi(1-\pi)\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T$.

For the second term $E[\mathrm{cov}(\mathbf{z}|\Delta)]$, since $\Delta^2 = \Delta$, $(1-\Delta)^2 = 1-\Delta$ and $\Delta(1-\Delta) = 0$, we have

$$\begin{aligned}&\mathrm{cov}(\mathbf{z}|\Delta) = E(\mathbf{z}\mathbf{z}^T|\Delta) - E(\mathbf{z}|\Delta)E(\mathbf{z}|\Delta)^T \\ &= \Delta^2\{E(\mathbf{z}^{(1)}\mathbf{z}^{(1)^T}) - \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T\} + (1-\Delta)^2 E(\mathbf{z}^{(2)}\mathbf{z}^{(2)^T}) = \Delta\boldsymbol{\Sigma}_1 + (1-\Delta)\boldsymbol{\Sigma}_2.\end{aligned}$$

After taking expectation on both sides, we get

$$E[\mathrm{cov}(\mathbf{z}|\Delta)] = \pi\boldsymbol{\Sigma}_1 + (1-\pi)\boldsymbol{\Sigma}_2.$$

Thus we have $\mathrm{cov}(\mathbf{z}) = \pi\boldsymbol{\Sigma}_1 + (1-\pi)\boldsymbol{\Sigma}_2 + \pi(1-\pi)\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T$.

Recall that $\tilde{\mathbf{z}} = \boldsymbol{\Omega}_1\mathbf{z}$. It follows that

$$\mathrm{cov}(\tilde{\mathbf{z}}) = \boldsymbol{\Omega}_1\mathrm{cov}(\mathbf{z})\boldsymbol{\Omega}_1 = \pi\boldsymbol{\Omega}_1 + (1-\pi)\boldsymbol{\Omega}_1\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_1 + \pi(1-\pi)\boldsymbol{\Omega}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T\boldsymbol{\Omega}_1.$$

Therefore, under Condition 2, we get

$$\begin{aligned}
\lambda_{\max}[\mathrm{cov}(\tilde{\mathbf{z}})] &\le \pi\lambda_{\max}(\boldsymbol{\Omega}_1) + (1-\pi)\lambda_{\max}(\boldsymbol{\Omega}_1\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_1) + \pi(1-\pi)\lambda_{\max}(\boldsymbol{\Omega}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T\boldsymbol{\Omega}_1)\\
&\le \pi\tau_1^{-1} + (1-\pi)\tau_1^{-2}\tau_{2,p} + \pi(1-\pi)\tau_1^{-2}\|\boldsymbol{\mu}_1\|_2^2;\\
\lambda_{\min}[\mathrm{cov}(\tilde{\mathbf{z}})] &\ge \pi\lambda_{\min}(\boldsymbol{\Omega}_1) + (1-\pi)\lambda_{\min}(\boldsymbol{\Omega}_1\boldsymbol{\Sigma}_2\boldsymbol{\Omega}_1) + \pi(1-\pi)\lambda_{\min}(\boldsymbol{\Omega}_1\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T\boldsymbol{\Omega}_1)\\
&\ge \pi\tau_{2,p}^{-1} + (1-\pi)\tau_1\tau_{2,p}^{-2}.
\end{aligned}$$

It completes the proof of Lemma 6. □

B.2. Lemma 7 and its proof.

LEMMA 7. Under model setting (2) and conditions in Proposition 1, for sufficiently large n, with probability at least $1-p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa})$, it holds that

$$\max_{1\le j\le p} |\tilde{\sigma}_j^2/\sigma_j^2 - 1| \le cn^{-\kappa}/8,$$

for some positive constant C, where c and κ are constants defined in Condition 3.

PROOF. We will first decompose $\tilde{\sigma}_j^2 - \sigma_j^2$ into several terms, and then prove deviation bounds for each term. Denote $\|\boldsymbol{\Omega}_1\|_2$ by the operator norm of $\boldsymbol{\Omega}_1$. Note that under Condition 2, $\|\boldsymbol{\Omega}_1\|_2$ is bounded from above by constant τ_1^{-1}. So $\tilde{\mathbf{z}}^{(k)} = \boldsymbol{\Omega}_1\mathbf{z}^{(k)}$ for $k = 1, 2$ are also sub-Gaussian distributed. Recall that $\widetilde{\mathbf{Z}} = \mathbf{Z}\boldsymbol{\Omega}_1$ is the transformed data matrix. Denote by $\tilde{\mathbf{Z}} = (\tilde{z}_{ij})_{n\times p}$. Then $\tilde{z}_{ij}$ are independent and identically distributed across i with mixture sub-Gaussian distribution and variance σ_j^2. Since $\tilde{\sigma}_j^2$ is the pooled sample variance estimate for the jth transformed feature $\widetilde{Z}_j$, we have

$$\tilde{\sigma}_j^2 = \sum_{i=1}^{n}(\tilde{z}_{ij} - \bar{\tilde{z}}_j)^2/n,$$

where $\bar{\tilde{z}}_j = \sum_{i=1}^n \tilde{z}_{ij}/n$ is the pooled sample mean estimate for $\widetilde{Z}_j$. Let $\tilde{\mu}_j = E(\tilde{z}_{ij})$ and $\tilde{\mu}_j^{(1)} = E(\tilde{z}_{ij}^{(1)})$. It is clear that $\tilde{\mu}_j = \pi\tilde{\mu}_j^{(1)}$. By some simple calculation, we have the following decomposition for $\tilde{\sigma}_j^2 - \sigma_j^2$,

$$\tilde{\sigma}_j^2 - \sigma_j^2 = \sum_{i=1}^n ([\tilde{z}_{ij} - \tilde{\mu}_j]^2 - \sigma_j^2)/n - (\bar{\tilde{z}}_j - \tilde{\mu}_j)^2.$$

Since $\tilde{z}_{ij} = \Delta_i\tilde{z}_{ij}^{(1)} + (1-\Delta_i)\tilde{z}_{ij}^{(2)}$, we have $\tilde{z}_{ij} - \tilde{\mu}_j = \Delta_i(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}) + (1-\Delta_i)\tilde{z}_{ij}^{(2)} + (\Delta_i - \pi)\tilde{\mu}_j^{(1)}$, where $\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}$ and $\tilde{z}_{ij}^{(2)}$ are sub-Gaussian distributed with mean 0 and variances $(\sigma_j^{(1)})^2$ and $(\sigma_j^{(2)})^2$, respectively. By the proof of Lemma 6, we know that $\sigma_j^2 = \pi(\sigma_j^{(1)})^2 + (1-\pi)(\sigma_j^{(2)})^2 + \pi(1-\pi)(\tilde{\mu}_j^{(1)})^2$. As $\Delta_i^2 = \Delta_i$, $(1-\Delta_i)^2 = 1-\Delta_i$ and $\Delta_i(1-\Delta_i) = 0$, by replacing $\tilde{z}_{ij}$ with $\Delta_i\tilde{z}_{ij}^{(1)} + (1-\Delta_i)\tilde{z}_{ij}^{(2)}$ and spreading out the terms, we can further decompose $\tilde{\sigma}_j^2 - \sigma_j^2$ as

$$\begin{aligned}\tilde{\sigma}_j^2 - \sigma_j^2 =& \frac{1}{n}\sum_{i=1}^n \big\{\Delta_i[(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})^2 - (\sigma_j^{(1)})^2] + (1-\Delta_i)[(\tilde{z}_{ij}^{(2)})^2 - (\sigma_j^{(2)})^2] \\ &+ [(\Delta_i - \pi)^2 - \pi(1-\pi)](\tilde{\mu}_j^{(1)})^2 + (\Delta_i - \pi)[(\sigma_j^{(1)})^2 - (\sigma_j^{(2)})^2] \\ &+ 2(\Delta_i - \pi)\tilde{\mu}_j^{(1)}[\Delta_i(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}) + (1-\Delta_i)\tilde{z}_{ij}^{(2)}]\big\} - (\bar{\tilde{z}}_j - \tilde{\mu}_j)^2.\end{aligned}$$

Let $S = \{1 \le i \le n : \Delta_i = 1\}$ and ε_n be any positive sequence such that $\varepsilon_n \to 0$ and $\varepsilon_n\tilde{\tau}_{1,p} \to 0$ as $n \to \infty$. It follows from the above decomposition

that

(A.21)

$$
\begin{aligned}
P(|\tilde{\sigma}_j^2 - \sigma_j^2| > \varepsilon_n \sigma_j^2/2) \le\ & P(\frac{1}{n}|\sum_{i\in S}[(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})^2 - (\sigma_j^{(1)})^2]| > \frac{\varepsilon_n\sigma_j^2}{12}) \\
& + P(\frac{1}{n}|\sum_{i\in S^c}[(\tilde{z}_{ij}^{(2)})^2 - (\sigma_j^{(2)})^2]| > \frac{\varepsilon_n\sigma_j^2}{12}) \\
& + P(\frac{1}{n}|\sum_{i=1}^{n}[(\Delta_i - \pi)^2 - \pi(1-\pi)](\tilde{\mu}_j^{(1)})^2| > \frac{\varepsilon_n\sigma_j^2}{12}) \\
& + P(\frac{1}{n}|(\sum_{i=1}^{n}\Delta_i - n\pi)[(\sigma_j^{(1)})^2 - (\sigma_j^{(2)})^2]| > \frac{\varepsilon_n\sigma_j^2}{12}) \\
& + P(\frac{1}{n}|\sum_{i=1}^{n}2(\Delta_i - \pi)\tilde{\mu}_j^{(1)}[\Delta_i(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}) + (1-\Delta_i)\tilde{z}_{ij}^{(2)}]| > \frac{\varepsilon_n\sigma_j^2}{12}) \\
& + P\{(\bar{\tilde{z}}_j - \tilde{\mu}_j)^2 > \frac{\varepsilon_n\sigma_j^2}{12}\}.
\end{aligned}
$$

We will bound the six terms on the right hand side above one by one using some deviation results. The same notation C will be used to denote a generic positive constant without loss of generality.

Recall that $n_1 = \sum_{i=1}^n \Delta_i$ and $n_2 = n - n_1$. By Lemma 6, σ_j^2 are uniformly bounded from below by $\tilde{\tau}_{1,p}$. Thus, by the deviation of sub-Gaussian in (A.1), conditioning on any realization of $\boldsymbol{\Delta} = (\Delta_1, \cdots, \Delta_n)^T$, we obtain

$$
\begin{aligned}
& P(\frac{1}{n}|\sum_{i\in S}[(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})^2 - (\sigma_j^{(1)})^2]| > \frac{\varepsilon_n\sigma_j^2}{12}|\boldsymbol{\Delta}) \\
\le\ & P(\frac{1}{n_1}|\sum_{i\in S}[(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})^2 - (\sigma_j^{(1)})^2]| > \frac{\varepsilon_n\tilde{\tau}_{1,p}}{12}|\boldsymbol{\Delta}) \le C\exp(-\rho\varepsilon_n^2\tilde{\tau}_{1,p}^2 n_1/12^2) \\
\le\ & \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2\sum_{i=1}^{n}\Delta_i).
\end{aligned}
$$

Since $\varepsilon_n\tilde{\tau}_{1,p} \to 0$ as $n \to \infty$ and $\sum_{i=1}^n \Delta_i$ is a Binomial random variable with probability of success π, for sufficiently large n such that $\varepsilon_n\tilde{\tau}_{1,p}$ is small

enough, taking expectation on both sides above yields

(A.22)
$$P\Big(\frac{1}{n}\Big|\sum_{i\in S}[(\tilde{z}_{ij}^{(1)}-\tilde{\mu}_j^{(1)})^2-(\sigma_j^{(1)})^2]\Big|>\frac{\varepsilon_n\sigma_j^2}{12}\Big)\le E\{\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2\sum_{i=1}^n\Delta_i)\}$$
$$=\{1-\pi+\pi\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2)\}^n=\exp\{n\ln[1-\pi+\pi\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2)]\}$$
$$\le\ \exp\{-n\pi[1-\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2)]\}\le\exp(-2n\pi C\varepsilon_n^2\tilde{\tau}_{1,p}^2)=\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2n),$$

where we have used the inequalities that $\log(1+x)\le x$ for any $x>-1$, and $1-\exp(-x)\le 2x$ for sufficiently small $x>0$. This gives an upper bound on the first term in (A.21).

Similar to (A.22), the second term in (A.21) can be bounded as

(A.23)
$$P\Big(\frac{1}{n}\Big|\sum_{i\in S^c}[(\tilde{z}_{ij}^{(2)})^2-(\sigma_j^{(2)})^2]\Big|>\frac{\varepsilon_n\sigma_j^2}{12}\Big)\le\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2n).$$

As $\tilde{\mathbf{z}}^{(1)}$ is sub-Gaussian distributed, it follows from Lemma 11 that $|\tilde{\mu}_j^{(1)}|$ are uniformly bounded from above by some positive constant across j. Since Δ_i are independently Bernoulli distributed with success probability π, we know that $E\{(\Delta_i-\pi)^2\}=\pi(1-\pi)$ and $(\Delta_i-\pi)^2$ are i.i.d. and bounded from above by 1. By Hoeffding's inequality, we have the following bound for the third term in (A.21),

(A.24)
$$P\Big(\frac{1}{n}\Big|\sum_{i=1}^n[(\Delta_i-\pi)^2-\pi(1-\pi)](\tilde{\mu}_j^{(1)})^2\Big|>\frac{\varepsilon_n\sigma_j^2}{12}\Big)$$
$$\le P\Big(\Big|\frac{1}{n}\sum_{i=1}^n(\Delta_i-\pi)^2-\pi(1-\pi)\Big|>\frac{\varepsilon_n\tilde{\tau}_{1,p}}{12(\tilde{\mu}_j^{(1)})^2}\Big)\le 2\exp\Big(-\frac{2\varepsilon_n^2\tilde{\tau}_{1,p}^2n}{12^2(\tilde{\mu}_j^{(1)})^4}\Big)$$
$$\le\ \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2n).$$

Due to the fact that $\sigma_j^2=\pi(\sigma_j^{(1)})^2+(1-\pi)(\sigma_j^{(2)})^2+\pi(1-\pi)(\tilde{\mu}_j^{(1)})^2$, we have $\sigma_j^2\ge\pi(1-\pi)[(\sigma_j^{(1)})^2+(\sigma_j^{(2)})^2]$. Similar to (A.24), by applying

Hoeffding's inequality, the fourth term in (A.21) can be bounded as

(A.25)
$$\begin{aligned}
&P\Big(\frac{1}{n}\Big|\Big(\sum_{i=1}^{n}\Delta_i - n\pi\Big)[(\sigma_j^{(1)})^2 - (\sigma_j^{(2)})^2]\Big| > \frac{\varepsilon_n\sigma_j^2}{12}\Big)\\
&\le P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\Delta_i - \pi\Big|\Big[\frac{\sigma_j^2}{\pi(1-\pi)}\Big] > \frac{\varepsilon_n\sigma_j^2}{12}\Big) \le P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\Delta_i - \pi\Big| > \frac{\pi(1-\pi)\varepsilon_n}{12}\Big)\\
&\le 2\exp\Big(-\frac{2n\pi^2(1-\pi)^2\varepsilon_n^2}{12^2}\Big) \le \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n).
\end{aligned}$$

For the fifth term in (A.21), we first decompose it as

(A.26)
$$\begin{aligned}
&P\Big(\frac{1}{n}\Big|\sum_{i=1}^{n}2(\Delta_i-\pi)\tilde{\mu}_j^{(1)}[\Delta_i(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}) + (1-\Delta_i)\tilde{z}_{ij}^{(2)}]\Big| > \frac{\varepsilon_n\sigma_j^2}{12}\Big)\\
&\le P\Big(\frac{1}{n}\Big|\sum_{i\in S}2(1-\pi)(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}) - \sum_{i\in S^c}2\pi\tilde{z}_{ij}^{(2)}]\Big| > \frac{\varepsilon_n\sigma_j^2}{12\tilde{\mu}_j^{(1)}}\Big)\\
&\le P\Big(\frac{1}{n}\Big|\sum_{i\in S}2(1-\pi)(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})\Big| > \frac{\varepsilon_n\sigma_j^2}{24\tilde{\mu}_j^{(1)}}\Big) + P\Big(\frac{1}{n}\Big|\sum_{i\in S^c}2\pi\tilde{z}_{ij}^{(2)}\Big| > \frac{\varepsilon_n\sigma_j^2}{24\tilde{\mu}_j^{(1)}}\Big).
\end{aligned}$$

Recall that $\sigma_j^2 \ge \pi(\sigma_j^{(1)})^2$ and $\max_{1\le j\le p}|\tilde{\mu}_j^{(1)}|$ can be bounded from above by some positive constant. Applying Bernstein's inequality to the sum of independent sub-Gaussian random variables yields

$$\begin{aligned}
&P\Big(\frac{1}{n}\Big|\sum_{i\in S}2(1-\pi)(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})\Big| > \frac{\varepsilon_n\sigma_j^2}{24\tilde{\mu}_j^{(1)}}\Big|\boldsymbol{\Delta}\Big)\\
&\le P\Big(\frac{1}{n_1}\Big|\sum_{i\in S}\Big(\frac{\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)}}{\sigma_j^{(1)}}\Big)\Big| > \frac{\sqrt{\pi}\varepsilon_n\sigma_j}{48(1-\pi)\tilde{\mu}_j^{(1)}}\Big|\boldsymbol{\Delta}\Big)\\
&\le \exp\Big(-\frac{\pi\varepsilon_n^2\tilde{\tau}_{1,p}n_1}{96^2(1-\pi)^2(\tilde{\mu}_j^{(1)})^2}\Big) \le \exp\Big(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2\sum_{i=1}^{n}\Delta_i\Big).
\end{aligned}$$

Applying the same argument as in (A.22), taking expectation on both sides above then we can obtain

$$P\Big(\frac{1}{n}\Big|\sum_{i\in S}2(1-\pi)(\tilde{z}_{ij}^{(1)} - \tilde{\mu}_j^{(1)})\Big| > \frac{\varepsilon_n\sigma_j^2}{24\tilde{\mu}_j^{(1)}}\Big) \le \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n).$$

Similarly we have

$$P\left(\frac{1}{n}|\sum_{i\in S^c} 2\pi\tilde{z}_{ij}^{(2)}| > \frac{\varepsilon_n\sigma_j^2}{24\tilde{\mu}_j^{(1)}}\right) \le \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n).$$

In view of (A.26), the two bounds we obtained yield

(A.27)
$$P(\frac{1}{n}|\sum_{i=1}^{n} 2(\Delta_i-\pi)\tilde{\mu}_j^{(1)}[\Delta_i(\tilde{z}_{ij}^{(1)}-\tilde{\mu}_j^{(1)})+(1-\Delta_i)\tilde{z}_{ij}^{(2)}]| > \frac{\varepsilon_n\sigma_j^2}{12}) \le \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n),$$

which is the upper bound for the fifth term in (A.21).

Applying Bernstein's inequality similarly to the sixth term in (A.21) gives

$$P\{(\bar{\tilde{z}}_j-\tilde{\mu}_j)^2 > \frac{\varepsilon_n\sigma_j^2}{12}\} \le P(|\frac{\bar{\tilde{z}}_j-\tilde{\mu}_j}{\sigma_j}| > \sqrt{\frac{\varepsilon_n}{12}}) \le \exp(-\frac{\varepsilon_n n}{24}).$$

Combining the six bounds for the terms in (A.21) that we have obtained yields

(A.28)
$$P(|\tilde{\sigma}_j^2/\sigma_j^2-1| > \varepsilon_n/2) \le P(|\tilde{\sigma}_j^2-\sigma_j^2| > \varepsilon_n\sigma_j^2/2) \le \exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n).$$

It follows that

$$\begin{aligned}(\text{A.29})\qquad P(\max_{1\le j\le p}|\tilde{\sigma}_j^2/\sigma_j^2-1| > \varepsilon_n/2) &\le \sum_{1\le j\le p} P(|\tilde{\sigma}_j^2/\sigma_j^2-1| > \varepsilon_n/2)\\ &\le p\exp(-C\varepsilon_n^2\tilde{\tau}_{1,p}^2 n).\end{aligned}$$

Let $\varepsilon_n = cn^{-\kappa}/4$ with constants c and κ defined in Condition 3. Since $\tilde{\tau}_{1,p} \le 1$, we know that $\varepsilon_n\tilde{\tau}_{1,p} \to 0$ as $n \to \infty$. Thus we can replace ε_n with $cn^{-\kappa}/4$ in (A.29) and get

$$(\text{A.30})\qquad P(\max_{1\le j\le p}|\tilde{\sigma}_j^2/\sigma_j^2-1| > cn^{-\kappa}/8) \le p\exp(-C\tilde{\tau}_{1,p}^2 n^{1-2\kappa}).$$

This completes the proof of Lemma 7. □

B.3. Lemma 8 and its proof.

LEMMA 8. Under model setting (2) and Condition 2, for sufficiently large n, with probability at least $1-p^2\exp(-\tilde{C}_2 n)$, it holds that

$$\|(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\|_{\max}/n \le c_3\tau_{2,p}$$

for some positive constants $\tilde{C}_2$ and $c_3 > 2$, where $\overline{\mathbf{Z}} = \mathbf{1}_n(\bar{z}_1,\cdots,\bar{z}_p)$ with $\bar{z}_j = \sum_{i=1}^n z_{ij}/n$ and $\mathbf{1}_n$ the $n\times 1$ column vector with all components 1.

PROOF. The deviation bound of $\|(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\|_{\max}/n$ can be obtained by bounding each component of $(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})$. Recall that $\mu_j = E(\bar{z}_j) = \pi\mu_j^{(1)}$ with $\mu_j^{(1)} = E(z_{ij}^{(1)})$. Note that the jth diagonal component of $(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})/n$ can be bounded as

$$\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)^2/n = \sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n - (\bar{z}_j-\mu_j)^2 \le \sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n, \tag{A.31}$$

Let $\mathbf{Z} = (z_{ij})_{n\times p}$, then the components of each row in $\mathbf{Z}$ are independent and identically distributed (i.i.d.) with a mixture sub-Gaussian distribution which can be written as $z_{ij} = \Delta_i z_{ij}^{(1)} + (1-\Delta_i)z_{ij}^{(2)}$, where Δ_i are i.i.d. Bernoulli random variables. Denote by $\mathbf{\Sigma}_1 = (\sigma_{ij}^{(1)})_{p\times p}$ and $\mathbf{\Sigma}_2 = (\sigma_{ij}^{(2)})_{p\times p}$. Then for each $j = 1,\cdots,p$, the random variables $z_{ij}^{(1)} - \mu_j^{(1)}$ and $z_{ij}^{(2)}$ are independent across i with mean 0 and variances $\sigma_{jj}^{(1)}$ and $\sigma_{jj}^{(2)}$, respectively. We will then bound $\sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n$ by some deviation results. The same notation $\tilde{C}_2$ will be used to denote a generic constant without loss of generality.

For any $1 \le i \le n$, we have the following decomposition for $\sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n$,

$$\begin{aligned}\sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n = n^{-1}\sum_{i=1}^{n}\{&\Delta_i(z_{ij}^{(1)}-\mu_j^{(1)})^2 + (1-\Delta_i)(z_{ij}^{(2)})^2 + (\Delta_i-\pi)^2(\mu_j^{(1)})^2 \\ &+ 2(\Delta_i-\pi)\mu_j^{(1)}[\Delta_i(z_{ij}^{(1)}-\mu_j^{(1)}) + (1-\Delta_i)(z_{ij}^{(2)})]\}.\end{aligned} \tag{A.32}$$

Recall that $S = \{1 \le i \le n : \Delta_i = 1\}$, $n_1 = \sum_{i=1}^{n}\Delta_i$ and $n_2 = n - n_1$. By Condition 2, both $\sigma_{jj}^{(1)}$ and $\sigma_{jj}^{(2)}$ can be uniformly bounded from above by $\tau_{2,p}$ across j. For the first term in (A.32) and any positive constant c_3, applying the argument with sub-Gaussian deviation and Taylor expansion similarly as in (A.22) gives

$$\begin{aligned}&P(n^{-1}\sum_{i=1}^{n}\{\Delta_i(z_{ij}^{(1)}-\mu_j^{(1)})^2\} > \tau_{2,p}+c_3) \le P(n^{-1}\sum_{i\in S}(z_{ij}^{(1)}-\mu_j^{(1)})^2 > \sigma_{jj}^{(1)}+c_3) \\ &\le P(n^{-1}|\sum_{i\in S}[(z_{ij}^{(1)}-\mu_j^{(1)})^2 - \sigma_{jj}^{(1)}]| > c_3) \le \exp(-\tilde{C}_2 n).\end{aligned}$$

Similarly, we have for the second term in (A.32),

$$P(n^{-1}\sum_{i=1}^{n}(1-\Delta_i)(z_{ij}^{(2)})^2 > \tau_{2,p}+c_3) \le \exp(-\tilde{C}_2 n).$$

Since $|\mu_j^{(1)}|$ are uniformly bounded from above by some positive constant across j from Lemma 11, similar to (A.24), by Hoeffding's inequality, we have the following bound for the third term in (A.32),

$$P\{n^{-1}\sum_{i=1}^{n}(\Delta_i-\pi)^2(\mu_j^{(1)})^2>\pi(1-\pi)(\mu_j^{(1)})^2+c_3\}\le\exp(-\tilde{C}_2n).$$

For the last term in (A.32), similar to (A.27), by Bernstein's inequality we have

$$P\{n^{-1}\sum_{i=1}^{n}2(\Delta_i-\pi)\mu_j^{(1)}[\Delta_i(z_{ij}^{(1)}-\mu_j^{(1)})+(1-\Delta_i)(z_{ij}^{(2)})]>c_3\}\le\exp(-\tilde{C}_2n).$$

In view of (A.32), by the similar argument as in (A.21), combining the four bounds we have obtained gives

$$P(\sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n>2\tau_{2,p}+C_3)\le\exp(-\tilde{C}_2n),$$

where $C_3=\pi(1-\pi)(\mu_j^{(1)})^2+4c_3$. As shown in (A.31), $\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)^2/n$ can be bounded from above by $\sum_{i=1}^{n}(z_{ij}-\mu_j)^2/n$. Thus we have

$$P(\sum_{i=1}^{n}(z_{ij}-\bar{z}_j)^2/n>2\tau_{2,p}+C_3)\le\exp(-\tilde{C}_2n).$$

Then for the (i,j)th component of $(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})$, by the Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}
&P(\sum_{k=1}^{n}(z_{ki}-\bar{z}_i)(z_{kj}-\bar{z}_j)/n>2\tau_{2,p}+C_3)\\
\le&P\{[n^{-1}\sum_{k=1}^{n}(z_{ki}-\bar{z}_i)^2][n^{-1}\sum_{k=1}^{n}(z_{kj}-\bar{z}_j)^2]>[2\tau_{2,p}+C_3]^2\}\\
\le&P\{n^{-1}\sum_{k=1}^{n}(z_{ki}-\bar{z}_i)^2>2\tau_{2,p}+C_3\}+P\{n^{-1}\sum_{k=1}^{n}(z_{kj}-\bar{z}_j)^2>2\tau_{2,p}+C_3\}\\
\le&\exp(-\tilde{C}_2n).
\end{aligned}$$

Since the above inequality holds for any (i,j)th component of $(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})$, similar to (A.29), we conclude that

$$P(\|(\mathbf{Z}-\overline{\mathbf{Z}})^T(\mathbf{Z}-\overline{\mathbf{Z}})\|_{\max}/n>2\tau_{2,p}+C_3)\le p^2\exp(-\tilde{C}_2n)$$

As $\tau_{2,p}$ is allowed to diverge, we can choose some constant $c_3 > 2$ such that $c_3\tau_{2,p} > 2\tau_{2,p} + C_3$, which gives

$$P(\|(\mathbf{Z} - \overline{\mathbf{Z}})^T(\mathbf{Z} - \overline{\mathbf{Z}})\|_{\max}/n > c_3\tau_{2,p}) \le p^2 \exp(-\tilde{C}_2 n). \tag{A.33}$$

It completes the proof of Lemma 8.

□

B.4. Lemma 9 and its proof.

LEMMA 9. Assume that $\mathbf{z}^{(1)} \in \mathbb{R}^p$ and $\mathbf{z}^{(2)} \in \mathbb{R}^p$ are sub-Gaussian, and Δ follows a Bernoulli distribution with probability of success π. Let $\mathbf{z} = \Delta\mathbf{z}^{(1)} + (1 - \Delta)\mathbf{z}^{(2)}$. Then $\mathbf{z}$ is also sub-Gaussian.

PROOF. Since $\mathbf{z}^{(1)} \in \mathbb{R}^p$ is sub-Gaussian, there exists some positive constants a_1 and b_1 such that $P(|\mathbf{v}^T\mathbf{z}^{(1)}| > t) \le a_1 \exp(-b_1 t^2)$ for any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$. Similarly, there exists some positive constants a_2 and b_2 such that $P(|\mathbf{v}^T\mathbf{z}^{(2)}| > t) \le a_2 \exp(-b_2 t^2)$ for any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$ since $\mathbf{z}^{(2)} \in \mathbb{R}^p$ is sub-Gaussian.

Let $a_3 = \max\{a_1, a_2\}$ and $b_3 = \min\{b_1, b_2\}$. Then we have

$$P(|\mathbf{v}^T\mathbf{z}^{(k)}| > t) \le a_3 \exp(-b_3 t^2)$$

for $k = 1, 2$. This, together with the law of total probability, yields

$$\begin{aligned} &P(|\mathbf{v}^T\mathbf{z}| > t) \\ =&P(|\mathbf{v}^T\mathbf{z}| > t|\Delta = 1)P(\Delta = 1) + P(|\mathbf{v}^T\mathbf{z}| > t|\Delta = 0)P(\Delta = 0) \\ =&\pi P(|\mathbf{v}^T\mathbf{z}^{(1)}| > t) + (1 - \pi)P(|\mathbf{v}^T\mathbf{z}^{(2)}| > t) \le a_3 \exp(-b_3 t^2). \end{aligned}$$

Thus $\mathbf{z}$ is sub-Gaussian. □

B.5. Lemma 10 and its proof.

LEMMA 10. If a random vector $\mathbf{w} = (W_1, \cdots, W_p)^T \in \mathbb{R}^p$ is sub-Gaussian, then so is $\mathbf{w} - E(\mathbf{w})$.

PROOF. If $\mathbf{w}$ is sub-Gaussian, then there exist some positive constants a_1 and b_1 such that

$$P(|\mathbf{v}^T\mathbf{w}| > t) \le a_1 \exp(-b_1 t^2)$$

for any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$. Let $\boldsymbol{\mu} = E(\mathbf{w})$. Thus from the Cauchy-Schwarz inequality and Lemma 11, we have

$$(\mathbf{v}^T\boldsymbol{\mu})^2 = E^2(\mathbf{v}^T\mathbf{w}) \le E[(\mathbf{v}^T\mathbf{w})^2] \le (1+a_1)c^{-1}$$

with $c = b_1/2$. Note that $(|a-b|)^2 \le (|a|+|b|)^2 \le 2(a^2+b^2)$ for any a and b. Thus, for any vector $\mathbf{v} \in \mathbb{R}^p$ satisfying $\|\mathbf{v}\|_2 = 1$ and any $t > 0$,

$$\begin{aligned}
&P\{|\mathbf{v}^T[\mathbf{w} - E(\mathbf{w})]| > t\} \le P\{|\mathbf{v}^T\mathbf{w}| + |\mathbf{v}^T\boldsymbol{\mu}| > t\} \\
\le& P\{2(\mathbf{v}^T\mathbf{w})^2 + 2(\mathbf{v}^T\boldsymbol{\mu})^2 > t^2\} \le P\{e^{c(\mathbf{v}^T\mathbf{w})^2} > e^{ct^2/2 - c(\mathbf{v}^T\boldsymbol{\mu})^2}\} \\
\le& e^{-ct^2/2 + c(\mathbf{v}^T\boldsymbol{\mu})^2} E[e^{c(\mathbf{v}^T\mathbf{w})^2}] \le (1+a_1)e^{1+a_1} \cdot e^{-ct^2/2} = a_2 \exp(-b_2 t^2)
\end{aligned}$$

with $a_2 = (1+a_1)e^{1+a_1}$ and $b_2 = c/2 = b_1/4$. Thus $\mathbf{w} - E(\mathbf{w})$ is sub-Gaussian. □

B.6. Lemma 11 and its proof.

LEMMA 11. Let W be a nonnegative random variable such that $P(W > t) \le a\exp(-bt^2)$ for any $t > 0$, where a and b are positive constants. Then $E[\exp(2^{-1}bW^2)] \le 1 + a$ and $E(W^{2m}) \le (1+a)(2/b)^m m!$ for any integer $m \ge 0$.

PROOF. Denote by $F(t)$ the cumulative distribution function of W. Then for all $x > 0$, we have $1 - F(t) = P(W > t) \le a\exp(-bt^2)$. For any constant $0 < c < b$, integration by parts yields

$$\begin{aligned}
E(e^{cW^2}) &= -\int_0^\infty e^{ct^2}\, d[1-F(t)] = 1 + \int_0^\infty 2cte^{ct^2}[1-F(t)]\, dt \\
&\le 1 + \int_0^\infty 2cate^{-(b-c)t^2}\, dt = 1 + \frac{ca}{b-c}.
\end{aligned}$$

Thus letting $c = b/2$ gives $E[\exp(2^{-1}bW^2)] \le 1 + a$.

Note that

$$\frac{2^{-m}b^m E(W^{2m})}{m!} \le \sum_{k=0}^\infty \frac{E(2^{-1}bW^2)^k}{k!} = E[\exp(2^{-1}bW^2)] \le 1 + a.$$

This implies $E(W^{2m}) \le (1+a)(2/b)^m m!$ for any integer $m \ge 0$. This completes the proof of Lemma 11. □

B.7. Lemma 12.

LEMMA 12 (Lemma 8 of Hao and Zhang [2]). Let $W_1, \cdots, W_n$ be independent random variables with zero mean. If $E[\exp(T_0|W_i|^\alpha)] \le \tilde{c}_1$ for some constants $T_0 > 0$, $\tilde{c}_1 > 0$ and $0 < \alpha \le 1$, then there exist positive constants $\tilde{c}_2$ and $\tilde{c}_3$ such that

$$P(|n^{-1}\sum_{i=1}^{n} W_i| > \varepsilon) \le \tilde{c}_2 \exp(\tilde{c}_3 n^\alpha \varepsilon^2)$$

for any $0 < \varepsilon \le 1$.

B.8. Lemma 13 and its proof.

LEMMA 13. Assume that $\max_{1\le j\le p} E[\exp(\tilde{c}_1 Z_{1j}^2)] \le \tilde{c}_2$ holds with some positive constants $\tilde{c}_1$ and $\tilde{c}_2$. If for each $1 \le j \le p$, the random variables $Z_{1j}, \cdots, Z_{nj}$ are independent and identically distributed, then

$$P\left\{|n^{-1}\sum_{i=1}^{n}[Z_{ij} - E(Z_{ij})]| \ge \varepsilon\right\} \le \tilde{c}_3 \exp(-\tilde{c}_4 n\varepsilon^2)$$

$$P\left\{|n^{-1}\sum_{i=1}^{n}[Z_{ij}Z_{ik} - E(Z_{ij}Z_{ik})]| \ge \varepsilon\right\} \le \tilde{c}_3 \exp(-\tilde{c}_4 n\varepsilon^2)$$

$$P\left\{|n^{-1}\sum_{i=1}^{n}[Z_{ij}Z_{ik}Z_{i\ell} - E(Z_{ij}Z_{ik}Z_{i\ell})]| \ge \varepsilon\right\} \le \tilde{c}_3 \exp(-\tilde{c}_4 n^{2/3}\varepsilon^2)$$

$$P\left\{|n^{-1}\sum_{i=1}^{n}[Z_{ik}Z_{i\ell}Z_{ik'}Z_{i\ell'} - E(Z_{ik}Z_{i\ell}Z_{ik'}Z_{i\ell'})]| \ge \varepsilon\right\} \le \tilde{c}_3 \exp(-\tilde{c}_4 n^{1/2}\varepsilon^2)$$

for any $0 < \varepsilon < 1$, where $1 \le j, k, \ell, k', \ell' \le p$, and $\tilde{c}_3$ and $\tilde{c}_4$ are generic positive constants which may vary from line to line.

PROOF. The idea of the proof is to use Lemma 12 and similar to that for Lemma 9 in Hao and Zhang [2]. So we omit the details here. □

REFERENCES

[1] Cai, T., Zhang, C.-H., and Zhou, H. (2008). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, 2118–2144.

[2] Hao, N. and Zhang, H. H. (2014). Interaction Screening for Ultra-High Dimensional Data. *J. Amer. Statist. Assoc.*, to appear.

[3] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
USA
E-mail: fanyingy@marshall.usc.edu
daojili@marshall.usc.edu

Department of Preventive Medicine
Keck School of Medicine
University of Southern California
Los Angeles, CA 90089
USA
E-mail: yinfeiko@usc.edu

Department of Mathematics
University of Southern California
Los Angeles, CA 90089
USA
E-mail: zeminzhe@usc.edu