SUPPLEMENT TO "LINEAR HYPOTHESIS TESTING FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS"

By Chengchun Shi*, Rui Song*, Zhao Chen†, and Runze Li†

North Carolina State University and Pennsylvania State University

In this Web Appendix, we compare the power of the proposed test statistics with existing tests in the literature in Section S1. Section S2 proves the convergence rate of $\hat{\phi}$ proposed in Section 3.3.2. Section S3 discusses the asymptotic powers of partial penalized test statistics. Discussions of Condition (A1), (A2), (A3) and (A4) are presented in Section S4. Proofs and technical lemmas are given in Section S5 and Section S6 respectively. Section S7 contains the real data analysis. Section S8 includes additional simulation studies for Poisson regression model and additional tables and plots.

APPENDIX S1: POWER COMPARISONS

In this section, we consider the following class of null hypothesis: $H_0 : \beta_{0,j} = 0$ and compare the power of our tests with the Wald test in van de Geer et al. (2014) and the decorrelated score test statistic in Ning and Liu (2017). Without loss of generality, we fix $j = 1$. Consider the following local alternative hypothesis $H_a : \beta_{0,1} = n^{-1/2}h$ for some $h \neq 0$. According to Corollary 3.1, with some calculations, we can show that for $T = T_W, T_S$ and T_L, the power function takes the form

$$\begin{aligned}\Pr(T > \chi^2_\alpha(1)) &= \Pr(\chi^2(r,\gamma_n) > \chi^2_\alpha(1)) + o(1) = \Pr(|\mathbb{Z} + \sqrt{\gamma_n}| > z_{\frac{\alpha}{2}}) + o(1)\\ &= \Pr(\mathbb{Z} > z_{\frac{\alpha}{2}} - \sqrt{\gamma}_n) + \Pr(\mathbb{Z} < -z_{\frac{\alpha}{2}} - \sqrt{\gamma}_n) + o(1)\\ &= g(\alpha,\gamma_n) + o(1), \qquad \text{(S1.1)}\end{aligned}$$

where $g(\alpha,\gamma) = \Pr(\mathbb{Z} > z_{\alpha/2} - \sqrt{\gamma}) + \Pr(\mathbb{Z} \geq z_{\alpha/2} + \sqrt{\gamma})$ for any $\gamma \geq 0$, $\mathbb{Z}$ denotes a standard normal variable, z_α is the upper αth quantile of a standard normal distribution, and $\gamma_n = \left(\phi_0 \mathbf{e}^T_{1,1+s}\mathbf{\Omega}_n\mathbf{e}_{1,1+s}\right)^{-1} h^2$ where $\mathbf{e}_{1,q}$ denotes the basis vector of length q, with the first element equal to 1 and other elements equal to 0. To simplify the presentation, we consider a

*Supported by NSF grant DMS 1555244, NCI grant P01 CA142538.

†Supported by NSF grant DMS 1512422, NIH grants P50 DA039838 and P50 DA036107, and T32 LM012415, and NNSFC grants 11690014 and 11690015.

random design setting where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed, according to the distribution of $\boldsymbol{X}_0$. Define

$$\boldsymbol{\Phi} = \mathrm{E}\left\{\boldsymbol{X}_0 b''(\boldsymbol{X}_0\beta_0)\boldsymbol{X}_0^T\right\}.$$

Under certain regularity conditions, it follows from (S1.1) that

$$\Pr(T > \chi^2_\alpha(1)) = g\left(\alpha, \{\phi_0 \mathbf{e}_{1,1+s}^T(\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S})^{-1}\mathbf{e}_{1,1+s}\}^{-1}h^2\right) + o(1).$$

S1.0.1. *Comparison with the Wald test based on the de-sparsified Lasso estimator.* van de Geer et al. (2014) proposed to de-sparsify the Lasso estimator for constructing statistical tests for $\beta_{0,j}$ in a high dimensional generalized linear model. Define the Lasso estimator

$$\hat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n}\sum_{i=1}^n \{b(\boldsymbol{\beta}^T\boldsymbol{X}_i) - Y_i\boldsymbol{\beta}^T\boldsymbol{X}_i\} + \lambda_n\|\boldsymbol{\beta}\|_1 \right).$$

The de-sparsified Lasso estimator $\hat{\beta}^{DL}$ is given by

$$\hat{\boldsymbol{\beta}}^{DL} = \hat{\boldsymbol{\beta}}^L + \frac{1}{n}\widehat{\boldsymbol{\Theta}}\boldsymbol{X}^T\{\boldsymbol{Y} - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}}^L)\},$$

where $\widehat{\boldsymbol{\Theta}}$ is some consistent estimator for $\boldsymbol{\Phi}^{-1}$ obtained by nodewise regression. It follows from Theorem 3.1 in van de Geer et al. (2014) that

$$\sqrt{n}(\hat{\beta}_j^{DL} - \beta_{0,j})/\hat{\sigma}_j \sim N(0,1), \tag{S1.2}$$

where

$$\hat{\sigma}_j^2 = \frac{1}{n}\left(\boldsymbol{e}_{j,p}^T\widehat{\boldsymbol{\Theta}}\boldsymbol{X}^T\{\boldsymbol{Y} - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}}^L)\}\{\boldsymbol{Y} - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}}^L)\}^T\boldsymbol{X}\widehat{\boldsymbol{\Theta}}\boldsymbol{e}_{j,p}\right).$$

Moreover, under the model assumption (1.1), it follows from Corollary 3.1 and the proof of Theorem 3.1 in van de Geer et al. (2014) that

$$\hat{\sigma}_j^2 = \phi_0\boldsymbol{e}_{j,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{j,p} + o_p(1). \tag{S1.3}$$

To test $H_0 : \beta_{0,1} = 0$, van de Geer et al. (2014) considered the Wald-type statistic and proposed to reject H_0 when $\sqrt{n}|\hat{\beta}_1^{DL}| > z_{\alpha/2}\hat{\sigma}_1$ for a given α. It follows from (S1.2) that such test statistic has correct size under H_0. Assume $\liminf_n \mathbf{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\mathbf{e}_{1,p} > 0$, under the local alternative, it follows from (S1.2)

and (S1.3) that

$$\begin{aligned}
& \Pr(\sqrt{n}|\hat{\beta}_1^{DL}| > z_{\alpha/2}\hat{\sigma}_1) = \Pr(\sqrt{n}(\hat{\beta}_1^{DL} - \beta_{0,1}) > z_{\alpha/2}\hat{\sigma}_1 - h) \\
+ \quad & \Pr(\sqrt{n}(\hat{\beta}_1^{DL} - \beta_{0,1}) < -z_{\alpha/2}\hat{\sigma}_1 - h) \\
= \quad & \Pr\left(\frac{\sqrt{n}(\hat{\beta}_1^{DL} - \beta_{0,1})}{\hat{\sigma}_1} > z_{\alpha/2} - \frac{h}{\sqrt{\phi_0 \mathbf{e}_{1,p}^T \mathbf{\Phi}^{-1} \mathbf{e}_{1,p}}} \right) \\
+ \quad & \Pr\left(\frac{\sqrt{n}(\hat{\beta}_1^{DL} - \beta_{0,1})}{\hat{\sigma}_1} < -z_{\alpha/2} - \frac{h}{\sqrt{\phi_0 \mathbf{e}_{1,p}^T \mathbf{\Phi}^{-1} \mathbf{e}_{1,p}}} \right) + o(1) \\
= \quad & g\left(\alpha, (\phi_0 \mathbf{e}_{1,p}^T \mathbf{\Phi}^{-1} \mathbf{e}_{1,p})^{-1} h^2\right) + o(1).
\end{aligned}$$

LEMMA S.1. *For any $p \times p$ positive definite matrix $\mathbf{\Phi}$ and any set $S \subseteq [2, 3, \ldots, p]$, we have*

$$\mathbf{e}_{1,p}^T \mathbf{\Phi}^{-1} \mathbf{e}_{1,p} \geq \mathbf{e}_{1,1+s}^T (\mathbf{\Phi}_{\{1\}\cup S, \{1\}\cup S})^{-1} \mathbf{e}_{1,1+s},$$

with equality if and only if

$$\mathbf{\Phi}_{S,\{1\}} = \mathbf{\Phi}_{S,(\{1\}\cup S)^c} (\mathbf{\Phi}_{(\{1\}\cup S)^c, (\{1\}\cup S)^c})^{-1} \mathbf{\Phi}_{(\{1\}\cup S)^c, \{1\}}.$$

REMARK S1.1. Observe that for any $0 < \alpha < 1$, $g(\alpha, \gamma)$ is strictly increasing as a function of γ. Lemma S.1 therefore suggests that

$$g\left(\alpha, (\phi_0 \mathbf{e}_{1,p}^T \mathbf{\Phi}^{-1} \mathbf{e}_{1,p})^{-1} h^2\right) \leq g\left(\alpha, \{\phi_0 \mathbf{e}_{1,1+s}^T (\mathbf{\Phi}_{\{1\}\cup S, \{1\}\cup S})^{-1} \mathbf{e}_{1,1+s}\}^{-1} h^2\right),$$

with equality if and only if

$$\mathbf{\Phi}_{S,\{1\}} = \mathbf{\Phi}_{S,(\{1\}\cup S)^c} (\mathbf{\Phi}_{(\{1\}\cup S)^c, (\{1\}\cup S)^c})^{-1} \mathbf{\Phi}_{(\{1\}\cup S)^c, \{1\}}.$$

This implies that our test statistics are asymptotically more powerful than the Wald-type statistic based on the de-sparsified Lasso estimator.

S1.0.2. *Comparison with the decorrelated score test.* Let $\mathcal{M}_1 = [2, \ldots, p]$. For any $\theta \in \mathbb{R}$, the decorrelated score function (Ning and Liu, 2017) is given by

$$\hat{S}_D(\theta) = \{\boldsymbol{X}_{\mathcal{M}_1} \widehat{\boldsymbol{\omega}} - \boldsymbol{X}^1\}^T \{\boldsymbol{Y} - \mu(\theta \boldsymbol{X}^1 + \boldsymbol{X}_{\mathcal{M}_1} \hat{\beta}_{\mathcal{M}_1})\},$$

for some penalized regression estimator $\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \sum_{i=1}^n \{b(\boldsymbol{\beta}^T \boldsymbol{X}_i) - Y_i \boldsymbol{\beta}^T \boldsymbol{X}_i\} + \sum_{j=1}^p p_\lambda(|\beta_j|) \right),$$

and a Dantzig type estimator $\widehat{\boldsymbol{\omega}}$,

$$\widehat{\boldsymbol{\omega}} = \arg\min \|\boldsymbol{\omega}\|_1 \quad s.t.\ \|(\boldsymbol{\omega}^T \boldsymbol{X}_{\mathcal{M}_1}^T - \boldsymbol{X}^1)\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\boldsymbol{X}_{\mathcal{M}_1}\|_\infty \le n\lambda'.$$

For testing $H_0 : \beta_{0,1} = 0$, Ning and Liu (2017) proposed the following decorrelated score test statistic,

$$T_{SD} = \frac{\sqrt{n}\hat{S}_D(0)}{\hat{\sigma}_s},$$

for some $\hat{\sigma}_s^2$ that consistently estimates $\sigma_s^2 = \phi_0(\boldsymbol{v}^*)^T\boldsymbol{\Phi}\boldsymbol{v}^*$ where $\boldsymbol{v}^* = (1, -(\boldsymbol{\omega}^*)^T)^T$ and $\boldsymbol{\omega}^* = (\boldsymbol{\Phi}_{\mathcal{M}_1,\mathcal{M}_1})^{-1}\boldsymbol{\Phi}_{\{1\}\mathcal{M}_1}$. The null hypothesis is rejected when $|T_{SD}| > z_{\alpha/2}$. It was shown in Theorem 3.1 in Ning and Liu (2017) that such tests has correct size under H_0.

Moreover, under $H_a : \beta_{0,1} = n^{-1/2}h$, it follows from Corollary D.1 in their supplementary article that, for any $0 < \alpha < 1$,

$$\begin{aligned}\Pr(|T_{SD}| > z_{\alpha/2}) &= \Pr\left(\mathbb{Z} > z_{\alpha/2} - h\sqrt{(\boldsymbol{v}^*)^T\boldsymbol{\Phi}\boldsymbol{v}^*/\phi_0}\right)\\ -\ \Pr&\left(\mathbb{Z} > z_{\alpha/2} + h\sqrt{(\boldsymbol{v}^*)^T\boldsymbol{\Phi}\boldsymbol{v}^*/\phi_0}\right) + o(1) = g(\alpha, h^2(\boldsymbol{v}^*)^T\boldsymbol{\Phi}\boldsymbol{v}^*/\phi_0) + o(1),\end{aligned}$$

where $\mathbb{Z}$ denotes a standard normal random variable. By the definition of $\boldsymbol{v}^*$, we have

$$(\boldsymbol{v}^*)^T\boldsymbol{\Phi}\boldsymbol{v}^* = \boldsymbol{\Phi}_{\{1\}\{1\}} - \boldsymbol{\Phi}_{\{1\},\mathcal{M}_1}(\boldsymbol{\Phi}_{\mathcal{M}_1,\mathcal{M}_1})^{-1}\boldsymbol{\Phi}_{\mathcal{M}_1,\{1\}} = \left(\boldsymbol{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{1,p}\right)^{-1},$$

where the last equality follows by the block matrix inversion formula (Lemma S.9). Therefore, it follows from Lemma S.1 and the monotonicity of $g(\cdot)$ that our tests are more powerful than the decorrelated score test.

APPENDIX S2: ADDITIONAL DETAILS REGARDING $\hat{\phi}$ IN LINEAR REGRESSION MODELS

By Theorem 2.1, we have with probability tending to 1,

$$\widehat{S}_a = S. \tag{S2.1}$$

Besides,

$$\|\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{a,S\cup\mathcal{M}}\|_2 = O_p\left(\sqrt{(s+m)/n}\right). \tag{S2.2}$$

Let $\varepsilon_i = Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}_0$. Under the event defined in (S2.1), we have

$$\hat{\phi} = \underbrace{\frac{1}{n-s-m}\sum_{i=1}^{n}\varepsilon_i^2}_{I_1} + \underbrace{\frac{2}{n-s-m}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{X}_{i,S\cup\mathcal{M}}^T(\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{a,S\cup\mathcal{M}})}_{I_2}$$
$$+ \underbrace{\frac{1}{n-s-m}(\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{0,S\cup\mathcal{M}})^T\left(\sum_{i=1}^{n}\boldsymbol{X}_{i,S\cup\mathcal{M}}\boldsymbol{X}_{i,S\cup\mathcal{M}}^T\right)(\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{0,S\cup\mathcal{M}})}_{I_3}.$$

By Chebyshev's inequality, we have for any $t_0 > 0$,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - \phi_0) > t_0\right) \le \frac{\sum_{i=1}^{n}\mathrm{Var}(\varepsilon_i^2)}{n^2 t_0^2}.$$

This implies

$$I_1 = \frac{n}{n-s-m}\phi_0 + O_p\left(\frac{\sqrt{n}}{n-s-m}\right).$$

Since we require $\max(s,m) = o(n)$, we have

$$I_1 = \phi_0 + O_p(n^{-1/2}). \tag{S2.3}$$

By Cauchy-Schwarz inequality,

$$|I_2| \le \underbrace{\left\|\frac{2}{n-s-m}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{X}_{i,S\cup\mathcal{M}}\right\|_2}_{I_2^*} \|\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{a,S\cup\mathcal{M}}\|_2.$$

With some calculations, we have

$$\mathrm{E}(I_2^*)^2 = \frac{4}{(n-s-m)^2}\mathrm{E}\sum_{i,j=1}^{n}\varepsilon_i\varepsilon_j \boldsymbol{X}_{i,S\cup\mathcal{M}}^T\boldsymbol{X}_{j,S\cup\mathcal{M}}$$
$$= \frac{4\phi_0}{(n-s-m)^2}\sum_{i=1}^{n}\|\boldsymbol{X}_{i,S\cup\mathcal{M}}\|_2^2 = \frac{\phi_0}{(n-s-m)^2}\mathrm{tr}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right)$$
$$\le \frac{4\phi_0(s+m)}{(n-s-m)^2}\lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O\left(\frac{n(s+m)}{(n-s-m)^2}\right) = O\left(\frac{(s+m)}{n}\right),$$

where the fourth equality is due to Condition (A1), and the last equality is due to the condition that $\max(s,m) = o(n)$. This implies

$$I_2^* = O_p\left(\frac{\sqrt{s+m}}{\sqrt{n}}\right),$$

which together with (S2.2) yields

$$I_2 = O_p\left(\frac{s+m}{n}\right). \tag{S2.4}$$

Besides, it follows (S2.2) and Condition (A1) that

$$\begin{aligned} I_3 &\leq \frac{1}{n-s-m}\|\hat{\boldsymbol{\beta}}_{a,S\cup\mathcal{M}} - \boldsymbol{\beta}_{a,S\cup\mathcal{M}}\|_2^2 \lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T \boldsymbol{X}_{S\cup\mathcal{M}}\right) \\ &= O\left(\frac{s+m}{n-s-m}\right) = O\left(\frac{s+m}{n}\right). \end{aligned}$$

Combining this together with (S2.3) and (S2.4) gives that

$$\hat{\phi} = \phi_0 + O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{s+m}{n}\right) = \phi_0 + O_p\left(\frac{1}{\sqrt{n}}\right),$$

where the last equality is due to the condition that $\max(s,m) = o(n^{1/2})$.

APPENDIX S3: ADDITIONAL DETAILS REGARDING POWERS OF PARTIAL PENALIZED TEST STATISTICS

The proposed test statistics are based on the partial penalized estimators $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_a$ defined in (2.2) and (2.3). Alternatively, we can construct our test statistics based on

$$\begin{aligned} \hat{\boldsymbol{\beta}}_0^{\mathcal{N}} &= \arg\max_{\boldsymbol{\beta}} Q_n^{\mathcal{N}}(\boldsymbol{\beta}, \lambda_{n,0}) \text{ subject to } \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}, \\ \hat{\boldsymbol{\beta}}_a^{\mathcal{N}} &= \arg\max_{\boldsymbol{\beta}} Q_n^{\mathcal{N}}(\boldsymbol{\beta}, \lambda_{n,a}), \end{aligned}$$

where

$$Q_n^{\mathcal{N}}(\boldsymbol{\beta}, \lambda) = \frac{1}{n}\sum_{i=1}^n \{Y_i \boldsymbol{\beta}^T \boldsymbol{X}_i - b(\boldsymbol{\beta}^T \boldsymbol{X}_i)\} - \sum_{j \notin \mathcal{M}\cup\mathcal{N}} p_\lambda(|\beta_j|),$$

for some set $\mathcal{N} \subseteq [1, \ldots, p]$. The set $\mathcal{N}$ can be chosen by the domain knowledge. For example, it can contain variables that the analyst thinks are important for ensuring robustness. Let $\mathcal{N}^* = \mathcal{N} \cap (\mathcal{M} \cup S)^c$. Define

$$\begin{aligned} T_L^{\mathcal{N}} &= 2n\{L_n(\hat{\boldsymbol{\beta}}_a^{\mathcal{N}}) - L_n(\hat{\boldsymbol{\beta}}_0^{\mathcal{N}})\}/\hat{\phi}, \\ T_W^{\mathcal{N}} &= (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}^{\mathcal{N}} - \boldsymbol{t})^T \left(\boldsymbol{C}\widehat{\boldsymbol{\Omega}}_{a,mm}^{\mathcal{N}} \boldsymbol{C}^T\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{a,\mathcal{M}}^{\mathcal{N}} - \boldsymbol{t})/\hat{\phi}, \\ T_S^{\mathcal{N}} &= \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0^{\mathcal{N}})\}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}\cup\mathcal{N}^*} \\ \boldsymbol{X}_{\widehat{S}_0^{\mathcal{N}}} \end{pmatrix} \widehat{\boldsymbol{\Omega}}_0^{\mathcal{N}} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}\cup\mathcal{N}^*} \\ \boldsymbol{X}_{\widehat{S}_0^{\mathcal{N}}} \end{pmatrix}^T \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0^{\mathcal{N}})\}/\hat{\phi}, \end{aligned}$$

where $\hat{\phi}$ is some constant estimators for ϕ_0, $\widehat{\boldsymbol{\Omega}}^{\mathcal{N}}_{a,mm}$ is the first m rows and columns of $\widehat{\boldsymbol{\Omega}}^{\mathcal{N}}_a$,

$$\widehat{\boldsymbol{\Omega}}^{\mathcal{N}}_a = n\begin{pmatrix} \boldsymbol{X}^T_{\mathcal{M}}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}^T_{\mathcal{M}}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_a)\boldsymbol{X}_{\widehat{S}^{\mathcal{N}}_a\cup\mathcal{N}^*} \\ \boldsymbol{X}^T_{\widehat{S}^{\mathcal{N}}_a\cup\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_a)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}^T_{\widehat{S}^{\mathcal{N}}_a\cup\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_a)\boldsymbol{X}_{\widehat{S}^{\mathcal{N}}_a\cup\mathcal{N}^*} \end{pmatrix}^{-1},$$

$$\widehat{\boldsymbol{\Omega}}^{\mathcal{N}}_0 = n\begin{pmatrix} \boldsymbol{X}^T_{\mathcal{M}\cup\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_0)\boldsymbol{X}_{\mathcal{M}\cup\mathcal{N}^*} & \boldsymbol{X}^T_{\mathcal{M}\cup\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_0)\boldsymbol{X}_{\widehat{S}^{\mathcal{N}}_0} \\ \boldsymbol{X}^T_{\widehat{S}^{\mathcal{N}}_0}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_0)\boldsymbol{X}_{\mathcal{M}\cup\mathcal{N}^*} & \boldsymbol{X}^T_{\widehat{S}^{\mathcal{N}}_0}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathcal{N}}_0)\boldsymbol{X}_{\widehat{S}^{\mathcal{N}}_0} \end{pmatrix}^{-1},$$

and

$$\widehat{S}^{\mathcal{N}}_a = \{j \in (\mathcal{M}\cup\mathcal{N}^*)^c : \hat{\beta}^{\mathcal{N}}_{a,j} \neq 0\}, \quad \widehat{S}^{\mathcal{N}}_0 = \{j \in (\mathcal{M}\cup\mathcal{N}^*)^c : \hat{\beta}^{\mathcal{N}}_{0,j} \neq 0\}.$$

For a given significance level α, we reject the null hypothesis when $T^{\mathcal{N}} > \chi^2_\alpha(r)$ for $T^{\mathcal{N}} = T^{\mathcal{N}}_L, T^{\mathcal{N}}_W$ or $T^{\mathcal{N}}_S$.

Similar to Corollary 3.1, we can show the Type I error rates of $T^{\mathcal{N}}_L, T^{\mathcal{N}}_W$ and $T^{\mathcal{N}}_S$ are close to the nominal level. Besides, under the alternative $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} - \boldsymbol{t} = \boldsymbol{h}_n$, the asymptotic power functions of these test statistics are equal to

$$\Pr(\chi^2(r, \gamma^{\mathcal{N}}_n) > \chi^2_\alpha(r)), \tag{S3.1}$$

where $\gamma^{\mathcal{N}}_n = n\boldsymbol{h}^T_n\left(\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T\right)^{-1}\boldsymbol{h}_n/\phi_0$, $\boldsymbol{\Omega}^{\mathcal{N}}_{mm}$ is the first m rows and columns of

$$\boldsymbol{\Omega}^{\mathcal{N}}_n = \left\{\frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T_{\mathcal{M}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}^T_{\mathcal{M}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{S}} & \boldsymbol{X}^T_{\mathcal{M}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{N}^*} \\ \boldsymbol{X}^T_{\mathcal{S}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}^T_{\mathcal{S}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{S}} & \boldsymbol{X}^T_{\mathcal{S}}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{N}^*} \\ \boldsymbol{X}^T_{\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}^T_{\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{S}} & \boldsymbol{X}^T_{\mathcal{N}^*}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{N}^*} \end{pmatrix}\right\}^{-1}.$$

We now show the test statistic achieves its greatest power if $\mathcal{N}^* = 0$. This means the partial penalized tests are most advantageous if each unpenalized variable is either an important variable (i.e., in $\mathcal{S}$) or a variable in $\mathcal{M}$. By (S3.1), it suffices to prove $\gamma^{\mathcal{N}}_n \leq \gamma_n$ for any $\mathcal{N}$. Assume for now, we've shown

$$\inf_{\boldsymbol{a}\in\mathbb{R}^m}\left(\boldsymbol{a}^T\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{\Omega}_{mm}\boldsymbol{a}\right) \geq 0. \tag{S3.2}$$

This implies

$$\inf_{\boldsymbol{a}\in\mathbb{R}^r}\left(\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\boldsymbol{a}\right) \geq 0, \tag{S3.3}$$

and hence the matrix $\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T - \boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$ is positive semidefinite.

Note that $\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$ is positive definite. By the eigenvalue decomposition theorem, we can find some positive definite matrix $\boldsymbol{Q}$ such that $\boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$. By (S3.3), we can similarly show that the matrix

$$\boldsymbol{Q}^{-1}\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T\boldsymbol{Q}^{-1} - \boldsymbol{I}$$

is positive semidefinite. As a result, the eigenvalues of $\boldsymbol{Q}^{-1}\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T\boldsymbol{Q}^{-1}$ are all greater than or equal to 1. This implies that the eigenvalues of $\boldsymbol{Q}(\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T)^{-1}\boldsymbol{Q}^{-1}$ are all smaller than or equal to 1. Therefore, the matrix

$$\boldsymbol{I} - \boldsymbol{Q}(\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T)^{-1}\boldsymbol{Q}$$

is positive semidefinite. This further implies that the matrix $\left(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\right)^{-1} - \left(\boldsymbol{C}\boldsymbol{\Omega}^{\mathcal{N}}_{mm}\boldsymbol{C}^T\right)^{-1}$ is positive semidefinite. Therefore, we have $\gamma_n \geq \gamma^{\mathcal{N}}_n$.

It remains to prove (S3.2). Let $\boldsymbol{\Omega}^{\mathcal{N}}_{m+s,m+s}$ be the first $(s+m)$ rows and columns of $\boldsymbol{\Omega}^{\mathcal{N}}_n$. By Lemma S.1, the matrix $\boldsymbol{\Omega}^{\mathcal{N}}_{m+s,m+s} - \boldsymbol{\Omega}_n$ is positive semidefinite. The assertion (S3.2) thus follows.

APPENDIX S4: DISCUSSION OF THE TECHNICAL CONDITIONS

S4.1. Discussion of Condition (A1). For Gaussian linear regression models, Condition (A1) reduces to (A1*) given below.

(A1*) Assume

$$\begin{aligned}
&\max_{1\leq j\leq p}\|\boldsymbol{X}^j\|_\infty = O\left(\sqrt{n/\log(p)}\right), \quad \max_{1\leq j\leq p}\|\boldsymbol{X}^j\|_2 = O(\sqrt{n}),\\
&\lambda_{\min}\left(\boldsymbol{X}^T_{S\cup\mathcal{M}}\boldsymbol{X}_{S\cup\mathcal{M}}\right) \geq cn, \quad \lambda_{\max}\left(\boldsymbol{X}^T_{S\cup\mathcal{M}}\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O(n),\\
&\|\boldsymbol{X}^T_{(S\cup\mathcal{M})^c}\boldsymbol{X}_{S\cup\mathcal{M}}\|_{2,\infty} = O(n).
\end{aligned}$$

for some constants $c > 0$.

For logistic regression and Poisson regression models, Condition (A1) is implied by (A1*) and the following conditions:

$$\max_{1\leq i\leq n}|\boldsymbol{X}_i^T\boldsymbol{\beta}_0| = O(1), \tag{S4.1}$$

$$\max_{1\leq j\leq p}\|\boldsymbol{X}^j\|_\infty = O\left(\frac{n^{1/2}}{(s+m)\log^{1/2}n}\right), \tag{S4.2}$$

$$\max_{1\leq j\leq p}\lambda_{\max}\left\{\boldsymbol{X}^T_{S\cup\mathcal{M}}\mathrm{diag}(|\boldsymbol{X}^j|)\boldsymbol{X}_{S\cup\mathcal{M}}\right\} = O(n). \tag{S4.3}$$

We note that van de Geer et al. (2014) and Ning and Liu (2017) also assumed (S4.1) to establish the asymptotic properties of de-sparsified Lasso estimator

and the decorrelated score statistic under the high dimensional generalized linear models. Assume $\max(s,m) = O(n^{l_0})$ for some $0 < l_0 < 1/2$ and that each covariate has sub-exponential tail. Then we can show (S4.2) holds with probability tending to 1.

To further simplify Condition (A1*) and (S4.3), we consider a random design setting where $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are i.i.d copies of some random vector $\boldsymbol{X}_0 \in \mathbb{R}^p$. More specifically, we consider the following two cases: (i) The bounded case: $\max_{j=[1,\ldots,p]} |\boldsymbol{X}_{0,j}| \leq \omega_0$ for some constant $\omega_0 > 0$. (ii) The sub-Gaussian case: $\sup_{\|\boldsymbol{v}\|_2 \leq 1, \boldsymbol{v} \in \mathbb{R}^p} \|\boldsymbol{v}^T \boldsymbol{X}_0\|_{\psi_2} \leq \omega_0$ where for any random variable $\mathbb{Z}$,

$$\|\mathbb{Z}\|_{\psi_2} \equiv \inf_{C>0} \left\{ \mathrm{E} \exp\left(\frac{|\mathbb{Z}^2|}{C^2} \right) \leq 2 \right\}.$$

Let $\boldsymbol{\Lambda} = \mathrm{E}\boldsymbol{X}_0 \boldsymbol{X}_0^T$.

LEMMA S.2. *Assume that* $\lambda_{\max}(\boldsymbol{\Lambda}_{S\cup\mathcal{M}, S\cup\mathcal{M}}) = O(1)$, $\lambda_{\min}(\boldsymbol{\Lambda}_{S\cup\mathcal{M}, S\cup\mathcal{M}}) \geq \bar{c}$ *for some constant* $\bar{c} > 0$, $\max(s,m) = o\{(n/\log n)^{1/2}\}$ *and* $\max(s,m)\log p = o(n)$. *Then in the bounded case, Condition (A1*) and Condition* (S4.3) *hold with probability tending to* 1.

LEMMA S.3. *Assume that* $\lambda_{\min}(\boldsymbol{\Lambda}_{S\cup\mathcal{M}, S\cup\mathcal{M}}) \geq \bar{c}$ *for some constant* $\bar{c} > 0$, $\max(s,m)\log p = o(n/\log n)$ *and* $\log p = O(\sqrt{n})$. *Then in the sub-Gaussian case, Condition (A1*) and Condition* (S4.3) *hold with probability tending to* 1.

Proofs of Lemma S.2 and Lemma S.3 are given in Section S5.3 and Section S5.4, respectively.

S4.2. Discussion of Condition (A2). The condition $p'_{\lambda_{n,j}}(d_n) = o((s+m)^{-1/2} n^{-1/2})$, $\lambda_{n,j}\kappa_{0,j} = o(1)$ automatically holds for SCAD penalty function, since $p'_{\lambda_{n,j}}(d_n) = 0$ and $\kappa_{0,j} = 0$ when $d_n \gg \lambda_{n,j}$. The Lasso penalty function doesn't satisfy the condition $p'_{\lambda_{n,j}}(d_n) = o((s+m)^{-1/2} n^{-1/2})$. Hence, the corresponding (un)constrained estimators have relatively large biases and don't have the asymptotic distributions in Theorem 2.1.

S4.3. Discussion of Condition (A3). For logistic regression models, we have $\max_i |Y_i - \mu(\boldsymbol{X}_i^T \boldsymbol{\beta}_0)| \leq 1$. It is easy to show Condition (A3) holds

with $M = 1$ and $v_0 = 2\exp(1)$. For Gaussian linear models, we have

$$\begin{aligned} & \mathrm{E}\left\{\exp\left(\frac{|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_0|}{\phi_0}\right) - 1 - \frac{|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_0|}{\phi_0}\right\} \\ \leq \quad & \mathrm{E}\exp\left(\frac{|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_0|}{\phi_0}\right) \leq 2. \end{aligned}$$

Hence, (A3) holds with $M = \phi_0$ and $v_0 = 4\phi_0^2$. Assume

$$\max_{i\in[1,\ldots,n]} |\boldsymbol{X}_i^T\boldsymbol{\beta}_0| \leq K_0,$$

for some constant $K_0 > 0$. Then, for Poisson regression models, we have

$$\begin{aligned} & \mathrm{E}\exp(|Y_i - \exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)|) \leq \mathrm{E}\exp(|Y_i| + |\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)|) \\ \leq \quad & \mathrm{E}\exp(2Y_i) = \exp\left\{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)\,(\exp(2) - 1)\right\} \leq \exp\{\exp(K_0 + 2)\}, \end{aligned}$$

where the second inequality follows by Jensen's inequality. Condition (A3) is thus satisfied.

S4.4. Discussion of Condition (A4). Condition (A4) is not restrictive at all. When the matrix $\boldsymbol{C}$ doesn't vary with n, the condition $\lambda_{\max}\left((\boldsymbol{C}\boldsymbol{C}^T)^{-1}\right) = O(1)$ is automatically satisfied since $\boldsymbol{C}$ is of full row rank.

APPENDIX S5: PROOFS

S5.1. Proof of Theorem 2.1. We focus on proving the statistical properties of $\hat{\boldsymbol{\beta}}_0$. Properties of $\hat{\boldsymbol{\beta}}_a$ can be similarly proven. We divide the proof into three steps. In the first step, we show that there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$ with the constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$ and $\boldsymbol{\beta}_{(\mathcal{M}\cup S)^c} = \boldsymbol{0}$, such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(\sqrt{(s + m - r)/n})$. In the second step, we show $\hat{\boldsymbol{\beta}}$ is indeed a local maximizer of $Q_n(\boldsymbol{\beta})$ with the linear constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = t$. This implies $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. In the final step, we show

$$\begin{aligned} \sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} \quad = \quad & \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}(\boldsymbol{I} - \boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\beta_0)\} \\ - \quad & \sqrt{n}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix} \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n \\ \boldsymbol{0} \end{pmatrix} + o_p(1). \end{aligned}$$

Step 1: Define a p-dimensional vector $\boldsymbol{\beta}^*$ as

$$\begin{cases} \boldsymbol{\beta}^*_{\mathcal{M}} = \boldsymbol{\beta}_{0,\mathcal{M}} - \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n, \\ \boldsymbol{\beta}^*_{\mathcal{M}^c} = \boldsymbol{\beta}_{0,\mathcal{M}^c}. \end{cases}$$

It is immediate to see that

$$\text{(S5.1)}\quad \boldsymbol{C}\boldsymbol{\beta}^*_{\mathcal{M}} - \boldsymbol{t} = \boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} - \boldsymbol{C}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n - \boldsymbol{t} = \boldsymbol{h}_n - \boldsymbol{h}_n = 0.$$

Besides, it follows from Assumption (A4) that

$$\begin{aligned}\text{(S5.2)}\qquad & \|\boldsymbol{\beta}^*_{\mathcal{M}\cup S} - \boldsymbol{\beta}_{0,\mathcal{M}\cup S}\|_2^2 \\ = \;& \|\boldsymbol{h}_n^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\|_2^2 = O(\|\boldsymbol{h}_n\|_2^2) = O\left(\frac{m+s-r}{n}\right).\end{aligned}$$

Therefore, it suffices to show there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$ with the constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$ and $\boldsymbol{\beta}_{(\mathcal{M}\cup S)^c} = \boldsymbol{0}$, such that $\|\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S} - \boldsymbol{\beta}^*_{\mathcal{M}\cup S}\|_2^2 = O_p((m+s-r)/n)$.

Observe that for any $\boldsymbol{\beta}$ with $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$, it follows from (S5.1) that $\boldsymbol{C}(\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}) = \boldsymbol{0}$ and hence $\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}}$ belongs to the null space of $\boldsymbol{C}$. We take a basis matrix $\boldsymbol{Z} \in \mathbb{R}^{m\times(m-r)}$ of the null space $\boldsymbol{C}$. This implies $\boldsymbol{C}\boldsymbol{Z} = 0$. Further assume $\boldsymbol{Z}$ is orthogonalized such that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}_{m-r}$ where $\boldsymbol{I}_q$ stands for a $q \times q$ identity matrix. Hence, for any $\boldsymbol{\beta}_{\mathcal{M}}$ such that $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$, there exists some $(m-r)$-dimensional vector $\boldsymbol{\nu}$ such that $\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{\beta}^*_{\mathcal{M}} = \boldsymbol{Z}\boldsymbol{\nu}$. For any $\boldsymbol{\delta} \in \mathbb{R}^{m-r+s}$, we define $\overline{Q}_n(\boldsymbol{\delta}) = Q_n(\boldsymbol{\beta}(\boldsymbol{\delta}))$ where $\boldsymbol{\beta}(\boldsymbol{\delta})$ is defined as

$$\begin{cases} \boldsymbol{\beta}(\boldsymbol{\delta})_{\mathcal{M}} = \boldsymbol{\beta}^*_{\mathcal{M}} + \boldsymbol{Z}\boldsymbol{\delta}_{J_0}, \\ \boldsymbol{\beta}(\boldsymbol{\delta})_S = \boldsymbol{\beta}_{0,S} + \boldsymbol{\delta}_{J_0^c}, \\ \boldsymbol{\beta}(\boldsymbol{\delta})_{(\mathcal{M}\cup S)^c} = \boldsymbol{\beta}_{0,\mathcal{M}\cup S^c}, \end{cases}$$

where $J_0 = [1, 2, \ldots, m-r]$. Since $\|\boldsymbol{Z}\boldsymbol{\delta}_{J_0}\|_2^2 = \|\boldsymbol{\delta}_{J_0}\|_2^2$, it suffices to show that there exists a local maximizer $\boldsymbol{\delta}_0$ of $\overline{Q}_n(\boldsymbol{\delta})$ such that $\|\boldsymbol{\delta}_0\|_2 = O_p(\sqrt{(s+m-r)/n})$.

Define $\mathcal{N}_\tau = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 = \tau\}$, and

$$H_n = \left\{\overline{Q}_n(\boldsymbol{0}) > \max_{\boldsymbol{\delta}\in\partial\mathcal{N}_\tau} \overline{Q}_n(\boldsymbol{\delta})\right\},$$

where $\partial\mathcal{N}_\tau$ denotes the boundary of $\mathcal{N}_\tau$. Clearly, on the event H_n, there exists a local maximizer $\boldsymbol{\delta}$ in $\mathcal{N}_\tau$. Hence, it suffices to show $\Pr(H_n) \to 1$ as $n \to \infty$ and some sufficiently large τ.

For any $\boldsymbol{\delta}$, it follows from a second order Taylor's expansion that

$$\text{(S5.3)}\qquad \overline{Q}_n(\boldsymbol{\delta}) - \overline{Q}_n(\boldsymbol{0}) = \boldsymbol{\delta}^T\boldsymbol{v} - \frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{D}\boldsymbol{\delta},$$

where

$$\boldsymbol{v} = \frac{1}{n}\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}^T_{\mathcal{M}} \\ \boldsymbol{X}^T_S \end{pmatrix}\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\} - \begin{pmatrix} \boldsymbol{0}_{m-r} \\ \lambda_{n,0}\bar{\boldsymbol{\rho}}(\boldsymbol{\beta}^*_{\mathcal{S}}) \end{pmatrix}$$

and

$$\boldsymbol{D} = \frac{1}{n}\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T - \begin{pmatrix} \boldsymbol{O}_{(m-r)\times(m-r)} & \boldsymbol{O}_{(m-r)\times s} \\ \boldsymbol{O}_{s\times(m-r)} & \boldsymbol{\Lambda}^* \end{pmatrix},$$

where $\boldsymbol{\beta}^{**}$ lies in the line segment jointing $\boldsymbol{\beta}(\boldsymbol{\delta})$ and $\boldsymbol{\beta}^*$, and $\boldsymbol{\Lambda}^*$ is a diagonal matrix with nonnegative diagonal elements. By the definition of $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}(\boldsymbol{\delta})$, we have $\boldsymbol{\beta}^{**}_{(\mathcal{M}\cup S)^c} = 0$. Moreover, it follows from (S5.2) and $\|\boldsymbol{\beta}(\boldsymbol{\delta}) - \boldsymbol{\beta}^*\|_2 = \tau\sqrt{(m+s-r)/n}$ that

$$\begin{aligned} \|\boldsymbol{\beta}^{**} - \boldsymbol{\beta}_0\|_2 &\leq \tau\sqrt{(s+m-r)/n} + O(\sqrt{(s+m-r)/n}) \\ &\ll \tau\sqrt{(s+m-r)/n} + \frac{1}{2}\sqrt{(s+m)\log(n)/n}. \end{aligned}$$

Therefore, for sufficiently large n and any $\tau \leq \sqrt{\log n}/2$, we have $\boldsymbol{\beta}^{**} \in \mathcal{N}_0$. By Condition (A2), the maximum eigenvalue of $\boldsymbol{\Lambda}^*$ is upper bounded by $\lambda_n\kappa_0$. Let

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{Z} & \boldsymbol{O}_{m\times s} \\ \boldsymbol{O}_{s\times(m-r)} & \boldsymbol{I}_s \end{pmatrix},$$

for any $\boldsymbol{\delta} \in \mathbb{R}^{m+s-r}$, we have $\|\boldsymbol{L}\boldsymbol{\delta}\|_2 = \|\boldsymbol{\delta}\|_2$. Observe that

$$\frac{1}{n}\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T = \frac{1}{n}\boldsymbol{L}^T\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T\boldsymbol{L}.$$

Hence, for any $\boldsymbol{\delta} \in \mathbb{R}^{m+s-r}$ with $\|\boldsymbol{\delta}\|_2 = 1$, we have

$$\begin{aligned} &\frac{1}{n}\boldsymbol{\delta}^T\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T\boldsymbol{\delta} \\ \geq\ & \|L\boldsymbol{\delta}\|_2^2\lambda_{\min}\left\{\frac{1}{n}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T\right\} \\ =\ & \lambda_{\min}\left\{\frac{1}{n}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}^{**})\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T\right\} \geq c, \end{aligned}$$

where the last inequality is due to Condition (A1). Since $\lambda_n\kappa_0 = o(1)$, we have $\lambda_{\min}(\boldsymbol{D}) \geq \bar{c}$ for some constant $\bar{c}$ and sufficiently large n. Therefore, it follows from (S5.3) that

$$\begin{aligned} &\sup_{\boldsymbol{\delta}\in\mathcal{N}_\tau}\overline{Q}_n(\boldsymbol{\delta}) - \overline{Q}_n(\boldsymbol{0}) \leq \sup_{\boldsymbol{\delta}\in\mathcal{N}_\tau}\left(\|\boldsymbol{\delta}\|_2\|\boldsymbol{v}\|_2 - \frac{\bar{c}}{2}\|\boldsymbol{\delta}\|_2^2\right) \\ =\ & \tau\|\boldsymbol{v}\|_2\sqrt{\frac{m+s-r}{n}} - \frac{\bar{c}\tau^2}{2}\frac{(m+s-r)}{n}. \end{aligned}$$

By Markov's inequality, for any fixed $\tau > 0$ and sufficiently large n, we have

$$\text{(S5.4)}\quad \Pr(H_n) = \Pr\left(\|\boldsymbol{v}\|_2 < \bar{c}\tau\sqrt{\frac{m+s-r}{n}} \right) = 1 - \frac{n\mathrm{E}\|\boldsymbol{v}\|_2^2}{(m+s-r)\bar{c}^2\tau^2}.$$

Assume we can show

$$\text{(S5.5)}\qquad \mathrm{E}\|\boldsymbol{v}\|_2^2 = O\left(\frac{m+s-r}{n}\right).$$

Let $\tau \to \infty$, it follows from (S5.4) that $\Pr(H_n) \to 1$ and the proof is hence completed. It remains to show (S5.5).

By the definition of $\boldsymbol{v}$, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
\mathrm{E}\|\boldsymbol{v}\|_2^2 &\le 2\|\lambda_{n,0}\bar{\boldsymbol{\rho}}(\boldsymbol{\beta}_S^*)\|_2^2 + 2\mathrm{E}\left\| \frac{1}{n} \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\} \right\|_2^2 = 2\|\lambda_{n,0}\bar{\boldsymbol{\rho}}(\boldsymbol{\beta}_S^*)\|_2^2 \\
&+ \frac{2}{n^2}\mathrm{tr}\left[\mathrm{E}\left\{ \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\}\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\}^T \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \right\} \right] \\
&= 2\|\lambda_{n,0}\bar{\boldsymbol{\rho}}(\boldsymbol{\beta}_S^*)\|_2^2 + \frac{2\phi_0}{n^2}\mathrm{tr}\left\{ \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \right\} \\
&+ \frac{2}{n^2}\mathrm{tr}\left\{ \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0) - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\}\{\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0) - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*)\}^T \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \right\} \\
&\triangleq 2I_0 + 2I_1 + 2I_2,
\end{aligned}$$

where tr denotes the trace of a matrix. Observe that $\boldsymbol{\beta}_{\mathcal{S}}^* = \boldsymbol{\beta}_{0,\mathcal{S}}$. It follows from the monotonicity of ρ, the definition of d_n, and Condition (A2) that

$$I_0 \le s\{\lambda_{n,0}\rho'(d_n)\}^2 = o\left(\frac{1}{n}\right).$$

Besides,

$$\begin{aligned}
I_1 &\le \frac{\phi_0(m+s-r)}{n^2}\lambda_{\max}\left\{ \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \right\} \\
&= \frac{\phi_0(m+s-r)}{n^2}\lambda_{\max}\left\{ \boldsymbol{L}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \boldsymbol{L} \right\} \\
&= \frac{\phi_0(m+s-r)}{n^2}\lambda_{\max}\left\{ \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \right\} = O\left(\frac{s+m-r}{n}\right),
\end{aligned}$$

where the last equality is due to Condition (A1).

To prove (S5.5), it remains to show $I_2 = O((m+s-r)/n)$. A first order Taylor expansion gives

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0) - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}^*) &= \int_0^1 \boldsymbol{\Sigma}(\boldsymbol{X}\{\boldsymbol{\beta}^* + t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)\})dt\boldsymbol{X}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*) \\ &= \int_0^1 \boldsymbol{\Sigma}(\boldsymbol{X}\{\boldsymbol{\beta}^* + t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)\})dt\boldsymbol{X}_{\mathcal{M}}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n.\end{aligned}$$

Let $\mathcal{N}_0^* = \{\boldsymbol{\beta}^* + t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*) : \forall 0 \le t \le 1\}$. By (S5.2), we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 = O\left(\frac{\sqrt{s+m-r}}{\sqrt{n}}\right).$$

It follows from Cauchy-Schwarz inequality that

$$\begin{aligned}&\left\|\begin{pmatrix}\boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T\end{pmatrix}\int_0^1 \boldsymbol{\Sigma}(\boldsymbol{X}\{\boldsymbol{\beta}^* + t(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*)\})dt\boldsymbol{X}_{\mathcal{M}}\right\|_2^2 \\ \le\ &\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\lambda_{\max}\left\{\begin{pmatrix}\boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T\end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\begin{pmatrix}\boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T\end{pmatrix}^T\right\} \\ \times\ &\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}}\right) \le \sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\left\{\lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\right)\right\}^2.\end{aligned}$$

By (A1), we have

$$\lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\right) = O(n). \tag{S5.6}$$

Besides, it follows from Taylor's theorem that

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}}\left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right| \\ \le\ &\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}}\sum_{j\in\mathcal{M}\cup S}\left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\mathrm{diag}\{|\boldsymbol{X}^j| \circ |b'''(\boldsymbol{X}\boldsymbol{\beta}^{**})|\}\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right||\beta_{0,j} - \beta_j|,\end{aligned}$$

for some $\boldsymbol{\beta}^*{*}$ lying on the line segment joining $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$. For any $\boldsymbol{\beta} \in \mathcal{N}_0^*$, we have $\boldsymbol{\beta}^{**} \in \mathcal{N}_0^*$. Therefore,

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}}\left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right| \\ \le\ &\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}}\sum_{j\in\mathcal{M}\cup S}\left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\mathrm{diag}\{|\boldsymbol{X}^j| \circ |b'''(\boldsymbol{X}\boldsymbol{\beta})|\}\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right||\beta_{0,j} - \beta_j| \\ \le\ &O(1)n\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*}\sum_j|\beta_{0,j} - \beta_j| = O\left(\sqrt{n}(s+m)\right) = O(n),\end{aligned}$$

where $O(1)$ denotes a positive constant, the second inequality is due to Condition (A1), the first equality follows by the definition of $\mathcal{N}_0^*$, and the last equality is due to the condition that $\max(s, m) = o(n^{1/2})$. Combining this together with (S5.6), we have

$$\begin{aligned}
\text{(S5.7)} \quad & \sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\right) \\
= \; & \sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right| \\
\leq \; & \sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right| \\
+ \; & \sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \left|\boldsymbol{a}^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{a}\right| = O(n),
\end{aligned}$$

and hence

$$\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \left\{\lambda_{\max}\left(\boldsymbol{X}_{\mathcal{M}\cup S}^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}\cup S}\right)\right\}^2 = O(n^2).$$

Using Cauchy-Schwarz inequality again, we obtain

$$\begin{aligned}
I_2 \; &\leq \; \frac{1}{n^2}\sup_{\boldsymbol{\beta}\in\mathcal{N}_0^*} \left\|\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{X}_{\mathcal{M}}\right\|_2^2 \|\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\|_2^2 \\
&= \; O(\|\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\|_2^2) = O(\boldsymbol{h}_n^2) = O((s+m-r)/n),
\end{aligned}$$

by Conditions (A4). This completes the proof for the first step.

Step 2: In this step, we show that with probability tending to 1, the local maximizer $\hat{\boldsymbol{\beta}}$ of $Q_n(\boldsymbol{\beta})$ with the constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$ and $\boldsymbol{\beta}_{(\mathcal{M}\cup S)^c} = 0$ is indeed a local maximizer of $Q_n(\boldsymbol{\beta})$ with the linear constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$. This implies $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}$. From the proof in the first step, we have shown that with probability at least $1 - 1/\tau^2$,

$$\text{(S5.8)} \quad \|\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S} - \boldsymbol{\beta}_{0,\mathcal{M}\cup S}\|_2 \leq \bar{c}\tau\left(\sqrt{\frac{s+m-r}{n}}\right), \; \hat{\boldsymbol{\beta}}_{(\mathcal{M}\cup S)^c} = \boldsymbol{0},$$

for some constant $\bar{c} > 0$ and any $0 < \tau \leq \sqrt{\log n}/2$.

Similar to the proof of Theorem 1 in Fan and Lv (2011), it suffices to show with probability tending to 1, we have

$$\|\boldsymbol{X}_{(\mathcal{M}\cup S)^c}^T\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\}\|_\infty < n\lambda_{n,0}\rho'(0+),$$

or

$$\frac{1}{n}\|\boldsymbol{X}_{(\mathcal{M}\cup S)^T}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\}\|_\infty \ll \lambda_{n,0}. \tag{S5.9}$$

Since $\hat{\boldsymbol{\beta}}_{(\mathcal{M}\cup S)^c} = \boldsymbol{\beta}_{0,(\mathcal{M}\cup S)^c} = \mathbf{0}$, using a second-order Taylor expansion around $\boldsymbol{\beta}_0$, we obtain for any $j \in (\mathcal{M}\cup S)^c$

$$\begin{aligned}(\boldsymbol{X}^j)^T\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\} &= (\boldsymbol{X}^j)^T\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\ - \quad (\boldsymbol{X}^j)^T\boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}&(\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S}) + R_j,\end{aligned} \tag{S5.10}$$

where the absolute value of each element in the remainder term R_j is bounded by

$$\max_j(\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S})^T\boldsymbol{X}_{\mathcal{M}\cup S}^T\mathrm{diag}(|\boldsymbol{X}_j|\circ|b'''(\boldsymbol{X}\bar{\boldsymbol{\beta}}_j)|)\boldsymbol{X}_{\mathcal{M}\cup S}(\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S}),$$

for some $\bar{\boldsymbol{\beta}}_j$ lying on the line segment joining $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Under the event defined in (S5.8), we have $\hat{\boldsymbol{\beta}} \in \mathcal{N}_0$. Since $\boldsymbol{\beta}_0 \in \mathcal{N}_0$, we have $\bar{\boldsymbol{\beta}}_j \in \mathcal{N}_0$. Therefore, under the event defined in (S5.8), it follows from Condition (A1) that

$$\sup_{j\in(\mathcal{M}\cup S)^c}|R_j| \le \bar{c}_1\tau^2(s+m) \ll \bar{c}_1\tau^2\sqrt{n(s+m)}, \tag{S5.11}$$

where $\bar{c}_1$ is some positive constant and the last equality is due to that $\max(s,m) = o(n^{1/2})$.

Besides, by Condition (A1) and (S5.8), there exists some constant $\bar{c}_2 > 0$ such that

$$\begin{aligned}&\|\boldsymbol{X}^T_{(\mathcal{M}\cup S)^c}\Sigma(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}(\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\|_\infty\\ &\quad\le \|\boldsymbol{X}^T_{(\mathcal{M}\cup S)^c}\Sigma(\boldsymbol{X}\boldsymbol{\beta}_0)\boldsymbol{X}_{\mathcal{M}\cup S}\|_{2,\infty}\|\hat{\boldsymbol{\beta}}_{\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S}\|_2 \le \bar{c}_2\tau\sqrt{n(s+m)},\end{aligned} \tag{S5.12}$$

with probability at least $1-1/\tau^2$.

Moreover, it follows from the condition $\max_{1\le j\le p}\|\boldsymbol{X}^j\|_\infty = O(\sqrt{n/\log(p)})$, $\max_{1\le j\le p}\|\boldsymbol{X}^j\|_2 = O(\sqrt{n})$ in (A1), Condition (A3) and Proposition 4 in Fan and Lv (2011) that

$$\begin{aligned}&\max_j \Pr\left(|(\boldsymbol{X}^j)^T\{\boldsymbol{Y}-\mu(\boldsymbol{X}\boldsymbol{\beta}_0)\}| > \gamma\sqrt{n\log p}\right)\\ \le\ & 2\exp\left(-\frac{1}{2}\frac{\gamma^2 n\log p}{O(nv_0)+O(\sqrt{n/\log p})M\gamma\sqrt{n\log p}}\right)\\ \le\ & 2\exp\left(-\frac{1}{2}\frac{\gamma^2\log p}{O(v_0)+O(\gamma M)}\right),\end{aligned}$$

where the constants M and v_0 are defined in Condition (A3). By Bonferroni's inequality, we have

$$\Pr\left(\max_j |X_j^T\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}| > \gamma\sqrt{n\log p}\right) \le 2\exp\left(-\frac{1}{2}\frac{\gamma^2\log p}{O(v_0+\gamma M)}+\log p\right). \tag{S5.13}$$

Since $p\to\infty$, for sufficiently large γ, the RHS of (S5.13) converges to 0. This implies we have with probability tending to 1,

$$\|\boldsymbol{X}^T_{(\mathcal{M}\cup S)^c}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\|_\infty \le \bar{c}_3\sqrt{n\log p}, \tag{S5.14}$$

for some constant $\bar{c}_3>0$.

Let $\bar{c}_4 = \max(\bar{c}_1,\bar{c}_2,\bar{c}_3)$. Combining (S5.11), (S5.12) with (S5.14), we obtain

$$\begin{aligned}\Pr\Big(\|\boldsymbol{X}^T_{(\mathcal{M}\cup S)^c}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\}\|_\infty \le \bar{c}_4\sqrt{n\log p}+\bar{c}_4(\tau+\tau^2)\sqrt{n(s+m)}\Big)\\ \ge 1-\frac{1}{\tau^2}+o(1).\end{aligned} \tag{S5.15}$$

By (A2), we have $\lambda_{n,0}\gg \max(\sqrt{\log p},\sqrt{s+m})/\sqrt{n}$. Let τ_n to be any diverging sequence that $\tau_n \ll \sqrt{\log n}$ and $\tau_n \ll \sqrt{n}\lambda_{n,0}/\max(\sqrt{\log p},\sqrt{s+m})$. By (S5.15), we have

$$\Pr\left(\|\boldsymbol{X}^T_{(\mathcal{M}\cup S)^c}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\}\|_\infty \ll \lambda_{n,0}\right) \ge 1-\frac{1}{\tau_n^2}+o(1)\to 1.$$

Thus, we have with probability tending to 1, (S5.9) holds.

Step 3: We've shown that for any $0<\tau\le\sqrt{\log n}/2$ and sufficiently large n,

$$\Pr\left\{\|\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S}-\boldsymbol{\beta}_{0,\mathcal{M}\cup S}\|_2 \le \bar{c}\tau\left(\frac{s+m-r}{n}\right)\right\} \ge 1-\frac{1}{\tau^2}, \tag{S5.16}$$

$$\Pr\left(\hat{\boldsymbol{\beta}}_0\in\mathcal{N}_0\right)\to 1, \tag{S5.17}$$

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}} = \boldsymbol{t}. \tag{S5.18}$$

Finally, we show that (S5.19) holds.

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix} &= \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}^T_{\mathcal{M}}\\ \boldsymbol{X}^T_S\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\ &- \sqrt{n}\boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n\\ \boldsymbol{0}\end{pmatrix}+o_p(1).\end{aligned} \tag{S5.19}$$

Theorem 2.1 therefore follows.

In the first step, we have shown that $\hat{\boldsymbol{\beta}}_0$ is the local maximizer of $Q_n(\boldsymbol{\beta})$ with the constraints $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$, $\boldsymbol{\beta}_{(\mathcal{M}\cup S)^c} = \mathbf{0}$. This implies that there exists some vector $\boldsymbol{\nu} \in \mathbb{R}^r$ such that

$$\text{(S5.20)} \quad \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup S})\} = \begin{pmatrix} \sqrt{n}\boldsymbol{C}^T\boldsymbol{\nu} \\ n\lambda_n\bar{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{0,S}) \end{pmatrix}.$$

Similar to (S5.10) and (S5.11), we can show that the left-hand side (LHS) of (S5.20) is equal to

$$\begin{aligned} \text{(S5.21)} \qquad & \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\} \\ - \quad & \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} + \boldsymbol{R}, \end{aligned}$$

where the remainder term $\boldsymbol{R}$ satisfies $\|\boldsymbol{R}\|_\infty = O_p(s+m)$. Hence we have

$$\|\boldsymbol{R}\|_2 = O_p\left((s+m)^{3/2}\right) = o_p(\sqrt{n}),$$

under the conditions that $s + m = o(n^{1/3})$. Recall that

$$\boldsymbol{K}_n = \frac{1}{n} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{X}\boldsymbol{\beta}_0) \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}^T.$$

Under Condition (A1), we have $\liminf_n \lambda_{\min}(\boldsymbol{K}_n) > 0$. This implies $\|\boldsymbol{K}_n^{-1}\boldsymbol{R}\|_2 = o_p(\sqrt{n})$. Combining this together with (S5.20) and (S5.21), we obtain

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{0,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} &= \frac{1}{\sqrt{n}} \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\} \\ \text{(S5.22)} \qquad &- \boldsymbol{K}_n^{-1} \begin{pmatrix} \boldsymbol{C}^T\boldsymbol{\nu} \\ \sqrt{n}\lambda_n\bar{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{0,S}) \end{pmatrix} + o_p(1). \end{aligned}$$

Besides, by (S5.16), it follows from Condition (A2) that $\|\hat{\boldsymbol{\beta}}_{0,S} - \boldsymbol{\beta}_{0,S}\|_\infty \ll d_n$ with probability tending to 1. This implies for sufficiently large n, we have $\min_{j\in S}|\hat{\beta}_{0,j}| > \min_{j\in S}|\beta_{0,j}| - d_n = d_n$. By the monotonicity of ρ' and Condition (A2), we obtain

$$|\sqrt{n}\lambda_n\bar{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{0,S})| \le s^{1/2}|\sqrt{n}\lambda_n\rho'(d_n)| = o(1),$$

with probability tending to 1. This together with (S5.22) suggests that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix} = \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\}$$
$$- \boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\nu}+o_p(1). \tag{S5.23}$$

Since $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}}-\boldsymbol{t}=\boldsymbol{h}_n$, by (S5.18), we have $\boldsymbol{C}(\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}})=-\boldsymbol{h}_n$. Therefore, it follows from (S5.23) that

$$\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\nu}-\sqrt{n}\boldsymbol{h}_n \tag{S5.24}$$
$$= \frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\mu(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\}+\boldsymbol{C}\boldsymbol{R}^*,$$

for some m-dimensional vector $\boldsymbol{R}^*$ such that $\|\boldsymbol{R}^*\|_2=o_p(1)$.

By definition, we have

$$\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}=\boldsymbol{C}^T\boldsymbol{\Omega}_{mm}\boldsymbol{C}.$$

Multiplying $(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}$ on both sides of (S5.24), we have

$$\boldsymbol{\nu} = \frac{1}{\sqrt{n}}\boldsymbol{\Psi}^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\}$$
$$+ \boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{R}^*+\sqrt{n}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n,$$

where $\boldsymbol{\Psi}=\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T$. This together with (S5.23) yields

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix} = \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\} \tag{S5.25}$$
$$- \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}^T\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\}$$
$$- \sqrt{n}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n-\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{R}^*+o_p(1)$$
$$= \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}_{\mathcal{M}\cup S}\boldsymbol{\beta}_{0,\mathcal{M}\cup S})\}$$
$$- \sqrt{n}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n-\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{R}^*+o_p(1).$$

In the following, we prove

$$\text{(S5.26)}\quad \boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{R}^*=\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T\boldsymbol{\Psi}^{-1}\boldsymbol{C}\boldsymbol{R}^*=o_p(1).$$

Under Condition (A1), we have $\lambda_{\max}(\boldsymbol{K}_n)=O(1)$. This implies $\liminf_n \lambda_{\min}(\boldsymbol{\Omega}_n)>0$, or equivalently,

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1}\liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_n\boldsymbol{a}>0.$$

Hence, we have

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1,\boldsymbol{a}_{J_0}^c=0}\liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_n\boldsymbol{a}>0,$$

where $J_0=[1,\ldots,m]$. Note that this implies

$$\inf_{\boldsymbol{a}\in\mathbb{R}^{m+s}:\|\boldsymbol{a}\|_2=1}\liminf_n \boldsymbol{a}^T\boldsymbol{\Omega}_{mm}\boldsymbol{a}>0.$$

Therefore, we obtain

$$\text{(S5.27)}\qquad \liminf_n \lambda_{\min}(\boldsymbol{\Omega}_{mm})>0.$$

Similarly, we can show

$$\text{(S5.28)}\qquad \lambda_{\max}(\boldsymbol{\Omega}_{mm})=O(1).$$

Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\|\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}\boldsymbol{C}\boldsymbol{R}^*\|_2^2\\
=\;&\|\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{\Omega}_{mm}^{-1/2}\boldsymbol{R}^*\|_2^2\\
\le\;&\|\boldsymbol{\Omega}_{mm}^{1/2}\|_2^2\|\boldsymbol{\Omega}_{mm}^{1/2}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{\Omega}_{mm}\boldsymbol{C}^T)^{-1}\boldsymbol{C}\boldsymbol{\Omega}_{mm}^{1/2}\|_2^2\|\boldsymbol{\Omega}_{mm}^{-1/2}\|_2^2\|\boldsymbol{R}^*\|_2^2\\
\le\;&\lambda_{\max}\left(\boldsymbol{\Omega}_{mm}\right)\lambda_{\max}\left(\boldsymbol{\Omega}_{mm}^{-1}\right)\|\boldsymbol{R}^*\|_2^2=o_p(1),
\end{aligned}$$

by (S5.27) and (S5.28). This proves (S5.26). Hence, it follows from (S5.25) that

$$\begin{aligned}
\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}}-\boldsymbol{\beta}_{0,\mathcal{M}}\\ \hat{\boldsymbol{\beta}}_{0,S}-\boldsymbol{\beta}_{0,S}\end{pmatrix}&=\frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_n)\boldsymbol{K}_n^{-1/2}\begin{pmatrix}\boldsymbol{X}_{\mathcal{M}}^T\\ \boldsymbol{X}_S^T\end{pmatrix}\{\boldsymbol{Y}-\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\}\\
&\quad-\sqrt{n}\boldsymbol{K}_n^{-1}\begin{pmatrix}\boldsymbol{C}^T\\ \boldsymbol{O}_{r\times s}^T\end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n+o_p(1).\qquad\text{(S5.29)}
\end{aligned}$$

Moreover, observe that

$$\boldsymbol{h}_n = \boldsymbol{C}\boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n = \begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \begin{pmatrix} \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n \\ \boldsymbol{0} \end{pmatrix}.$$

Hence, we have

$$\begin{aligned} \boldsymbol{K}_n^{-1}\begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}\boldsymbol{\Psi}^{-1}\boldsymbol{h}_n &= \boldsymbol{K}_n^{-1}\begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}\boldsymbol{\Psi}^{-1}\begin{pmatrix} \boldsymbol{C}^T \\ \boldsymbol{O}_{r\times s}^T \end{pmatrix}^T \begin{pmatrix} \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n \\ \boldsymbol{0} \end{pmatrix} \\ &= \boldsymbol{K}_n^{-1/2}\boldsymbol{P}_n\boldsymbol{K}_n^{-1/2}\begin{pmatrix} \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{h}_n \\ \boldsymbol{0} \end{pmatrix}. \end{aligned}$$

This together with (S5.29) proves (S5.19).

Similarly, we can show $\hat{\boldsymbol{\beta}}_a$ satisfies (i) $\Pr(\hat{\boldsymbol{\beta}}_{a,(\mathcal{M}\cup S)^c} = 0) \to 1$; (ii) $\|\hat{\boldsymbol{\beta}}_{a,\mathcal{M}\cup S} - \boldsymbol{\beta}_{0,\mathcal{M}\cup S}\|_2 = O_p(\sqrt{(s+m)/n})$; (iii)

$$\begin{aligned} \sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} &= \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + O\left(\frac{(s+m)^{3/2}}{\sqrt{n}}\right) \\ &+ O_p\left((ns)^{1/2}\lambda_{n,a}\rho'(d_n,\lambda_{n,a})\right). \end{aligned}$$

Under the condition that $\max(s,m) = o(n^{1/3})$ and $\lambda_{n,a}\rho'(d_n,\lambda_{n,a}) = o(n^{-1/2}(s+m)^{-1/2})$ in (A2), we obtain

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{\beta}_{0,\mathcal{M}} \\ \hat{\boldsymbol{\beta}}_{a,S} - \boldsymbol{\beta}_{0,S} \end{pmatrix} = \frac{1}{\sqrt{n}}\boldsymbol{K}_n^{-1}\begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_S^T \end{pmatrix}\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}_0)\} + o_p(1).$$

The proof is hence completed.

S5.2. Proof of Lemma S.1. Let $\boldsymbol{\Theta} = \boldsymbol{\Phi}^{-1}$. It follows from the block matrix inversion formula (Lemma S.9) that

$$\begin{aligned} &\boldsymbol{\Theta}^{-1}_{\{1\}\cup S,\{1\}\cup S} \\ =\ &\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S} - \boldsymbol{\Phi}_{\{1\}\cup S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,\{1\}\cup S}. \end{aligned}$$

Observe that $\boldsymbol{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{1,p} = \boldsymbol{e}_{1,p}^T\boldsymbol{\Theta}\boldsymbol{e}_{1,p} = \boldsymbol{e}_{1,1+s}^T\boldsymbol{\Theta}_{1\cup\{S\},1\cup\{S\}}\boldsymbol{e}_{1,1+s}$. Hence,

$$\begin{aligned} &\boldsymbol{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{1,p} = \boldsymbol{e}_{1,1+s}^T \\ &\left\{\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S} - \boldsymbol{\Phi}_{\{1\}\cup S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,\{1\}\cup S}\right\}^{-1}\boldsymbol{e}_{1,1+s}. \end{aligned}$$

With some calculation, we have

$$\begin{aligned}
&\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S} - \boldsymbol{\Phi}_{\{1\}\cup S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,\{1\}\cup S} \\
= &\begin{pmatrix} \boldsymbol{\Phi}_{\{1\},\{1\}} & \boldsymbol{\Phi}_{S,\{1\}}^T \\ \boldsymbol{\Phi}_{S,\{1\}} & \boldsymbol{\Phi}_{S,S} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Phi}_{\{1\},(\{1\}\cup S)^c} \\ \boldsymbol{\Phi}_{S,(\{1\}\cup S)^c} \end{pmatrix} \boldsymbol{\Phi}^{-1}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c} \begin{pmatrix} \boldsymbol{\Phi}_{\{1\},(\{1\}\cup S)^c} \\ \boldsymbol{\Phi}_{S,(\{1\}\cup S)^c} \end{pmatrix}^T \\
= &\begin{pmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{21}^T \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{K}_{11} &= \boldsymbol{\Phi}_{\{1\},\{1\}} - \boldsymbol{\Phi}_{\{1\},(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}^T_{\{1\},(\{1\}\cup S)^c}, \\
\boldsymbol{K}_{21} &= \boldsymbol{\Phi}_{S,\{1\}} - \boldsymbol{\Phi}_{S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,\{1\}}, \\
\boldsymbol{K}_{22} &= \boldsymbol{\Phi}_{S,S} - \boldsymbol{\Phi}_{S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,S}.
\end{aligned}$$

By Lemma S.9, we have

$$\boldsymbol{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{1,p} = (\boldsymbol{K}_{11} - \boldsymbol{K}_{21}^T\boldsymbol{K}_{22}^{-1}\boldsymbol{K}_{21})^{-1}.$$

Similarly we have

$$\boldsymbol{e}_{1,1+s}^T(\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S})^{-1}\boldsymbol{e}_{1,1+s} = \boldsymbol{K}_{11}^{-1}.$$

Therefore, to show $\boldsymbol{e}_{1,p}^T\boldsymbol{\Phi}^{-1}\boldsymbol{e}_{1,p} \geq \boldsymbol{e}_{1,1+s}^T(\boldsymbol{\Phi}_{\{1\}\cup S,\{1\}\cup S})^{-1}\boldsymbol{e}_{1,1+s}$, it suffices to show $\boldsymbol{K}_{11} \geq \boldsymbol{K}_{11} - \boldsymbol{K}_{21}^T\boldsymbol{K}_{22}^{-1}\boldsymbol{K}_{21}$. However, this is immediate to see since $\boldsymbol{K}_{22}$ is positive definite. Besides, the equality holds if and only if $\boldsymbol{K}_{21} = 0$, or equivalently

$$\boldsymbol{\Phi}_{S,\{1\}} = \boldsymbol{\Phi}_{S,(\{1\}\cup S)^c}(\boldsymbol{\Phi}_{(\{1\}\cup S)^c,(\{1\}\cup S)^c})^{-1}\boldsymbol{\Phi}_{(\{1\}\cup S)^c,\{1\}}.$$

This completes the proof.

S5.3. Proof of Lemma S.2. To prove Lemma S.2, we need to show

$$\text{(S5.30)} \quad \Pr\left(\max_{1\leq j\leq p}\|\boldsymbol{X}^j\|_\infty = O\left(\sqrt{n/\log(p)}\right)\right) \to 1,$$

$$\text{(S5.31)} \quad \Pr\left(\max_{1\leq j\leq p}\|\boldsymbol{X}^j\|_2 = O(\sqrt{n})\right) \to 1,$$

$$\text{(S5.32)} \quad \Pr\left(\lambda_{\min}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right) \geq cn\right) \to 1,$$

$$\text{(S5.33)} \quad \Pr\left(\lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O(n)\right) \to 1,$$

$$\text{(S5.34)} \quad \Pr\left(\|\boldsymbol{X}_{(S\cup\mathcal{M})^c}^T\boldsymbol{X}_{S\cup\mathcal{M}}\|_{2,\infty} = O(n)\right) \to 1,$$

$$\text{(S5.35)} \quad \Pr\left(\max_{1\leq j\leq p}\lambda_{\max}\left\{\boldsymbol{X}_{S\cup\mathcal{M}}^T\text{diag}(|\boldsymbol{X}^j|)\boldsymbol{X}_{S\cup\mathcal{M}}\right\} = O(n)\right) \to 1,$$

for some constant $c > 0$.

Under the given conditions, we have

$$\max_{1\le j\le p} \|\boldsymbol{X}^j\|_\infty \le \omega_0 \quad \text{and} \quad \max_{1\le j\le p} \|\boldsymbol{X}^j\|_2 \le \omega_0\sqrt{n}.$$

Since $\log p = O(n^a)$ for some $0 < a < 1$, we have $\log p = o(n)$. This proves (S5.30) and (S5.31).

Note that

$$\begin{aligned}
\frac{1}{n}\lambda_{\min}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right) &= \frac{1}{n}\inf_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^T\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\boldsymbol{a} \ge \inf_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^T\boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\boldsymbol{a} \\
&\quad - \left|\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}} \boldsymbol{a}^T\left(\frac{1}{n}\boldsymbol{X}_{i,S\cup\mathcal{M}}\boldsymbol{X}_{i,S\cup\mathcal{M}}^T - \boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\right)\boldsymbol{a}\right| \\
&\ge \bar{c} - \left\|\frac{1}{n}\boldsymbol{X}_{i,S\cup\mathcal{M}}\boldsymbol{X}_{i,S\cup\mathcal{M}}^T - \boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\right\|_2 \\
&\ge \bar{c} - \left\|\frac{1}{n}\boldsymbol{X}_{i,S\cup\mathcal{M}}\boldsymbol{X}_{i,S\cup\mathcal{M}}^T - \boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\right\|_\infty, \qquad \text{(S5.36)}
\end{aligned}$$

where the third inequality follows by Lemma S.5 and the condition that $\lambda_{\min}(\boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}) \ge \bar{c}$, the last inequality is due to Lemma S.8.

For any $j_1, j_2 \in S\cup\mathcal{M}$, we have

$$\begin{aligned}
&\mathrm{E}\exp(|X_{i,j_1}X_{i,j_2} - \boldsymbol{\Lambda}_{j_1,j_2}|) - 1 - |X_{i,j_1}X_{i,j_2} - \boldsymbol{\Lambda}_{j_1,j_2}| \\
\le\ &\mathrm{E}\exp(|X_{i,j_1}X_{i,j_2} - \boldsymbol{\Lambda}_{j_1,j_2}|) \le \mathrm{E}\exp(2|X_{i,j_1}X_{i,j_2}|) \le \exp(2\omega_0^2),
\end{aligned}$$

where the second inequality is due to Jensen's inequality. Similar to (S5.13), we can show

$$\begin{aligned}
&\Pr\left(\max_{j_1,j_2\in S\cup\mathcal{M}}\left|\frac{1}{n}\sum_{i=1}^n X_{i,j_1}X_{i,j_2} - \boldsymbol{\Lambda}_{j_1,j_2}\right| > \gamma\frac{\sqrt{\log n}}{\sqrt{n}}\right) \\
\le\ &2\exp\left(-\frac{1}{2}\frac{\gamma^2\log n}{\exp(2\omega_0^2) + \gamma\sqrt{n\log p}} + 2\log(s+m)\right).
\end{aligned}$$

Since $s + m = o(n)$, we can show for sufficiently large γ that

$$\Pr\left(\max_{j_1,j_2\in S\cup\mathcal{M}}\left|\frac{1}{n}\sum_{i=1}^n X_{i,j_1}X_{i,j_2} - \boldsymbol{\Lambda}_{j_1,j_2}\right| > \gamma\frac{\sqrt{\log n}}{\sqrt{n}}\right) \to 0. \qquad \text{(S5.37)}$$

This further implies

$$\Pr\left(\max_{j_1,j_2\in S\cup\mathcal{M}}\left\|\frac{1}{n}\sum_{i=1}^{n}X_{i,j_1}X_{i,j_2}-\mathbf{\Lambda}_{j_1,j_2}\right\|_{\infty}>\gamma(s+m)\frac{\sqrt{\log n}}{\sqrt{n}}\right)\to 0.$$

Under the given conditions, we have $(s+m)\sqrt{\log n}=o(\sqrt{n})$. Hence, we have for sufficiently large n,

$$\Pr\left(\max_{j_1,j_2\in S\cup\mathcal{M}}\left\|\frac{1}{n}\sum_{i=1}^{n}X_{i,j_1}X_{i,j_2}-\mathbf{\Lambda}_{j_1,j_2}\right\|_{\infty}>\frac{\bar{c}}{2}\right)\to 0.$$

This together with (S5.36) implies that

$$\Pr\left\{\frac{1}{n}\lambda_{\min}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^{T}\boldsymbol{X}_{S\cup\mathcal{M}}\right)>\frac{\bar{c}}{2}\right\}\to 1. \tag{S5.38}$$

(S5.32) is hence proven. Similarly, we can show (S5.33) holds.

By condition, we have $\lambda_{\max}(\mathbf{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}})=O(1)$. It then follows from Cauchy-Schwarz inequality that

$$\begin{aligned}&\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2\le 1}}\left|\mathrm{E}\boldsymbol{X}_{0,j}\boldsymbol{X}_{0,S\cup\mathcal{M}}^{T}\boldsymbol{a}\right|\le\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2\le 1}}\sqrt{\mathrm{E}\boldsymbol{X}_{0,j}^{2}}\sqrt{\mathrm{E}(\boldsymbol{X}_{0,S\cup\mathcal{M}}^{T}\boldsymbol{a})^{2}}\\ \le\ &\omega_0^2\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2\le 1}}\boldsymbol{a}^{T}\mathbf{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\boldsymbol{a}=O(\omega_0^2).\end{aligned}$$

This implies

$$\left\|\mathrm{E}\boldsymbol{X}_{0,(S\cup\mathcal{M})^c}\boldsymbol{X}_{0,S\cup\mathcal{M}}^{T}\right\|_{2,\infty}=O(1). \tag{S5.39}$$

Similar to (S5.37), we can show

$$\Pr\left(\max_{1\le j_1\le p,j_2\in S\cup\mathcal{M}}\left|\sum_{i=1}^{n}X_{i,j_1}X_{i,j_2}-n\mathrm{E}X_{0,j_1}X_{0,j_2}\right|>\gamma\sqrt{n\log p}\right)\to 0,$$

for some constant $\gamma>0$. This further implies that

$$\Pr\left(\left\|\sum_{i=1}^{n}\boldsymbol{X}_{i,(S\cup\mathcal{M})^c}\boldsymbol{X}_{i,S\cup\mathcal{M}}-n\boldsymbol{X}_{0,(S\cup\mathcal{M})^c}\boldsymbol{X}_{0,S\cup\mathcal{M}}\right\|_{2,\infty}>\gamma\sqrt{(s+m)n\log p}\right)\to 0. \tag{S5.40}$$

Under the given conditions, we have $\sqrt{(s+m)n\log p} = O(n)$. This together with (S5.39) implies that

$$\Pr\left(\left\|\sum_{i=1}^{n} \boldsymbol{X}_{i,(S\cup\mathcal{M})^c}\boldsymbol{X}_{i,S\cup\mathcal{M}}\right\|_{2,\infty} = O(n)\right) \to 1.$$

(S5.34) is hence proven.

Finally, note that

$$\max_{1\le j\le p} \lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T \mathrm{diag}\{|\boldsymbol{X}^j|\}\boldsymbol{X}_{S\cup\mathcal{M}}\right) \le \omega_0 \lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T\boldsymbol{X}_{S\cup\mathcal{M}}\right).$$

By (S5.33), we have

$$\Pr\left(\max_{1\le j\le p} \lambda_{\max}\left(\boldsymbol{X}_{S\cup\mathcal{M}}^T \mathrm{diag}\{|\boldsymbol{X}^j|\}\boldsymbol{X}_{S\cup\mathcal{M}}\right) = O(n)\right) \to 1.$$

This proves (S5.35). The proof is hence completed.

S5.4. Proof of Lemma S.3. It suffices to show (S5.30)-(S5.35) hold. By Bonferroni's inequality, Markov's inequality and the definition of the Orlicz norm, we have

$$\begin{aligned}
& \Pr\left(\max_{1\le i\le n, 1\le j\le p} |X_{i,j}| > \sqrt{2}\omega_0\sqrt{\log p + \log n}\right) \\
\le\ & np \max_{1\le i\le n, 1\le j\le p} \Pr(|X_{i,j}| > \sqrt{2}\omega_0\sqrt{\log p + \log n}) \\
\le\ & np\mathrm{E}\frac{\exp(X_{i,j}|^2/\omega_0^2)}{\exp\left(2\omega_0^2(\log p + \log n)/\omega_0^2\right)} \le \frac{2np}{\exp\left(\log(np)^2\right)} = \frac{2}{np} \to 0.
\end{aligned}$$

Therefore,

$$\Pr\left(\max_{1\le i\le n, 1\le j\le p} |X_{i,j}| = O(\sqrt{\log p + \log n})\right) \to 1. \tag{S5.41}$$

Under the given conditions, we have $\log p + \log n = O(n/(\log p))$. This together with (S5.41) yields (S5.30).

By the definition of the Orlicz norm, we have

$$\max_{1\le j\le p} \mathrm{E}X_{0,j}^2 \le \max_{1\le j\le p} \|X_{0,j}^2\|_{\psi_1} = \max_{1\le j\le p} \|X_{0,j}\|_{\psi_2}^2 \le \omega_0^2. \tag{S5.42}$$

Besides, it follows from the Bernstein's inequality (see Lemma G.3 in Shi et al., 2017) and Bonferroni's inequality that

$$\begin{aligned}&\Pr\left(\max_{j=1}^{p}\left|\sum_{i=1}^{n}X_{i,j}^2-n\mathrm{E}X_{0,j}^2\right|\le\gamma\sqrt{n\log p}\right)\\ \le\ &\max_{j=1}^{p}p\Pr\left(\left|\sum_{i=1}^{n}X_{i,j}^2-n\mathrm{E}X_{0,j}^2\right|\le\gamma\sqrt{n\log p}\right)\to 1,\end{aligned}$$

for some constant $\gamma>0$. Under the given conditions, this implies that

$$\Pr\left(\max_{j=1}^{p}\left|\sum_{i=1}^{n}X_{i,j}^2-n\mathrm{E}X_{0,j}^2\right|\le\gamma n\right)\to 1.$$

Combining this together with (S5.42) yields (S5.31).

It follows from Lemma C.1 in Shi et al. (2017) that we have with probability tending to 1,

$$\text{(S5.43)}\quad\sup_{\substack{\boldsymbol{a}\in\mathbb{R}^{s+m}\\ \|\boldsymbol{a}\|_2=1}}\left|\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{a}^T\boldsymbol{X}_{i,S\cup\mathcal{M}})^2-\boldsymbol{a}^T\boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\boldsymbol{a}\right|=O\left(\frac{\sqrt{s\log n}}{\sqrt{n}}\right).$$

Note that the RHS of (S5.43) is $o(1)$. By condition, $\lambda_{\min}(\boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}})\ge\bar{c}$. Following the arguments in (S5.36), we have

$$\Pr\left(\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i,S\cup\mathcal{M}}\boldsymbol{X}_{i,S\cup\mathcal{M}}^T\right)\ge\frac{\bar{c}}{2}\right)\to 1.$$

This proves (S5.32). Similarly, we can show (S5.33) holds.

Similar to (S5.42), we can show

$$\sup_{\boldsymbol{a}\in\mathbb{R}^{s+m},\|\boldsymbol{a}\|_2\le 1}\mathrm{E}|\boldsymbol{a}^T\boldsymbol{X}_{0,S\cup\mathcal{M}}|_2^2\le\sup_{\boldsymbol{a}\in\mathbb{R}^{s+m},\|\boldsymbol{a}\|_2\le 1}\|\boldsymbol{a}^T\boldsymbol{X}_{0,S\cup\mathcal{M}}\|_{\psi_2}^2\le\omega_0^2.$$

Hence, we have

$$\lambda_{\max}\left(\boldsymbol{\Lambda}_{S\cup\mathcal{M},S\cup\mathcal{M}}\right)=O(1).$$

Using similar arguments in (S5.39) and (S5.40), we can show (S5.34) holds.

Finally, by (S5.42), we have

$$\max_{1\le j\le p}\|X_{0,j}\|_{\psi_1}\le\max_{1\le j\le p}\|X_{0,j}\|_{\psi_2}/\sqrt{\log 2}\le\omega_0/\sqrt{\log 2}.$$

As a result, (S5.35) is directly implies by Lemma C.2 in Shi et al. (2017). This completes the proof.

APPENDIX S6: TECHNICAL LEMMAS

LEMMA S.4. *Let* $\widehat{S}_0 = \{j \in \mathcal{M}^c : \hat{\beta}_{0,j} \neq 0\}$. *Suppose that* $\hat{\boldsymbol{\beta}}_0$ *satisfies the following conditions. There exists some vector* $\boldsymbol{\nu} \in \mathbb{R}^r$ *such that*

$$\boldsymbol{X}_{\mathcal{M}}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}_{\widehat{S}\cup\mathcal{M}}\hat{\boldsymbol{\beta}}_{0,\widehat{S}\cup\mathcal{M}})) = \boldsymbol{C}^T\boldsymbol{\nu}, \tag{S6.1}$$

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}} = \boldsymbol{t}, \tag{S6.2}$$

$$\boldsymbol{X}_{\widehat{S}}^T(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}_{\widehat{S}\cup\mathcal{M}}\hat{\boldsymbol{\beta}}_{0,\widehat{S}\cup\mathcal{M}})) = n\bar{\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{0,\widehat{S}}, \lambda_{n,0}), \tag{S6.3}$$

$$\left\|\boldsymbol{X}_{(\widehat{S}\cup\mathcal{M})^c}^T\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\}\right\|_\infty < \lambda_{n,0}\rho'(0+). \tag{S6.4}$$

Furthermore, for any basis matrix $\boldsymbol{Z} \in \mathbb{R}^{m\times(m-r)}$ *for the null space of* $\boldsymbol{C}$,

$$\lambda_{\min}\left\{\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_{\widehat{S}}^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_{\widehat{S}}^T \end{pmatrix}^T\right\} > n\lambda_{n,0}\kappa(\rho, \hat{\boldsymbol{\beta}}_{0,\widehat{S}}, \lambda_{n,0}). \tag{S6.5}$$

Then $\hat{\boldsymbol{\beta}}_0$ *is a local maximizer of* $Q_n(\boldsymbol{\beta}, \lambda_{n,0})$ *with the constraint* $\boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{t}$.

REMARK S6.1. Note that solving (2.2) is equivalent to maximizing the following unconstrained problem:

$$\arg\max_{\boldsymbol{\beta},\boldsymbol{\nu}} = \left(\frac{1}{n}\sum_{i=1}^n\{Y_i\boldsymbol{\beta}^T\boldsymbol{X}_i - b(\boldsymbol{\beta}^T\boldsymbol{X}_i)\} - \sum_{j\notin\mathcal{M}} p_{\lambda_{n,0}}(|\beta_j|) - \boldsymbol{\nu}^T(\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t})\right).$$

Conditions (S6.1), (S6.2), (S6.3) and (S6.5) in Lemma S.4 guarantee that $\hat{\boldsymbol{\beta}}_0$ is a local optimum to (2.2) when constrained on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{(\widehat{S}\cup\mathcal{M})^c} = 0\}$ of $\mathbb{R}^p$. Condition (S6.4) ensures that $\hat{\boldsymbol{\beta}}_0$ is indeed a local maximizer of (2.2) with the linear constraints.

Proof: We first show (S6.1), (S6.2), (S6.3) and (S6.5) guarantee that $\hat{\boldsymbol{\beta}}_0$ is a local maximizer of $Q_n(\boldsymbol{\beta}, \lambda_{n,0})$ subject to $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$ when constrained on the subspace $\{\boldsymbol{\beta} : \boldsymbol{\beta}_{(\widehat{S}\cup\mathcal{M})^c} = 0\}$ of $\mathbb{R}^p$. For simplicity, we assume ρ is twice differentiable at $|\hat{\beta}_{0,j}|$ for any $j \in \mathcal{S}$. Since $\boldsymbol{Z}$ is a basis matrix for the null space of $\boldsymbol{C}$, the matrix

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{Z} & \boldsymbol{O}_{m\times s} \\ \boldsymbol{O}_{s\times m-r} & \boldsymbol{I}_s \end{pmatrix}$$

is a basis matrix for the null space of $(\boldsymbol{C}\ \boldsymbol{O}_{r\times s})$. The local optimality of

$\hat{\boldsymbol{\beta}}_{0,\mathcal{M}\cup\widehat{S}}$ requires:

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_{0,\mathcal{M}} = \boldsymbol{t}, \tag{S6.6}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_0, \lambda_{n,0})}{\partial \boldsymbol{\beta}_{\widehat{S}}} = \boldsymbol{0}, \tag{S6.7}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\beta}}_0, \lambda_{n,0})}{\partial \boldsymbol{\beta}_{\mathcal{M}}} = \boldsymbol{C}^T\boldsymbol{\nu}, \tag{S6.8}$$

$$\lambda_{\min}\left\{\boldsymbol{L}^T\begin{pmatrix} -\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_0,\lambda_{n,0})}{\partial\boldsymbol{\beta}_{\mathcal{M}}\partial\boldsymbol{\beta}_{\mathcal{M}}^T} & -\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_0,\lambda_{n,0})}{\partial\boldsymbol{\beta}_{\mathcal{M}}\partial\boldsymbol{\beta}_{\widehat{S}}^T} \\ -\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_0,\lambda_{n,0})}{\partial\boldsymbol{\beta}_{\widehat{S}}\partial\boldsymbol{\beta}_{\mathcal{M}}^T} & -\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}}_0,\lambda_{n,0})}{\partial\boldsymbol{\beta}_{\widehat{S}}\partial\boldsymbol{\beta}_{\widehat{S}}^T} \end{pmatrix}\boldsymbol{L}\right\} > 0. \tag{S6.9}$$

It is immediate to see that, (S6.1), (S6.2) and (S6.3) directly imply (S6.6), (S6.7) and (S6.8). In the following, we show that (S6.5) implies (S6.9). When ρ is twice differentiable at $|\hat{\beta}_{0,j}|$ for any $j \in \mathcal{S}$, the matrix in the left-hand side (LHS) of (S6.9) is equal to

$$\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_{\widehat{S}}^T \end{pmatrix}\boldsymbol{\Sigma}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)\begin{pmatrix} \boldsymbol{Z}^T\boldsymbol{X}_{\mathcal{M}}^T \\ \boldsymbol{X}_{\widehat{S}}^T \end{pmatrix}^T$$
$$- \quad n\lambda_{n,0}\mathrm{diag}\left(\underbrace{0,\cdots,0}_{\mathrm{m}}, -\rho''(|\hat{\boldsymbol{\beta}}_{0,\widehat{S}_1}|),\cdots,-\rho''(|\hat{\boldsymbol{\beta}}_{0,\widehat{S}_s}|)\right) \triangleq \boldsymbol{M}_1 - \boldsymbol{M}_2,$$

where $\widehat{S}_j$ denotes the jth element in the set $\widehat{S}$. When ρ is not twice differentiable, we can replace M_2 by a diagonal matrix whose absolute value is bounded by $\kappa(\rho, \hat{\boldsymbol{\beta}}_{0,\widehat{S}}, \lambda_{n,0})$. Hence, LHS of (S6.9) is larger than $\lambda_{\min}(\boldsymbol{M}_1) - \lambda_{\max}(\boldsymbol{M}_2) \geq \lambda_{\min}(\boldsymbol{M}_1) - n\lambda_n\kappa(\rho, \hat{\boldsymbol{\beta}}_{0,\widehat{S}}, \lambda_{n,0})$. Therefore, it follows from (S6.5) that (S6.9) is satisfied.

Under Condition (S6.4), using similar arguments in the proof of Theorem 1 in Lv and Fan (2009) or the proof of Theorem 1 in Fan and Lv (2011), we can show $\hat{\boldsymbol{\beta}}_0$ is indeed a local maximizer of $Q_n(\boldsymbol{\beta}, \lambda_{n,0})$ with $\boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}$. This completes the proof.

LEMMA S.5. *For any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{q\times q}$, we have*

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}|.$$

Proof: By Cauchy-Schwarz inequality, we have

$$\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}| \leq \sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} \|\boldsymbol{a}\|_2^2\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_2.$$

Hence, we have shown $\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}| \leq \|\boldsymbol{A}\|_2$. It remains to show $\|\boldsymbol{A}\|_2 \leq \sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}|$. Since $\boldsymbol{A}$ is symmetric, according to the eigen decomposition theorem, we have $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ for some orthogonal matrix $\boldsymbol{U}$ and diagonal matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_q)$. By definition, we have

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T)} = \sqrt{\lambda_{\max}(\boldsymbol{\Lambda}^2)} = \max_{j=1}^{q} |\lambda_j|. \tag{S6.10}$$

Assume $\boldsymbol{U} = (u_1, \ldots, u_q)$. Since $\boldsymbol{U}$ is orthogonal, we have $u_i^T u_j = 0$ when $i \neq j$ and $u_j^T u_j = 1$. Hence, we have

$$\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}| \geq \max_j |u_j^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T u_j| = \max_j |\lambda_j|.$$

This together with (S6.10) implies that $\sup_{\boldsymbol{a}:\|\boldsymbol{a}\|_2=1} |\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}| \geq \|\boldsymbol{A}\|_2$. The proof is hence completed.

LEMMA S.6. *Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_n$ be independent m-dimensional random vectors with $E\boldsymbol{X}_j = 0$, $\sum_j cov(\boldsymbol{X}_j) = \boldsymbol{I}_m$, let $\boldsymbol{Z}$ denote an m-dimensional multivariate normal vector with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{I}_m$, then*

$$\sup_C \left| Pr\left(\sum \boldsymbol{X}_i \in C\right) - Pr(\boldsymbol{Z} \in C) \right| \leq c_0 m^{1/4} \sum E\|\boldsymbol{X}_i\|_2^3,$$

for some constant c_0, where the supremum is taken for all convex subsets in $\mathbb{R}^m$.

Proof: This follows directly from Theorem 1 in Bentkus (2004).

LEMMA S.7. *Denoted by $\chi^2(r, \gamma)$ a χ^2 random variable with r degrees of freedom and a non-centrality parameter γ. Then we have*

$$\lim_{\varepsilon \to 0^+} \sup_{r \geq 1, \gamma \geq 0} |Pr(\chi^2(r, \gamma) \leq x + r\varepsilon) - Pr(\chi^2(r, \gamma) \leq x - r\varepsilon)| \to 0.$$

Proof: It suffices to show

$$\lim_{\varepsilon \to 0^+} \sup_{r \geq 1} |\Pr(\chi^2(r, 0) \leq x + r\varepsilon) - \Pr(\chi^2(r, 0) \leq x - r\varepsilon)| \to 0, \tag{S6.11}$$

since

$$\begin{aligned}
& |\Pr(\chi^2(r,\gamma) \le x + r\varepsilon) - \Pr(\chi^2(r,\gamma) \le x - r\varepsilon)| \\
\le & \left| \sum_{k=0}^{\infty} \exp(-\gamma/2) \frac{(\gamma/2)^k}{k!} \{\Pr(\chi^2(r+2k,0) \le x + r\varepsilon) - \Pr(\chi^2(r+2k,0) \le x - r\varepsilon)\} \right| \\
\le & \sum_{k=0}^{\infty} \exp(-\gamma/2) \frac{(\gamma/2)^k}{k!} \left| \Pr(\chi^2(r+2k,0) \le x + r\varepsilon) - \Pr(\chi^2(r+2k,0) \le x - r\varepsilon) \right| \\
\le & \sup_r \left| \Pr(\chi^2(r,0) \le x + r\varepsilon) - \Pr(\chi^2(r,0) \le x - r\varepsilon) \right|.
\end{aligned}$$

Below, we show (S6.11) holds. We first prove

$$\text{(S6.12)} \quad \lim_{\varepsilon \to 0^+} \sup_{r \ge 2, x} |\Pr(\chi^2(r,0) \le x + r\varepsilon) - \Pr(\chi^2(r,0) \le x - r\varepsilon)| \to 0.$$

Note that (S6.12) holds when the probability density function f_r of a χ^2 distribution with r degrees of freedom satisfies $\sup_{r \ge 2} \sup_x r f_r(x) = O(1)$. By definition, we have

$$f_r(x) = \frac{1}{2^{r/2}\Gamma(r/2)} x^{r/2-1} \exp(-x/2).$$

The supremum of f_r is achieved at $x = r - 2$. For $r = 2$, obviously we have $\sup_x 2|f_2(x)| \le 2|f_2(0)| = 1$. For $r \ge 3$, we have

$$\text{(S6.13)} \quad \sup_{x, r \ge 3} |f_r(x)| \le \sup_{r \ge 3} f_r(r-2) \le \sup_{r \ge 3} \frac{(r/2-1)^{r/2-1}}{\Gamma(r/2)} \exp(-r/2+1).$$

By Stirling's formula, we have

$$\Gamma(r/2) \ge \sqrt{2\pi} \left(\frac{r}{2} - 1\right)^{r/2-1/2} \exp\left(-\frac{r}{2} + 1\right),$$

which together with (S6.13) implies that

$$\sup_x \sup_{r \ge 3} r|f_r(x)| \le \frac{r}{\sqrt{2\pi}(r/2-1)} \le \frac{3\sqrt{2}}{\sqrt{\pi}}.$$

This proves (S6.12). It remains to show

$$\text{(S6.14)} \quad \lim_{\varepsilon \to 0^+} \sup_x |\Pr(\chi^2(1,0) \le x + \varepsilon) - \Pr(\chi^2(1,0) \le x - \varepsilon)| \to 0.$$

For any $\epsilon > 0$, there exists some $\delta > 0$ such that $\Pr(\chi^2(1,0) \leq 2\delta) \leq \epsilon$. Hence, for any $x \leq \delta$ and any $\varepsilon \leq \delta$, we have

$$\begin{aligned} & \Pr(\chi^2(1,0) \leq x+\varepsilon) - \Pr(\chi^2(1,0) \leq x-\varepsilon) \\ \leq\ & \Pr(\chi^2(1,0) \leq 2\delta) \leq \epsilon. \end{aligned} \tag{S6.15}$$

Moreover, $f_1(x)$ is continuous and monotonically decreasing on $[\delta, +\infty)$. This implies $\sup_{x \geq \delta} f_1(x) \leq f_1(\delta)$ and hence

$$\sup_{x \geq \delta} |\Pr(\chi^2(1,0) \leq x+\varepsilon) - \Pr(\chi^2(1,0) \leq x-\varepsilon)| \leq \frac{2\varepsilon}{f_1(\delta)}.$$

Therefore, for any $x \geq \delta$ and any $0 < \varepsilon \leq (2\varepsilon)/\{f_1(\delta)\}$, we have

$$\begin{aligned} & \Pr(\chi^2(1,0) \leq x+\varepsilon) - \Pr(\chi^2(1,0) \\ \leq\ & x-\varepsilon) \leq \Pr(\chi^2(1,0) \leq 2\delta) \leq \epsilon. \end{aligned} \tag{S6.16}$$

Equation (S6.14) now follows from (S6.15) and (S6.16). This completes the proof.

LEMMA S.8. *For any symmetric matrix $\boldsymbol{A}$, we have $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_\infty$.*

Proof : Note that $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_\infty \|\boldsymbol{A}\|_1$. Since $\boldsymbol{A}$ is symmetric, we have $\|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_1$. This proves the assertion.

LEMMA S.9 (Matrix inversion in block form). *For any positive definite matrix*

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix},$$

denote its inverse matrix as $\boldsymbol{\Omega}$ and partition it into $\boldsymbol{\Omega}_{11}, \ldots, \boldsymbol{\Omega}_{22}$ accordingly. Then

$$\boldsymbol{\Omega}_{11} = (\boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21})^{-1}, \qquad \boldsymbol{\Omega}_{22} = (\boldsymbol{\Psi}_{22} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12})^{-1}.$$

As a result, the matrices $\boldsymbol{\Omega}_{11} - \boldsymbol{\Psi}_{11}^{-1}$ and $\boldsymbol{\Omega}_{22} - \boldsymbol{\Psi}_{22}^{-1}$ are positive semidefinite.

APPENDIX S7: REAL DATA ANALYSIS

In this section, we apply our proposed testing procedures to the European American single nucleotide polymorphisms (SNPs) data set (Price et al., 2006), which consists of 488 European American samples. We use the height phenotype (0/1, binary variable) of these European American samples as

the response, and focus on finding variables that are associated with this phenotype among a set of 277 SNPs. The genotype for each SNP is a categorical variable, coded as 0/1/2. We removed the outlier individuals as in Price et al. (2006). This gives us a total of 361 observations. However, for each observation, approximately 2% of SNPs are missing on average. We imputed all the missing values using the R package `missForest` available in CRAN. This package uses a random forest trained based on the observed entries to predict those missing values.

To formulate the testing hypotheses, we adopt a data splitting procedure. More specifically, we first randomly sample 20% of the observations and perform some preliminary analysis based on these observations. Results are given in Section S7.1 and we find 13 SNPs that are highly correlated with the phenotype. Based on the remaining 80% of the observations, we focus on testing whether the regression coefficients of these 13 variables are zero using logistic regression. The p-values of the partial penalized Wald, score and likelihood ratio statistics are 6.8×10^{-3}, 7.4×10^{-3} and 7.0×10^{-3} respectively. Under the significance level of 0.05, we reject the null hypothesis and conclude that at least one of the 13 regression coefficients is not equal to zero.

Since each covariate X_i^j is discrete and takes value on $\{0, 1, 2\}$, we can define the dummy variables $Z_i^{(j,1)} = I(X_i^j = 1)$, $Z_i^{(j,2)} = I(X_i^j = 2)$ and use these $\boldsymbol{Z}^{(j,m)}$'s as covariates in our logistic regression model. This yields a total of 554 covariates. Denoted by $\beta_{j,m}$ the corresponding regression coefficient of $\boldsymbol{Z}^{(j,m)}$, our next goal is to test the following hypothesis based on the remaining 80% of the samples:

$$H_0 : 2\beta_{j,1} = \beta_{j,2}, \quad \forall j \in \mathcal{M},$$

where $\mathcal{M}$ denotes the set of the 13 SNPs selected in the preliminary analysis. Under H_0, we have $\beta_{j,1}\boldsymbol{X}^j = \beta_{j,1}\boldsymbol{Z}^{(j,1)} + \beta_{j,2}\boldsymbol{Z}^{(j,2)}$ for any $j \in \mathcal{M}$. This corresponds to a lack of fit test regarding these important variables. The p-values of the Wald, score and likelihood ratio statistics are $0.083, 0.069$ and 0.086 respectively. Hence, we fail to reject H_0.

S7.1. Preliminary analysis in the real data application. We apply sure independence screening (Fan and Song, 2010) across all 277 variables. Specifically, we independently fit 277 logistic regressions by maximizing the marginal likelihood with the response and each univariate covariate and obtain the p-values from each marginal model. We find out there are a total of 13 variables with p-values smaller than 0.05. We report these variables and their associated p-values in the table below.

SNP	22	50	68	126	134	138	141	151	152	187	199	260	272
p-value (%)	0.4	3.5	4.2	1.9	4.0	4.8	4.3	4.6	3.5	4.1	4.5	3.9	1.8

APPENDIX S8: ADDITIONAL SIMULATION RESULTS

S8.1. Simulations results for Poisson regression. We consider Poisson regression in this section. The data were generated from the following model,

$$\Pr(Y_i = k|\boldsymbol{X}_i) = \frac{\exp(-\lambda_i)\lambda_i^k}{k!},$$

where

$$\lambda_i = \exp\left(0.75X_i^{(1)} - (0.75 + h^{(1)})X_i^{(2)} + h^{(2)}(X_i^{(3)} + X_i^{(4)})\right),$$

and $\boldsymbol{X}_i \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ for some $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. Consider testing the following hypotheses:

$$H_0^{(1)} : \beta_{1,0} + \beta_{2,0} = 0, \quad v.s \quad H_a^{(1)} : \beta_{1,0} + \beta_{2,0} \neq 0$$
$$H_0^{(2)} : \beta_{3,0} = \beta_{4,0} = 0, \quad v.s \quad H_a^{(2)} : \beta_{3,0} \neq 0 \quad \text{or} \quad \beta_{4,0} \neq 0.$$

We set $h^{(2)} = 0$ when testing $H_0^{(1)}$, and set $h^{(1)} = 0$ when testing $H_0^{(2)}$. We use the following four settings: (i) $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$; (ii) $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$; (iii) $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}$; (iiii) $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}$. For each setting, we set $h^{(j)} = 0, 0.03, 0.06, 0.12$ when testing $H_0^{(j)}$. The sample size is set to be 500. We provide the rejection probabilities for $H_0^{(1)}$ and $H_0^{(2)}$ in Table S7. The kernel density estimates of three test statistics under $H_0^{(1)}$ and $H_0^{(2)}$ are plotted in Figure S7 and Figure S8.

It can be seen that the Type-I error rates are well controlled under the null hypotheses, and the powers increase as $h^{(1)}$ or $h^{(2)}$ increases.

S8.2. Tables and plots.

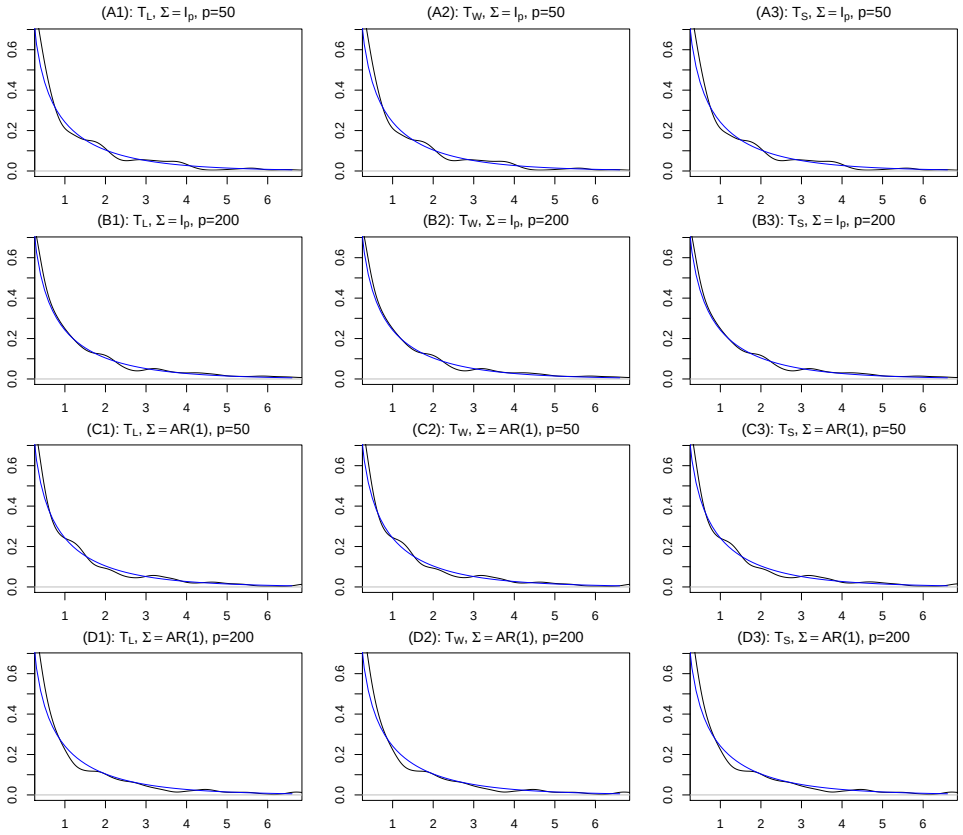

FIG S1. *Kernel density plot of three test statistics under* $H_0^{(1)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

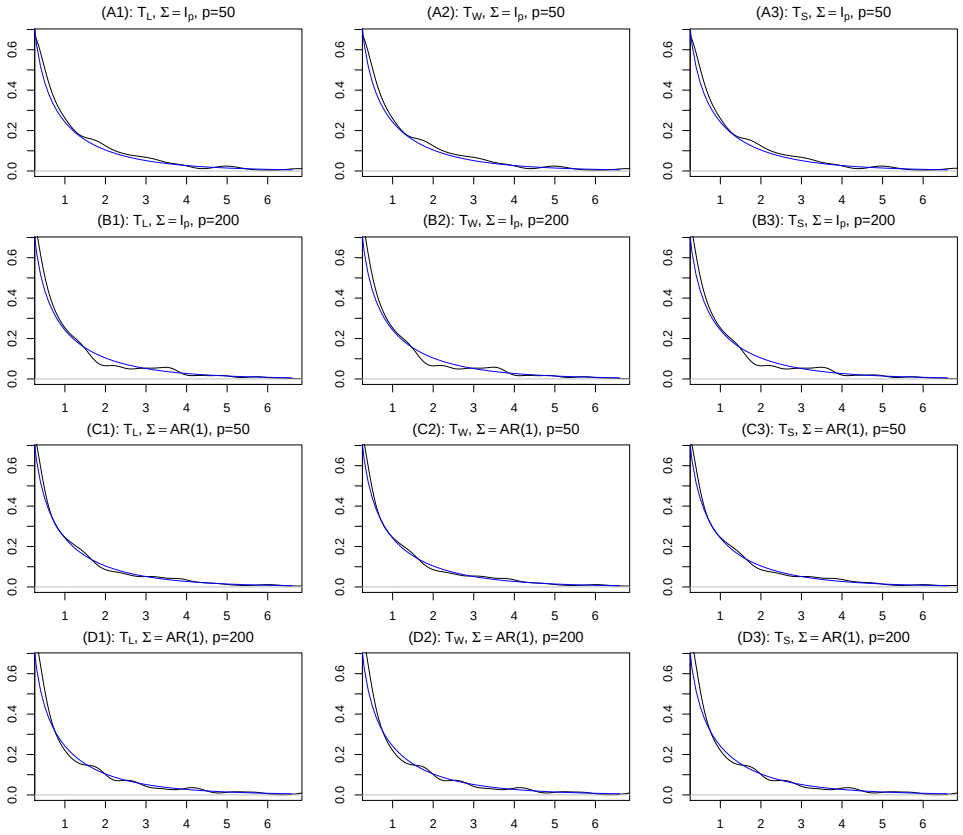

FIG S2. *Kernel density plot of three test statistics under* $H_0^{(2)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

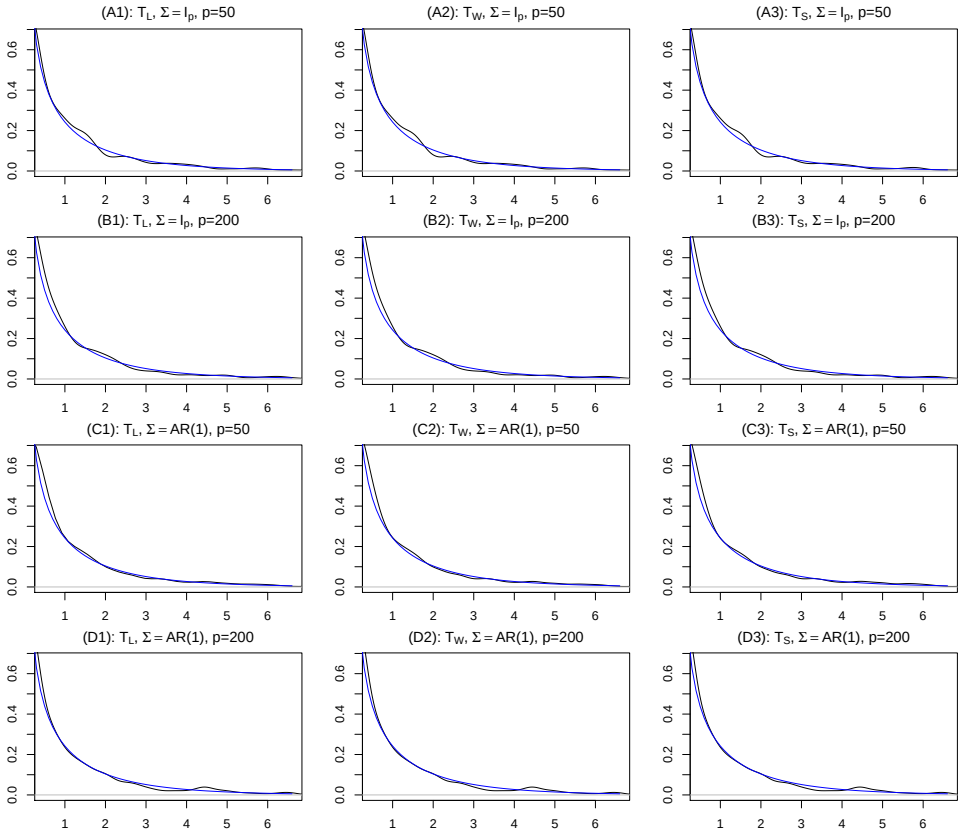

FIG S3. *Kernel density plot of three test statistics under* $H_0^{(2)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

TABLE S1

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%), under the setting where $\boldsymbol{\Sigma} = \boldsymbol{I}$.

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$			$H_0^{(1)}$			
0	3.50(0.75)	3.50(0.75)	3.50(0.75)	6.00(0.97)	6.00(0.97)	6.00(0.97)
0.1	10.33(1.24)	10.33(1.24)	10.33(1.24)	9.17(1.18)	9.17(1.18)	9.17(1.18)
0.2	26.67(1.81)	26.67(1.81)	26.67(1.81)	32.00(1.90)	32.00(1.90)	32.00(1.90)
0.4	83.17(1.53)	83.17(1.53)	83.33(1.52)	78.83(1.67)	78.83(1.67)	78.83(1.67)
$h^{(2)}$			$H_0^{(2)}$			
0	4.67(0.86)	4.67(0.86)	4.83(0.88)	4.17(0.82)	4.17(0.82)	4.33(0.83)
0.1	13.50(1.40)	13.50(1.40)	13.67(1.40)	10.17(1.23)	10.17(1.23)	10.33(1.24)
0.2	32.83(1.92)	32.67(1.91)	32.83(1.92)	27.80(1.83)	28.80(1.83)	28.80(1.83)
0.4	80.50(1.62)	80.50(1.62)	80.50(1.62)	80.50(1.62)	80.67(1.61)	80.67(1.61)
$h^{(2)}$			$H_0^{(3)}$			
0	5.00(0.89)	5.00(0.89)	5.17(0.90)	5.17(0.90)	5.17(0.90)	5.17(0.90)
0.1	10.83(1.27)	10.83(1.27)	11.00(1.28)	11.33(1.29)	11.33(1.29)	11.50(1.30)
0.2	28.67(1.85)	28.67(1.85)	28.83(1.85)	32.50(1.91)	32.50(1.91)	32.50(1.91)
0.4	80.67(1.61)	80.67(1.61)	80.67(1.61)	81.50(1.59)	81.50(1.59)	81.50(1.59)

TABLE S2

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics, the Wald test statistic based on the de-sparsified Lasso estimator and the decorrelated score statistic under the settings where $\mathbf{\Sigma} = \boldsymbol{I}$, with standard errors in parenthesis (%).

	T_L	T_W	T_S	T_W^D	T_S^D
$h^{(1)}$			$H_0^{(4)}$ and $p = 50$		
0	6.33(0.99)	6.50(1.01)	6.67(1.02)	5.17(0.90)	6.50(1.01)
0.1	18.67(1.59)	18.67(1.59)	18.83(1.60)	14.00(1.42)	18.50(1.59)
0.2	50.83(2.04)	50.83(2.04)	51.00(2.04)	41.33(2.01)	51.83(2.04)
0.4	95.67(0.83)	95.67(0.83)	95.67(0.83)	93.83(0.98)	96.00(0.80)
			$H_0^{(4)}$ and $p = 200$		
0	5.67(0.94)	5.67(0.94)	5.83(0.96)	5.17(0.90)	7.33(1.06)
0.1	15.67(1.48)	15.67(1.48)	15.67(1.48)	14.00(1.42)	15.67(1.48)
0.2	49.33(2.04)	49.33(2.04)	49.33(2.04)	41.33(2.01)	51.00(2.04)
0.4	97.33(0.66)	97.33(0.66)	97.33(0.66)	93.83(0.98)	97.67(0.62)
$h^{(2)}$			$H_0^{(5)}$ and $p = 50$		
0	6.33(0.99)	6.33(0.99)	6.50(1.00)	5.83(0.96)	6.50(1.01)
0.1	17.83(1.56)	17.83(1.56)	18.00(1.57)	14.17(1.42)	16.83(1.53)
0.2	53.00(2.04)	53.00(2.04)	53.00(2.04)	43.33(2.02)	52.17(2.04)
0.4	97.33(0.66)	97.33(0.66)	97.33(0.66)	95.17(0.88)	96.67(0.73)
			$H_0^{(5)}$ and $p = 200$		
0	4.00(0.80)	4.00(0.80)	4.00(0.80)	4.67(0.86)	3.17(0.71)
0.1	16.67(1.52)	16.67(1.52)	16.83(1.53)	13.67(1.40)	14.00(1.42)
0.2	49.17(2.04)	49.17(2.04)	49.17(2.04)	39.67(2.00)	43.17(2.02)
0.4	97.50(0.64)	97.50(0.64)	97.50(0.64)	92.50(1.08)	95.83(0.82)

TABLE S3

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%), under the settings where $\boldsymbol{\Sigma} = \boldsymbol{I}$.

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$	$H_0^{(6)}$					
0	5.50(0.93)	5.50(0.93)	5.67(0.93)	6.50(1.01)	6.50(1.01)	6.50(1.01)
0.2	15.00(1.46)	15.00(1.46)	15.67(1.46)	16.17(1.50)	16.17(1.50)	16.17(1.50)
0.4	50.33(2.04)	50.33(2.04)	50.50(2.04)	49.17(2.04)	49.17(2.04)	49.17(2.04)
0.8	97.50(0.64)	97.50(0.64)	97.50(0.64)	96.50(0.75)	96.50(0.75)	96.50(0.75)
$h^{(1)}$	$H_0^{(7)}$					
0	5.00(0.89)	5.00(0.89)	5.17(0.90)	5.83(0.96)	5.83(0.96)	5.83(0.96)
0.2	9.33(1.19)	9.17(1.18)	9.50(1.20)	13.00(1.37)	13.00(1.37)	13.00(1.37)
0.4	30.00(1.87)	30.00(1.87)	30.17(1.87)	26.83(1.81)	26.83(1.81)	26.83(1.81)
0.8	76.33(1.74)	76.67(1.73)	76.67(1.73)	76.67(1.73)	76.67(1.73)	76.67(1.73)
$h^{(1)}$	$H_0^{(8)}$					
0	4.83(0.88)	4.83(0.88)	5.00(0.89)	6.83(1.03)	6.83(1.03)	6.83(1.03)
0.2	8.83 (1.16)	8.67(1.15)	9.00(1.17)	10.33(1.24)	10.33(1.24)	10.33(1.24)
0.4	21.33(1.67)	21.33(1.67)	21.50(1.68)	20.50(1.65)	20.50(1.65)	20.50(1.65)
0.8	55.83(2.03)	56.17(2.03)	56.00(2.03)	58.33(2.01)	58.33(2.01)	58.33(2.01)

TABLE S4

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%).

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$	$H_0^{(1)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	5.33(0.92)	5.17(0.90)	5.33(0.92)	6.00(0.97)	5.33(0.92)	6.00(0.97)
0.2	11.33(1.29)	10.17(1.23)	10.83(1.27)	13.67(1.40)	13.00(1.37)	13.50(1.40)
0.4	33.33(1.92)	32.00(1.90)	33.17(1.92)	38.83(1.99)	37.67(1.98)	38.67(1.99)
0.8	88.50(1.30)	88.67(1.29)	88.83(1.29)	88.83(1.29)	88.17(1.32)	88.67(1.29)
	$H_0^{(1)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	6.50(1.01)	6.00(0.97)	6.33(0.99)	4.50(0.85)	4.33(0.83)	4.33(0.83)
0.2	19.17(1.61)	18.67(1.59)	19.00(1.60)	21.00(1.67)	20.33(1.64)	20.67(1.65)
0.4	61.50(1.99)	61.17(1.99)	61.50(1.99)	62.00(1.98)	62.17(1.98)	62.00(1.98)
0.8	98.33(0.52)	99.67(0.24)	99.67(0.24)	98.50(0.50)	99.33(0.33)	99.33(0.33)
$h^{(2)}$	$H_0^{(2)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	6.17(0.98)	6.00(0.97)	5.83(0.96)	5.67(0.94)	5.50(0.93)	5.83(0.96)
0.2	7.83(1.10)	9.17(1.12)	9.17(1.12)	8.00(1.11)	10.17(1.23)	10.00(1.22)
0.4	23.50(1.73)	26.50(1.80)	26.67(1.81)	23.83(1.74)	26.67(1.81)	26.83(1.81)
0.8	71.00(1.85)	76.00(1.74)	75.50(1.76)	74.00(1.79)	77.67(1.70)	77.50(1.70)
	$H_0^{(2)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	5.67(0.94)	5.33(0.92)	5.67(0.94)	4.50(0.85)	4.17(0.82)	4.17(0.82)
0.2	11.17(1.29)	12.67(1.36)	12.67(1.36)	11.00(1.28)	12.50(1.35)	12.50(1.35)
0.4	31.17(1.89)	35.33(1.95)	35.17(1.95)	33.00(1.92)	35.83(1.96)	35.67(1.96)
0.8	84.50(1.48)	88.33(1.31)	88.33(1.33)	88.67(1.29)	90.17(1.22)	90.00(1.22)
$h^{(2)}$	$H_0^{(3)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	4.17(0.82)	3.67(0.77)	3.67(0.77)	6.17(0.98)	5.67(0.94)	5.83(0.96)
0.2	8.67(1.15)	8.83(1.13)	8.83(1.13)	14.50(1.44)	13.67(1.40)	14.17(1.42)
0.4	35.83(1.96)	34.17(1.94)	35.50(1.95)	39.50(2.00)	39.00(1.99)	39.17(1.99)
0.8	90.00(1.22)	89.00(1.28)	89.67(1.24)	90.17(1.22)	89.67(1.24)	90.00(1.22)
	$H_0^{(3)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	4.83 (0.88)	4.67(0.86)	4.67(0.86)	5.83(0.96)	5.33(0.92)	5.67(0.94)
0.2	18.00 (1.57)	17.00(1.53)	17.67(1.56)	18.50(1.59)	18.17(1.57)	18.33(1.58)
0.4	55.17 (2.03)	54.50(2.03)	55.00(2.03)	53.83(2.04)	53.33(2.04)	53.83(2.04)
0.8	98.33 (0.52)	99.00(0.41)	99.00(0.41)	98.50(0.50)	98.33(0.44)	98.33(0.44)

TABLE S5

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics, the Wald test statistic based on the de-sparsified Lasso estimator and the decorrelated score statistic, with standard errors in parenthesis (%).

	T_L	T_W	T_S	T_W^D	T_S^D
$h^{(1)}$	$H_0^{(4)}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $p = 50$				
0	4.83(0.86)	4.67(0.86)	4.67(0.86)	51.17(2.04)	7.33 (1.06)
0.2	12.50(1.35)	8.00(1.11)	8.33(1.12)	23.33(1.73)	18.67 (1.59)
0.4	27.17(1.82)	20.83(1.66)	21.33(1.67)	9.00(1.17)	37.17 (1.97)
0.8	77.00(1.72)	70.67(1.86)	71.83(1.84)	15.50(1.48)	82.67 (1.55)
	$H_0^{(4)}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $p = 200$				
0	4.17(0.82)	4.33(0.83)	4.33(0.83)	70.67(1.86)	8.83(1.16)
0.2	12.00(1.33)	9.17(1.18)	9.17(1.18)	43.17(2.02)	22.33(1.70)
0.4	31.17(1.89)	23.33(1.73)	24.33(1.75)	24.00(1.74)	46.67(2.04)
0.8	78.33(1.67)	70.83(1.86)	71.83(1.84)	15.00(1.46)	89.33(1.26)
	$H_0^{(4)}$, $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$ and $p = 50$				
0	6.00(0.97)	5.33(0.92)	5.50(0.93)	31.00(1.89)	7.00(1.04)
0.2	13.50(1.40)	10.83(1.27)	11.17(1.29)	11.00(1.28)	14.67(1.44)
0.4	33.83(1.93)	28.17(1.84)	29.17(1.86)	8.00(1.11)	38.83(1.99)
0.8	82.50(1.55)	78.33(1.68)	78.67(1.67)	37.33(1.97)	89.33(1.26)
	$H_0^{(4)}$, $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$ and $p = 200$				
0	4.33(0.83)	4.00(0.80)	4.00(0.80)	57.83(2.02)	7.50(1.08)
0.2	10.67(1.26)	8.00(1.11)	8.50(1.14)	27.83(1.83)	18.67(1.59)
0.4	32.00(1.90)	27.00(1.81)	27.33(1.82)	13.17(1.38)	47.50(2.04)
0.8	79.67(1.64)	75.17(1.76)	75.17(1.76)	23.67(1.73)	88.50(1.30)
$h^{(2)}$	$H_0^{(5)}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $p = 50$				
0	5.33(0.92)	5.00(0.89)	5.17(0.90)	2.50(0.64)	4.17(0.82)
0.2	19.83(1.63)	19.33(1.61)	19.83(1.63)	9.83(1.22)	17.67(1.56)
0.4	62.50(1.98)	60.67(1.99)	61.83(1.98)	42.00(2.01)	60.33(2.00)
0.8	99.50(0.29)	99.50(0.29)	99.50(0.29)	97.50(0.64)	99.33(0.33)
	$H_0^{(5)}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $p = 200$				
0	5.00(0.89)	4.50(0.85)	4.83(0.88)	3.00(0.70)	3.00(0.70)
0.2	22.50(1.70)	21.83(1.69)	22.33(1.70)	13.17(1.38)	18.67(1.59)
0.4	64.00(1.96)	62.67(1.97)	63.67(1.96)	47.67(2.04)	58.17(2.01)
0.8	99.50(0.29)	99.67(0.24)	99.67(0.24)	98.00(0.57)	99.17(0.37)
	$H_0^{(5)}$, $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$ and $p = 50$				
0	5.50(0.93)	4.83(0.88)	5.50(0.93)	3.67(0.77)	3.00(0.70)
0.2	21.33(1.67)	20.83(1.66)	21.00(1.66)	8.00(1.11)	15.33(1.47)
0.4	63.33(1.97)	62.67(1.97)	63.50(1.97)	34.67(1.94)	53.00(2.04)
0.8	99.17(0.37)	99.83(0.17)	99.83(0.17)	95.17(0.88)	99.17(0.37)
	$H_0^{(5)}$, $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$ and $p = 200$				
0	7.67(1.09)	7.17(1.05)	7.67(1.09)	5.50(0.93)	4.00(0.80)
0.2	22.00(1.66)	21.00(1.69)	21.83(1.66)	9.00(1.17)	15.50(1.48)
0.4	64.50(1.95)	64.00(1.96)	64.33(1.96)	31.67(1.90)	48.00(2.04)
0.8	99.00(0.41)	99.50(0.29)	99.50(0.29)	94.33(0.94)	97.50(0.64)

TABLE S6

Rejection probabilities (%) of the partial penalized Wald, score and likelihood ratio statistics with standard errors in parenthesis (%).

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$	$H_0^{(6)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	5.67(0.94)	5.17(0.90)	5.50(0.93)	5.17(0.90)	4.83(0.88)	5.17(0.90)
0.4	19.17(1.61)	18.67(1.59)	19.17(1.61)	21.50(1.68)	21.00(1.66)	21.33(1.67)
0.8	61.50(1.99)	61.17(1.99)	61.33(1.99)	59.00(2.01)	57.50(2.02)	58.83(2.01)
1.6	97.67(0.62)	97.83(0.59)	97.83(0.59)	98.00(0.57)	98.00(0.57)	98.17(0.55)
	$H_0^{(6)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	4.33(0.83)	3.83(0.78)	4.33(0.80)	4.83(0.88)	4.67(0.86)	4.83(0.88)
0.4	45.33(2.03)	44.67(2.03)	45.17(2.03)	45.50(2.03)	44.33(2.03)	45.00(2.03)
0.8	93.50(1.01)	94.00(0.97)	94.17(0.96)	93.00(1.04)	92.67(1.06)	93.00(1.04)
1.6	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)
$h^{(1)}$	$H_0^{(7)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	5.83(0.96)	5.67(0.94)	5.83(0.96)	5.83(0.96)	5.17(0.90)	5.50(0.93)
0.4	15.33(1.47)	15.00(1.46)	15.00(1.46)	13.67(1.40)	13.17(1.38)	13.50(1.40)
0.8	36.00(1.96)	34.67(1.94)	35.50(1.95)	37.00(1.97)	35.67(1.96)	36.67(1.97)
1.6	82.67(1.55)	82.00(1.57)	82.50(1.55)	82.50(1.55)	81.17(1.60)	82.67(1.55)
	$H_0^{(7)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	6.00(0.97)	5.33(0.92)	5.50(0.93)	6.00(0.97)	5.83(0.96)	6.00(0.97)
0.4	32.83(1.92)	31.50(1.90)	32.33(1.91)	33.50(1.93)	32.67(1.91)	32.83(1.92)
0.8	78.83(1.67)	78.00(1.69)	78.83(1.68)	77.33(1.71)	76.83(1.72)	77.50(1.70)
1.6	99.50(0.29)	99.67(0.24)	99.67(0.24)	99.50(0.29)	99.83(0.17)	99.83(0.17)
$h^{(1)}$	$H_0^{(8)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	7.83(1.10)	7.33(1.06)	7.83(1.10)	5.83(0.96)	5.33(0.92)	5.67(0.94)
0.4	12.33(1.34)	12.00(1.33)	12.33(1.34)	12.00(1.33)	11.50(1.30)	11.67(1.31)
0.8	27.50(1.82)	26.33(1.80)	27.17(1.82)	26.00(1.79)	25.33(1.78)	26.00(1.79)
1.6	64.50(1.95)	61.00(1.99)	63.17(1.97)	66.50(1.93)	63.50(1.97)	65.67(1.94)
	$H_0^{(8)}$ and $\boldsymbol{\Sigma} = \{0.5^{\|i-j\|}\}$					
0	5.33 (0.92)	4.83(0.88)	5.00(0.89)	5.50(0.93)	5.50(0.93)	5.50(0.93)
0.4	23.00 (1.72)	22.00(1.70)	22.50(1.70)	25.00 (1.77)	24.00(1.74)	24.33(1.75)
0.8	66.00 (1.93)	63.50(1.97)	65.17(1.95)	64.17 (1.97)	62.83(1.96)	63.50(1.97)
1.6	98.00 (0.57)	98.17(0.55)	98.33(0.52)	97.67(0.62)	98.00(0.57)	98.00(0.57)

TABLE S7

Rejection probabilities (%) of the likelihood ratio (denoted by T_L), Wald (denoted by T_W) and score statistic (denoted by T_S) for testing $H_0^{(1)}$, with standard errors in parenthesis (%).

	$p = 50$			$p = 200$		
	T_L	T_W	T_S	T_L	T_W	T_S
$h^{(1)}$	$H_0^{(1)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	4.50(0.85)	4.67(0.86)	4.67(0.86)	4.17(0.82)	4.17(0.82)	4.17(0.82)
0.03	8.17(1.12)	8.17(1.12)	8.17(1.12)	10.83(1.27)	10.83(1.27)	10.83(1.27)
0.06	29.00(1.85)	29.50(1.86)	29.17(1.86)	28.50(1.84)	28.50(1.84)	28.50(1.84)
0.12	76.33(1.74)	76.67(1.73)	76.50(1.73)	73.50(1.80)	73.33(1.81)	73.50(1.80)
	$H_0^{(1)}$ and $\boldsymbol{\Sigma} = \{0.5^{\lvert i-j\rvert}\}$					
0	4.83(0.88)	4.83(0.88)	4.83(0.88)	7.17(1.05)	7.17(1.05)	7.17(1.05)
0.03	9.33(1.19)	9.50(1.20)	9.33(1.19)	11.00(1.28)	11.00(1.28)	11.00(1.28)
0.06	30.67(1.88)	30.67(1.88)	30.83(1.89)	29.50(1.86)	29.50(1.86)	29.50(1.86)
0.12	79.67(1.64)	79.67(1.64)	79.67(1.64)	77.17(1.71)	77.17(1.71)	77.17(1.71)
$h^{(2)}$	$H_0^{(2)}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$					
0	5.33(0.92)	5.17(0.9)	5.50(0.93)	4.50(0.85)	4.50(0.85)	5.00(0.89)
0.03	19.17(1.61)	19.00(1.60)	19.67(1.62)	20.33(1.64)	20.01(1.66)	21.67(1.68)
0.06	57.33(2.02)	57.67(2.02)	57.67(2.02)	61.33(1.99)	61.67(1.98)	62.00(1.98)
0.12	99.33(0.33)	99.33(0.33)	99.33(0.33)	99.00(0.41)	99.17(0.37)	99.33(0.33)
	$H_0^{(2)}$ and $\boldsymbol{\Sigma} = \{0.5^{\lvert i-j\rvert}\}$					
0	3.83(0.78)	3.83(0.78)	3.83(0.78)	6.17(0.98)	6.17(0.98)	6.67(1.02)
0.03	16.00(1.50)	16.50(1.52)	16.33(1.51)	18.33(1.58)	18.00(1.57)	18.33(1.58)
0.06	54.50(2.03)	54.33(2.03)	54.33(2.03)	55.50(2.03)	55.00(2.03)	56.33(2.02)
0.12	99.00(0.41)	99.17(0.37)	99.17(0.37)	99.33(0.33)	99.33(0.33)	99.33(0.33)

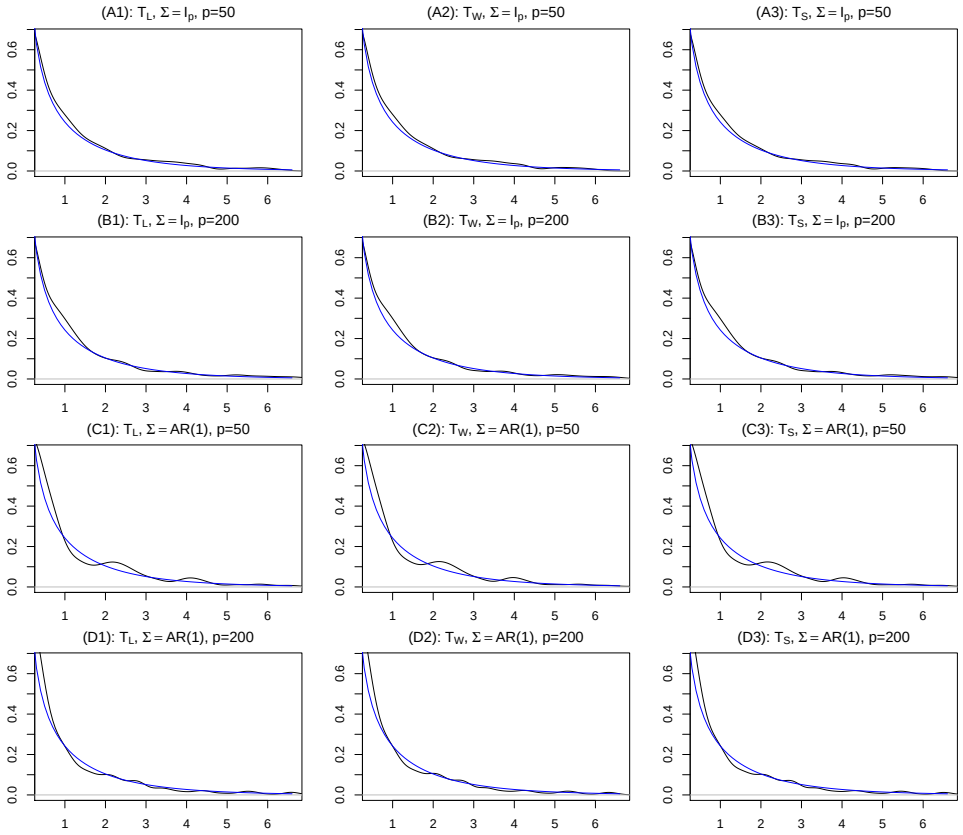

FIG S4. *Kernel density plot of three test statistics under* $H_0^{(1)}$ *with different combinations of* p *and the covariance matrices.* T_L, T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

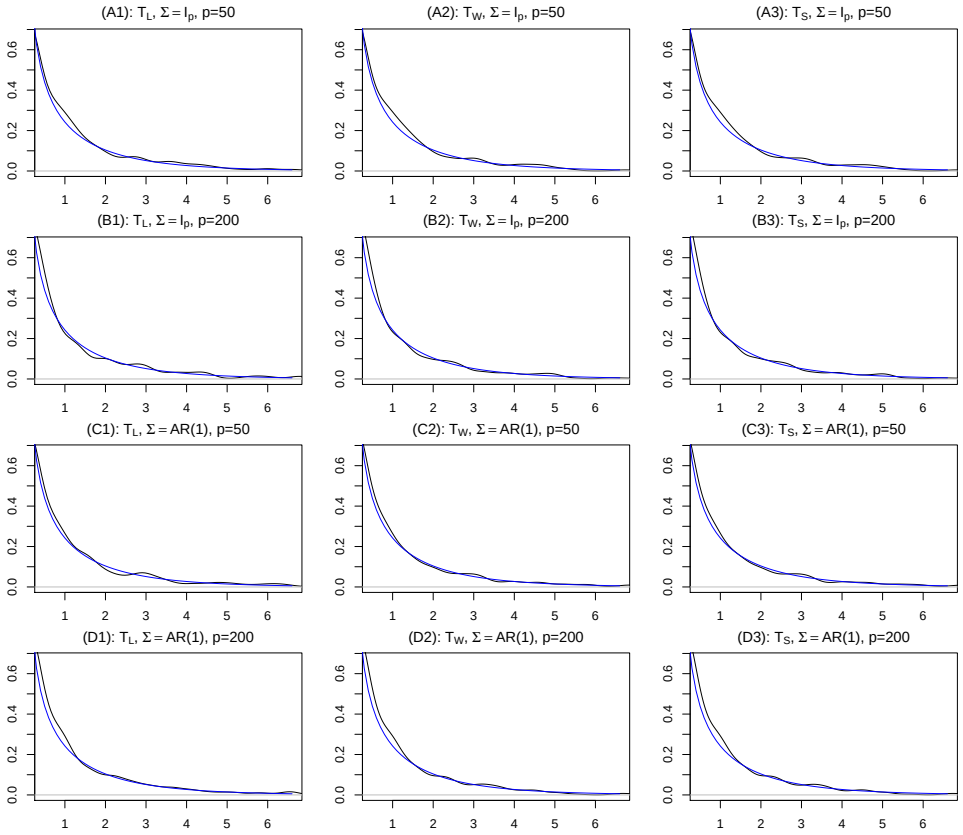

FIG S5. *Kernel density plot of three test statistics under* $H_0^{(2)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

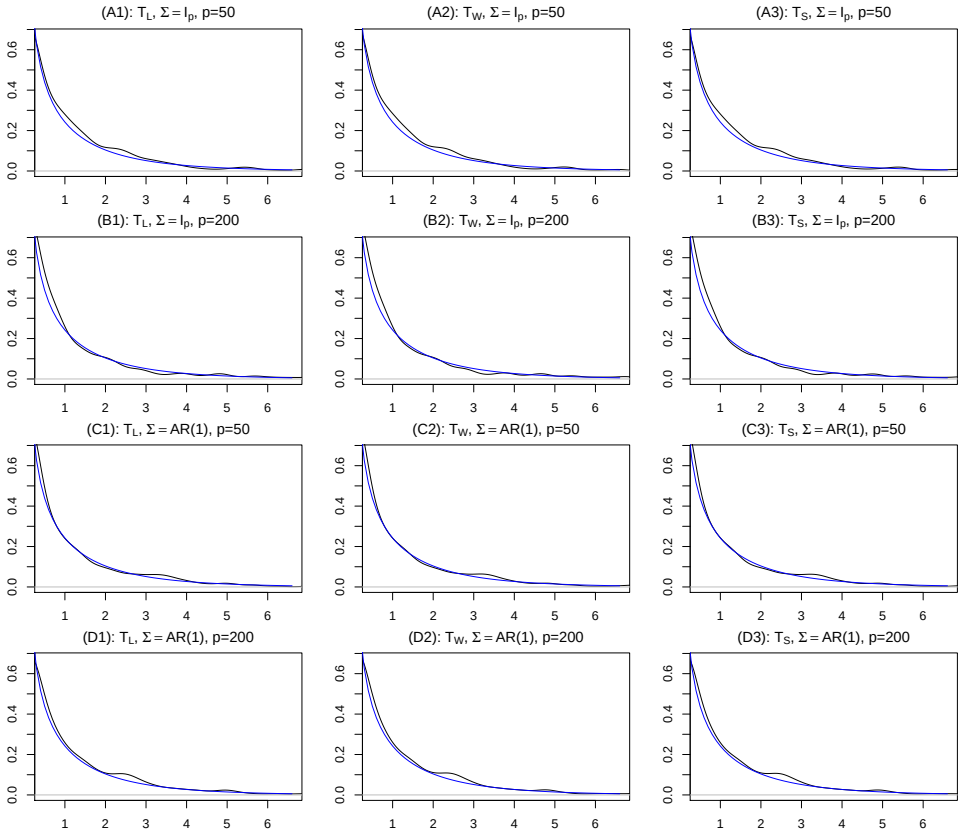

FIG S6. *Kernel density plot of three test statistics under* $H_0^{(2)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 200, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ *and* $p = 50, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \boldsymbol{\Sigma} = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

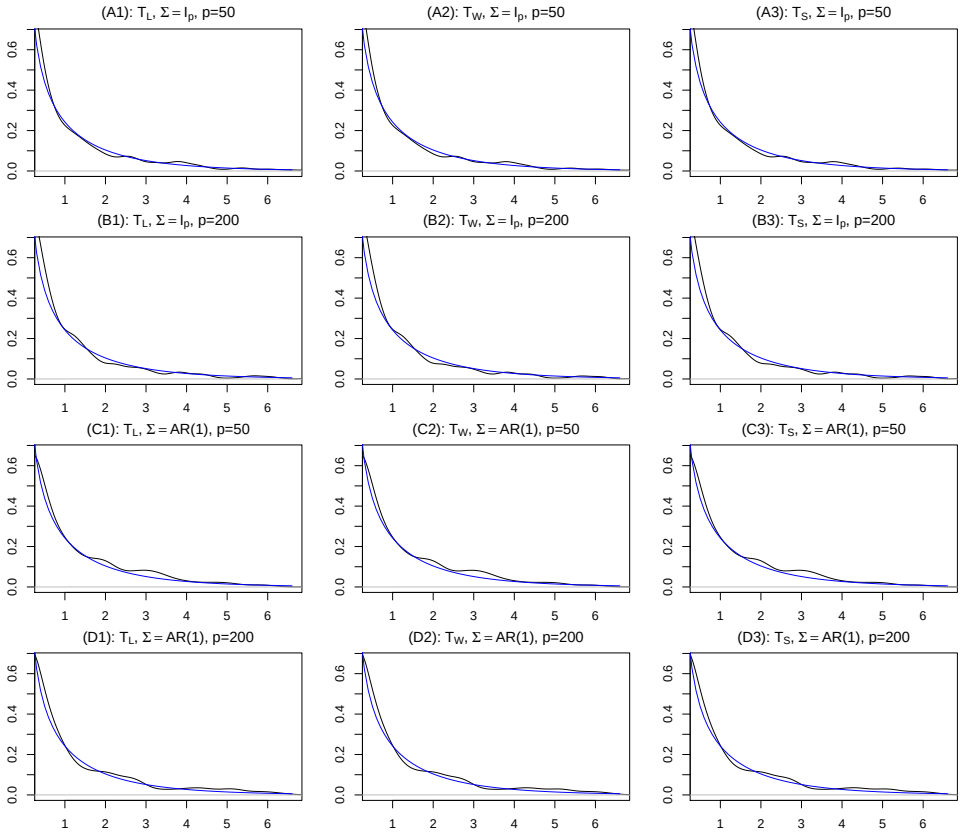

FIG S7. *Histograms of three test statistics under* $H_0^{(1)}$ *with different combinations of* p *and the covariance matrices.* T_L*,* T_W *and* T_S *from left to right.* $p = 50, \Sigma = I_p$ *and* $p = 200, \Sigma = I_p$ *and* $p = 50, \Sigma = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \Sigma = \Sigma = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 1 *degree of freedom.*

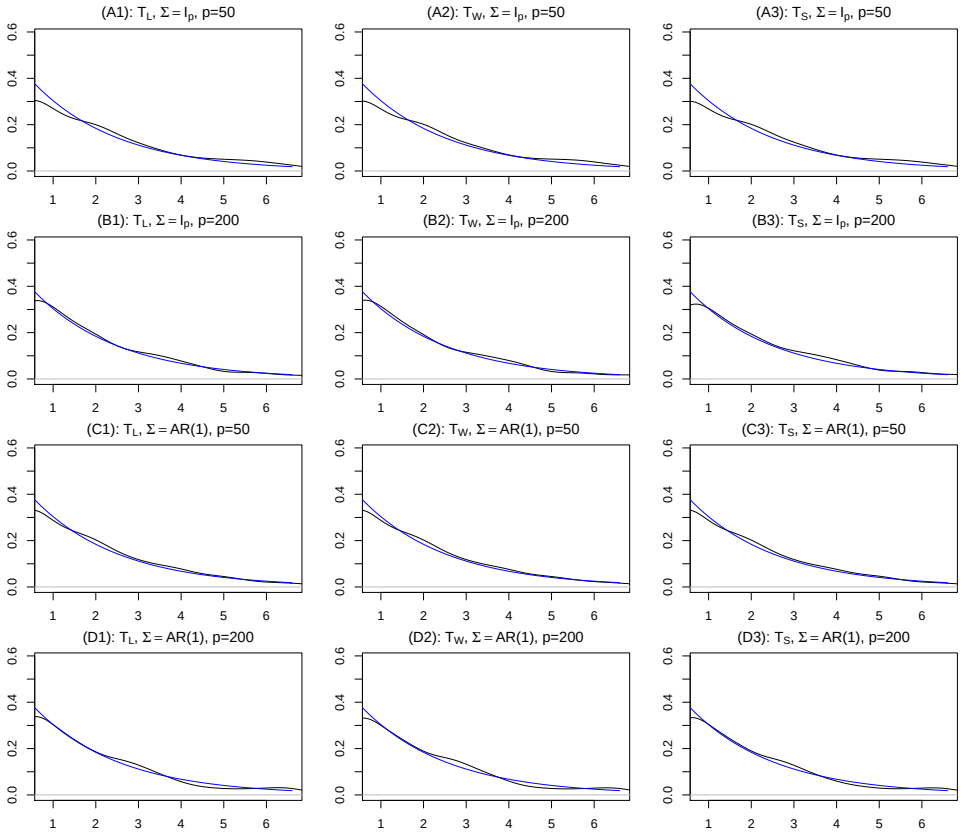

FIG S8. *Histograms of three test statistics under* $H_0^{(2)}$ *with different combinations of* p *and the covariance matrices.* T_L, T_W *and* T_S *from left to right.* $p = 50, \Sigma = I_p$ *and* $p = 200, \Sigma = I_p$ *and* $p = 50, \Sigma = \{0.5^{|i-j|}\}_{i,j}$ *and* $p = 200, \Sigma = \Sigma = \{0.5^{|i-j|}\}_{i,j}$ *from top to bottom. The black line plot the density function of a* χ^2 *distribution with* 2 *degrees of freedom.*

REFERENCES

Bentkus, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.* **49** 400–410.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604.

Lv, J. and Fan, Y. (2009). A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares. *Ann. Statist.* **37** 3498–3528.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38** 904–909.

Shi, C., Fan, A., Song, R. and Lu, W. (2017). High-dimensional A-learning for optimal dynamic treatment regimes. *Ann. Statist.* **Accepted**.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202.

Chengchun Shi and Rui Song
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203
USA
E-mail: cshi4@ncsu.edu
E-mail: rsong@ncsu.edu

Zhao Chen and Runze Li
Department of Statistics,
and The Methodology Center
the Pennsylvania State University,
University Park, PA 16802-2111
USA
E-mail: zuc4@psu.edu
E-mail: rzli@psu.edu