VARIABLE SELECTION FOR HIGH DIMENSIONAL MULTIVARIATE OUTCOMES

Tamar Sofer, Lee Dicker and Xihong Lin

Harvard School of Public Health and Rutgers University

Supplementary Material

Contents

S1 $p < n$ main results S4
S1.1 The technical conditions S4
S1.2 Proof of Theorem 1 . . . . . S4
S1.3 Proof of Lemma 1 . . . . . S8
S1.4 Proof of Theorem 2 . . . . . S10
S2 Consistency of the BIC S12
S2.1 Proof of Theorem 3 . . . . . S12
S3 Large p main results S17
S3.1 Modified technical conditions S17
S3.2 Proof of Theorem 4 . . . . . S17
S3.3 Proof of corollary S19
S4 Secondary lemmas S20
S5 Additional small-p simulations results S25

S6 Additional large-p simulations results S31

S7 Data set and code: description and instructions S31

List of Tables

1 Small-p simulation results for the 4 scenarios, by outcome correlation matrix and sample size, when using the two stage algorithm with the identity as the working correlation matrix. S26

2 Small-p simulation results. Estimated and empirical standard errors for the regression parameter estimators provided in Table 1. . . . . S27

3 Small-p simulations results for the 4 scenarios, by outcome correlation matrix and sample size, when using an iterative algorithm in which the regression parameters and the precision matrix are alternately estimated, and tuning parameters are selected at each iteration. S28

4 Small-p simulations results. Estimated and empirical standard errors for the regression parameter estimators provided in Table 3. . . . . S29

5 Small-p simulation results for the 4 scenarios, by outcome correlation matrix and sample size, when using the two stage algorithm with the true correlation structures as the working correlation matrix. S30

6 Large-p simulation results when the working correlation matrix is the true outcome correlation matrix. S32

7 Large-p simulation results when the working correlation matrix is the identity matrix. S33

S1 $p < n$ main results

Throughout this section , let $\boldsymbol{\beta}^*$ be the true regression parameters vector with $A = \{j : \beta_j^* \neq 0\}$, $|A| = s_n$, and $\boldsymbol{\beta}_A$ the sub-vector and of a vector $\boldsymbol{\beta}$ corresponding to the indices in A and $\mathbf{X}_A$ the sub-matrix of $\mathbf{X}$ with columns corresponding to the set A.

S1.1 The technical conditions

(C1) $n \to \infty$, $p = p_0 m$ may vary with n, and $p/n \to 0$.

(C2) There exists a positive constant R such that

$$0 < 1/R < \begin{array}{l} \lambda_{min}(\boldsymbol{\Lambda}_m), \lambda_{min}(\boldsymbol{\Omega}_m), \lambda_{min}(n^{-1}\mathbf{X}^T\mathbf{X}), \\ \lambda_{max}(\boldsymbol{\Lambda}_m), \lambda_{max}(\boldsymbol{\Omega}_m), \lambda_{max}(n^{-1}\mathbf{X}^T\mathbf{X}) \end{array} < R < \infty$$

where $\lambda_{min}(\boldsymbol{B})$ and $\lambda_{max}(\boldsymbol{B})$ are the smallest and largest eigenvalues of a matrix $\boldsymbol{B}$.

(C3) Let $\rho = \rho_n = \min\{|\beta_j|;\ j \in A\}$. $\rho/\sqrt{p/n} \to \infty$

(C4) $\max_{1 \leq i \leq n} n^{-1}||\boldsymbol{x}_i\boldsymbol{x}_i^T||_2 \to 0$ (where $||\boldsymbol{x}_i\boldsymbol{x}_i^T||_2 = ||\boldsymbol{x}_i||_2^2$)

(C5) There exists a δ such that $E(\epsilon_{ij}^{2+\delta}) < \infty$, $i = 1, ..., n$, $j = 1, ..., m$.

(C6) The function P_λ is concave on $[0, \infty)$ and differentiable on $(0, \infty)$. Furthermore, $P_\lambda(0) = 0$, $P_\lambda(\theta) = P_\lambda(-\theta)$ and $\lim_{\theta \to \infty} P_\lambda(\theta) \leq \frac{1}{n}$.

(C7) If $r_n/\sqrt{p/n} \to \infty$, then $P_\lambda'(r_n) = o(1/\sqrt{np})$.

(C8) If $r_n = O(\sqrt{p/n})$, then $\lim_{n \to \infty} (\sqrt{n/\max(p, k_n)})P_\lambda'(r_n) \to \infty$

(C9) $pm, m^2/n \to 0$

(C10) $\sup_{j,k} E(\epsilon_{ij}\epsilon_{ik})^{2+\delta} < \infty$, $i = 1, \ldots, n,\ \ j, k = 1, \ldots, m$.

(C11) If $r_n/\sqrt{(m+p)m/n} \to \infty$, then $mP_\gamma'(r_n) = O(1)$.

(C12) If $r_n = O(\sqrt{(m+p)m/n})$, then $\lim_{n \to \infty} \sqrt{(m+p)m/n}P_\gamma'(r_n) = \infty$.

S1.2 Proof of Theorem 1

part (a) (Consistency)

Fix $c > 0$ and let M be a positive constant. We will show that if M is sufficiently large, then

$$P\left\{\inf_{||\boldsymbol{u}||=1} Q(\boldsymbol{\beta}^* + M\sqrt{p/n}\boldsymbol{u}|\boldsymbol{\Lambda}_m) - Q(\boldsymbol{\beta}^*|\boldsymbol{\Lambda}_m) > 0\right\} \geq 1 - c$$

for n sufficiently large. This suffices to prove the (a).

Let

$$D(\boldsymbol{u}) = Q(\boldsymbol{\beta}^* + \sqrt{p/n}\boldsymbol{u}|\boldsymbol{\Lambda}_m) - Q(\boldsymbol{\beta}^*|\boldsymbol{\Lambda}_m) = J_1 + J_2 + J_3,$$

where $||\boldsymbol{u}|| = M$ and $J_1(\boldsymbol{u}) = -2\sqrt{p/n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$, $J_2(\boldsymbol{u}) = (p/n)\boldsymbol{u}^T\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$ and $J_3(\boldsymbol{u}) = 2n\sum_{j=1}^p \Big[P_\lambda(|\beta_j^* + \sqrt{p/n}u_j|) - P_\lambda(|\beta_j^*|)\Big]$. We bound the terms J_1, J_2, and J_3 separately. First,

$$|J_1| \le 2\sqrt{p/n}n||n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}||||\boldsymbol{u}|| = 2Mn\sqrt{p/n}O_p(\sqrt{p/n}) = O_p(p).$$

For $J_2(\boldsymbol{u})$, we have that $J_2(\boldsymbol{u}) \ge R^{-2}M^2p$.
Notice that there exist N_1, K_1 such that for $n > N_1$ $J_1 \le K_1 p$ and $J_2 \ge R^{-2}M^2p$. Therefore, $J_2 - J_1 \ge p(R^{-2}M^2 - K_1)$. Since M is arbitrary, for large enough M such that $J_2 - J_1 > K'p$ with probability tending to 1, for some constant K'.

$$\text{Decompose } J_3(\boldsymbol{u}) = 2n\sum_{j\in A}\Big[P_\lambda(|\beta_j^* + \sqrt{p/n}u_j|) - P_\lambda(|\beta_j^*|)\Big] + 2n\sum_{j\in A^c} P_\lambda(|\sqrt{p/n}u_j|).$$

Using the mean value theorem, and conditions (C7) and (C8), one can see that for some $0 < t_j < 1, j \in A$

$$\begin{aligned} 2n\sum_{j\in A}\Big[P_\lambda(|\beta_j^* + \sqrt{p/n}u_j|) - P_\lambda(|\beta_j^*|)\Big] &= 2n\sqrt{p/n}\sum_{j\in A}|u_j|P_\lambda(|\beta_j^* + t_j\sqrt{p/n}u_j|) \\ &\le 2\sqrt{np}||\boldsymbol{u}||_1 o(1/\sqrt{np}) = o(1), \end{aligned}$$

while

$$2n\sum_{j\in A^c} P_\lambda(|\sqrt{p_0/n}u_j|) = 2n(p - s_n)\Theta(\sqrt{p/n}) = 2(p - s_n)\Theta(\sqrt{p/n})$$

It follows that $P(D(\boldsymbol{u}) > 0) \to 1$ as $n \to \infty$ for M large enough.

To show that the global minimum of $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m)$ satisfies $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_p(\sqrt{p/n})$, let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{u}$, where $||\boldsymbol{u}|| = \Theta(\sqrt{p/n})$. In this case we have $J_1 = -2\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$, $J_2 = \boldsymbol{u}^T\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$ and $J_3 = 2n\sum_{j=1}^p \big[P_\lambda(|\beta_j^* + u_j|) - P_\lambda(|\beta_j^*|)\big]$. J_2 and J_1 satisfy (using Lemma 2, condition (C2) and Cauchy-Schwartz inequality)

$$\begin{aligned} J_2 &\ge n||\boldsymbol{u}||^2\lambda_{\min}\left(\frac{1}{n}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\right) \ge n||\boldsymbol{u}||^2R^{-2} \\ |J_1| &= 2n\left\langle\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}, \boldsymbol{u}\right\rangle \le 2n||\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}||||\boldsymbol{u}|| = 2O_p(\sqrt{np})||\boldsymbol{u}|| \end{aligned}$$

Since $\sqrt{np} < n$, J_2 dominates J_1. Further, by (C6) the negative part of the penalty term satisfies:

$$2n\sum_{j\in A} P_\lambda(|\beta_j^*|) \le 2npP_\lambda(\max_{j\in A}(|\beta_j^*|)) \le 2ns\frac{1}{n} \le 2s.$$

From these, it follows that $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda}_m) - Q(\boldsymbol{\beta}^*|\boldsymbol{\Lambda}_m) > 0$ as $n \to \infty$ with probability tending to 1, implying that the penalized maximum likelihood function is not maximized outside the ball of radius $\sqrt{p/n}$. ∎

part (b) (Sparsistency)

Let $\widehat{\boldsymbol{\beta}}$ be a consistent estimator of the true $\boldsymbol{\beta}^*$, such that $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| < M\sqrt{p/n}$. We will show that for all j such that $\beta_j^* = 0$, the derivative of the penalized likelihood is maximized at 0, so that $\hat{\beta}_j = 0$. We do it be showing that the sign of the penalized likelihood is dominated by the sign of the penalty $P'_\lambda(\hat{\beta}_j)$, which is the sign of $\hat{\beta}_j$.

consider:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\beta}) &= -2\mathbf{X}^T\boldsymbol{\Lambda}\boldsymbol{y} + 2\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{\beta} + 2n\sum_{j=1}^{p}P'_\lambda(\beta_j) \\ &= -2\mathbf{X}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon} + 2\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + 2n\sum_{j=1}^{p}P'_\lambda(\beta_j)\end{aligned}$$

The derivative by β_j, where the true $\beta^* = 0$ satisfies:

$$\frac{\partial}{\partial\beta_j}Q(\boldsymbol{\beta}) = -2\boldsymbol{x}_j^T\boldsymbol{\Lambda}\boldsymbol{\epsilon} + 2\boldsymbol{x}_j^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + 2n\text{sign}(\beta_j)P'_\lambda(|\beta_j|) = I_1 + I_2 + I_3.$$

Consider I_1. By Lemma 3 (part (2))

$$|I_1| = 2\left|\boldsymbol{x}_j^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}\right| \leq 2\sqrt{n}\sup_{j=1,\ldots p}\left|\frac{1}{\sqrt{n}}\boldsymbol{x}_j^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}\right| = o_p(\sqrt{nk_n}).$$

Consider now I_2.

$$\begin{aligned}I_2 = 2\boldsymbol{x}_j^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*) &\leq 2||\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|| \\ &\leq 2n\lambda_1\left(\frac{1}{n}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\right)||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| \leq 2Rn\sqrt{p/n} = 2R\sqrt{pn}\end{aligned}$$

Finally, using condition (C8),

$$I_3 = 2n\text{sign}(\beta_j)P'_\lambda(|\beta_j|) = \text{sign}(\beta_j)\Theta(\sqrt{\max(p,k_n)n}).$$

We get that for all β_j such that $\beta_j^* = 0$, $\hat{\beta}_j$ will be estimated as 0 with probability tending to 1. ∎

part (c) (Asymptotic normality)

Let $\widehat{\boldsymbol{\beta}}$ be a sequence of local minima of $Q(\boldsymbol{\beta}|\boldsymbol{\Lambda})$ satisfying $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_P(\sqrt{p/n})$. The existence of such $\widehat{\boldsymbol{\beta}}$ is guaranteed by part (a).

Let $\boldsymbol{u} \in \mathbb{R}^s$ be some vector where $s = s_n$. We will show that

$$\boldsymbol{u}^T\boldsymbol{B}\check{\boldsymbol{\Upsilon}}^{-1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \xrightarrow{D} N(0, \boldsymbol{u}^T\boldsymbol{G}\boldsymbol{u})$$

and if $||\mathbf{\Lambda}_m - \mathbf{\Omega}_m|| \to 0$, then $\boldsymbol{u}^T \boldsymbol{B}\mathbf{\Upsilon}^{-1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \xrightarrow{D} N(0, \boldsymbol{u}^T\boldsymbol{G}\boldsymbol{u})$.

Notice that basic calculus implies that on the event $\{j;\ \hat{\beta}_j \neq 0\} = A$,

$$\widehat{\boldsymbol{\beta}}_A = (\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{Y} - n(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{P}_A, \tag{S1.1}$$

where $\mathbf{P} = \mathbf{P}_\lambda(\widehat{\boldsymbol{\beta}}) = (P'_\lambda(\hat{\beta}_1), ..., P'_\lambda(\hat{\beta}_p))^T$.

Since $\lambda_{max}(\check{\mathbf{\Upsilon}}^{-1/2}),\ \lambda_{max}(\mathbf{\Upsilon}^{-1/2}) = O_P(\sqrt{N})$, we have

$$n\boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{P}_A,\quad n\boldsymbol{u}^T\boldsymbol{B}\mathbf{\Upsilon}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{P}_A = o_P(1), \tag{S1.2}$$

where we have made use of (C2), (C4), and (C8).

Now consider

$$\boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{Y} = \boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}\boldsymbol{\beta}_A^* + \boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\boldsymbol{\epsilon}. \tag{S1.3}$$

We use the Lindeberg-Feller central limit theorem to show that

$$\boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\boldsymbol{\epsilon} = \sum_{i=1}^n w_{i,n},$$

with $w_{i,n} = \boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_{A,i}\mathbf{\Lambda}_m\boldsymbol{\epsilon}_i$, converges in distribution to a normal random variable. Fix $c > 0$ and let

$$\eta_{i,n} = \boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\mathbf{X}_{A,i}\mathbf{\Lambda}_m\mathbf{\Sigma}_m\mathbf{\Lambda}_m\mathbf{X}_{A,i}^T(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\check{\mathbf{\Upsilon}}^{-1/2}\boldsymbol{B}^T\boldsymbol{u}.$$

Then for $\delta = 2$,

$$\begin{aligned} E\left(w_{i,n}^2;\ w_{i,n}^2 > c\right) &\leq \left\{E\left(w_{i,n}^{2+\delta}\right)\right\}^{2/(2+\delta)} P\left(w_{i,n}^2 > c\right)^{\delta/(2+\delta)} \\ &\leq \frac{1}{c}\eta_{i,n}^{1+\delta/(2+\delta)}\left\{E\left(\boldsymbol{\epsilon}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\epsilon}_i\right)^{(2+\delta)/2}\right\}^{2/(2+\delta)}. \end{aligned}$$

Since $\max_{1\leq i\leq n}\eta_{i,n} \leq R^3 n||(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\check{\mathbf{\Upsilon}}^{-1/2}\boldsymbol{B}\boldsymbol{u}||^2 \max_{1\leq i\leq n}\frac{1}{n}||\mathbf{X}_i\mathbf{X}_i^T|| \to 0$, by (C7), it follows that

$$\begin{aligned} \sum_{i=1}^n E(w_{i,n}^2;\ w_{i,n}^2 > c) &\leq \frac{1}{c}\left\{E\left(\boldsymbol{\epsilon}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\epsilon}_1\right)^{(2+\delta)/2}\right\}^{2/(2+\delta)}\left(\sum_{i=1}^n\eta_{i,n}\right)\max_{1\leq i\leq n}\eta_{i,n} \\ &= \frac{1}{c}\left\{E\left(\boldsymbol{\epsilon}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\epsilon}_1\right)^{(2+\delta)/2}\right\}^{2/(2+\delta)}\boldsymbol{u}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{u}\max_{1\leq i\leq n}\eta_{i,n} \to 0 \end{aligned}$$

Thus, $\{w_{i,n}\}_{i=1}^n$ satisfy the Lindeberg condition and it follows that

$$\boldsymbol{u}^T\boldsymbol{B}\check{\mathbf{\Upsilon}}^{-1/2}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A)^{-1}\boldsymbol{x}_{A,i}\mathbf{\Lambda}_m\boldsymbol{\epsilon}_i \xrightarrow{D} N(0, \boldsymbol{u}^T\boldsymbol{G}\boldsymbol{u}).$$

Combining this with (S1.1)-(S1.3) proves the theorem. ∎

S1.3 Proof of Lemma 1

Part (a) (Rate of covariance matrix estimation under $\widehat{\boldsymbol{\beta}}$)

We have $\hat{\boldsymbol{\Sigma}}_m = \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_3$, where

$$\begin{aligned}
\mathbf{K}_1 &= \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}^*)(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}^*)^T = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T \\
\mathbf{K}_2 &= \frac{1}{n}\sum_{i=1}^n \left[\boldsymbol{\epsilon}_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\mathbf{X}_i^T + \mathbf{X}_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\boldsymbol{\epsilon}_i^T\right] \\
\mathbf{K}_3 &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T\mathbf{X}_i^T.
\end{aligned}$$

We bound $||\mathbf{K}_1 - \boldsymbol{\Sigma}_m||_F$, $||\mathbf{K}_2||_F$, and $||\mathbf{K}_3||_F$. Since

$$\begin{aligned}
P\left\{\left|\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T - \boldsymbol{\Sigma}_m\right|\right|_F^2 > c\right\} &\leq \frac{1}{c}E\left|\left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T - \boldsymbol{\Sigma}_m\right|\right|_F^2 \\
&= \frac{1}{c}\sum_{j,k=1}^m E\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_{ij}\epsilon_{ik} - \sigma_{jk}\right\}^2 = O(m^2/n)
\end{aligned}$$

it follows that $||\mathbf{K}_1 - \boldsymbol{\Sigma}_m||_F = O_P(\sqrt{m^2/n})$. Notice that

$$\begin{aligned}
||\mathbf{K}_3||_F &\leq \frac{1}{n}\sum_{i=1}^n ||\mathbf{X}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\mathbf{X}_i^T||_F = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i\right)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&\leq \left(\frac{1}{n}\sum_{i=1}^n ||\boldsymbol{x}_i||^2\right)||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||^2 = O(R)O_P(p/n)
\end{aligned}$$

and

$$\begin{aligned}
||\mathbf{K}_2||_F &\leq \frac{2}{n}\sum_{i=1}^n ||\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\mathbf{X}_i^T||_F = \frac{2}{n}\sum_{i=1}^n \mathrm{tr}\left(\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\mathbf{X}_i^T\mathbf{X}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right)^{1/2} \\
&\leq \frac{2||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||}{n}\sum_{i=1}^n \lambda_{\max}(\mathbf{X}_i^T\mathbf{X}_i)^{1/2}||\boldsymbol{\epsilon}_i|| = \frac{2||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||}{n}\sum_{i=1}^n ||\boldsymbol{x}_i||||\boldsymbol{\epsilon}_i|| \\
&\leq 2||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||\left(\frac{1}{n}\sum_{i=1}^n ||\boldsymbol{\epsilon}_i||^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n ||\boldsymbol{x}_i||^2\right)^{1/2} = O_P(\sqrt{pm/n}).
\end{aligned}$$

The first transition is from the triangle inequality. The one before last is Cauchy-Schwartz inequality, and the last equality uses the rate of consistency of the $\widehat{\boldsymbol{\beta}}$ estimate and condition (C4). The result follows from combining the bounds on $||\mathbf{K}_1 - \boldsymbol{\Sigma}_m||_F, \mathbf{K}_2$ and $\mathbf{K}_3$. ∎

Part (b) (Consistency)

Assuming that $\hat{\mathbf{\Sigma}}_m = \mathbf{\Sigma}_m + O_p\left(\sqrt{(m+p)m/n}\right)$ we will show that the estimator of $\mathbf{\Omega}$, that minimizes $Q(\mathbf{\Omega}|\widehat{\boldsymbol{\beta}})$ (where $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| = O_p(\sqrt{p/n})$) is consistent and sparsistent.

Define $\mathbf{U}$ a symmetric matrix of size m, $\alpha_n = \sqrt{(m+p)m/n}$, and $\mathbf{\Delta}_n = \alpha_n \mathbf{U}$. Let M be a constant.

$$P\left(\inf_{||\mathbf{U}||_F = M} Q(\mathbf{\Omega} + \mathbf{\Delta}_u|\widehat{\boldsymbol{\beta}}) > Q(\mathbf{\Omega}|\widehat{\boldsymbol{\beta}})\right) \to 1$$

For sufficiently large constant M. This will imply that there is a local minimizer in the ball of radius $M\alpha_n$ around $\mathbf{\Omega}$.

Let

$$\begin{aligned}
Q(\mathbf{\Omega}_m + \mathbf{\Delta}_u|\widehat{\boldsymbol{\beta}}) - Q(\mathbf{\Omega}_m|\widehat{\boldsymbol{\beta}}) &= \operatorname{tr}(\hat{\mathbf{\Sigma}}_m(\mathbf{\Omega}_m + \mathbf{\Delta}_u)) - \operatorname{tr}(\hat{\mathbf{\Sigma}}_m\mathbf{\Omega}_m) + \log\det(\mathbf{\Omega}_m) \\
&\quad -\log\det(\mathbf{\Omega}_m + \mathbf{\Delta}_u) + \sum_{i\le j}(P_\gamma(|\omega_{ij} + \delta_{ij}|) - P_\gamma(|\omega_{ij}|)) \\
&= \operatorname{tr}((\hat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)\mathbf{\Delta}_u) + \operatorname{vec}(\mathbf{\Delta}_u)^T\left\{\int_0^1 g(v, \mathbf{\Omega}_v)(1-v)dv\right\}\operatorname{vec}(\mathbf{\Delta}_u) \\
&\quad + \sum_{i\le j}(P_\gamma(|\omega_{ij} + \delta_{ij}|) - P_\gamma(|\omega_{ij}|)) \\
&= L_1 + L_2 + L_3
\end{aligned}$$

Where we used the taylor expansion of $\log\det(\mathbf{\Omega}_m + \mathbf{\Delta}_u)$ around $\mathbf{\Omega}_m$ with the integral form of the remainder (see Rothman et al (2008)), and $\mathbf{\Omega}_v = \mathbf{\Omega}_m + v\mathbf{\Delta}_u$, and $g(v, \mathbf{\Omega}_v) = \mathbf{\Omega}_v^{-1} \otimes \mathbf{\Omega}_v^{-1}$. Examining the various terms:

$$\begin{aligned}
L_1 = \operatorname{tr}\left((\hat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)\mathbf{\Delta}_u\right) &\le \operatorname{tr}\left((\hat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)^T(\hat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)\right)^{1/2} \operatorname{tr}\left(\mathbf{\Delta}_u^T\mathbf{\Delta}_u\right)^{1/2} \\
&= ||\hat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m||_F||\mathbf{\Delta}_u||_F = O_p\left(\sqrt{(m+p)m/n}\right)O(\alpha_n) \\
&= O_P((m+p)m/n)
\end{aligned}$$

$$\begin{aligned}
\operatorname{vec}(\mathbf{U})^T\left\{\int_0^1 g(v, \mathbf{\Omega}_v)(1-v)dv\right\}\operatorname{vec}(\mathbf{U}) &\ge M^2/2 \min_{0\le v\le 1}\lambda_{\min}\left(\mathbf{\Omega}_v^{-1} \otimes \mathbf{\Omega}_v^{-1}\right) \\
&\ge M^2/2 \min_{0\le v\le 1}\lambda_{\min}(\mathbf{\Omega}_v^{-2}) \\
&\ge M^2/2\left(R + \alpha_n M\right)^{-2}
\end{aligned}$$

and this term is positive, so that

$$L_2 = \alpha_n^2 \operatorname{vec}(\mathbf{U})^T\left\{\int_0^1 g(v, \mathbf{\Omega}_v)(1-v)dv\right\}\operatorname{vec}(\mathbf{U}) = \Theta((m+p)m/n).$$

Note that in the last inequality we used the fact that $\lambda_{\max}(\mathbf{A}) = \lambda_{\min}(\mathbf{A}^{-1})$ for a positive definite matrix $\mathbf{A}$, and $\lambda_{\max}(A) \leq ||A||_F$.

Finally, for L_3:

$$\begin{aligned}\sum_{i\leq j}(P_\gamma(|\omega_{ij}+\delta_{ij}|)-P_\gamma(|\omega_{ij}|)) &= \sum_{i\leq j, w_{ij}\neq 0}(P_\gamma(|\omega_{ij}+\delta_{ij}|)-P_\gamma(|\omega_{ij}|)) \\ &\quad + \sum_{i\leq j, w_{ij}=0}(P_\gamma(|\omega_{ij}+\delta_{ij}|)\end{aligned}$$

Where the right term satisfies $\sum_{i\leq j, w_{ij}=0}(P_\gamma(|\omega_{ij}+\delta_{ij}|) > 0$. The left term is bounded by:

$$\begin{aligned}-\sum_{i\leq j, w_{ij}\neq 0} P'_\gamma(|w_{ij}+\delta_{ij}|)|\delta_{ij}| &\geq -\alpha_n P'_\gamma(\tau/2)\sum_{i\leq j, w_{ij}\neq 0}|\delta_{ij}| \\ &\geq -\sqrt{t_n}\alpha_n P'_\gamma(\tau/2) \geq -m\alpha_n P'_\gamma(\tau/2)\end{aligned}$$

Where t_n is the number of non-zero elements in $\mathbf{\Omega}$ and we used the relation between the L_1 and L_2 norms. From condition (C11), it follows that L_3, as well as L_1, is dominated by $L2$, and It follows that M may be chosen large enough so that the required probability is close to 1, whenever n is sufficiently large. The result follows. ∎

Part (C) (Uniform sparsity)

We examine the derivative of $Q(\mathbf{\Omega}|\hat{\mathbf{\Sigma}})$ at $\omega_{ij} = 0$, and show that it is dominated by the sign of the penalty term for all such terms. That will imply that the likelihood is maximized when $\hat{\omega}_{ij} = 0$.

$$\frac{\partial Q(\hat{\mathbf{\Omega}}|\hat{\mathbf{\Sigma}})}{\omega_{ij}} = 2(\hat{\sigma}_{ij} - \sigma_{ij} + P'_\gamma(|\hat{\omega}_{ij}|)\text{sign}(\hat{\omega}_{ij}))$$

From part (a), we have that $||\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}||_F = O_p(\sqrt{(m+p)m/n})$. Therefore, $\max_{\omega_{ij}=0}(\hat{\sigma}_{ij}-\sigma_{ij}) = O_p(\sqrt{(m+p)m/n})$. At the same time, since $\omega_{ij} = 0$, and $||\hat{\mathbf{\Omega}}-\mathbf{\Omega}||_F^2 = O_P((m+p)m/n)$, $\hat{\omega}_{ij} = O_p(\sqrt{(m+p)m/n})$. From condition (C12), the result hold. ∎

S1.4 Proof of Theorem 2

Preliminaries

First, we note that from Lemma 4, if $||\mathbf{\Lambda}_m - \mathbf{\Omega}_m|| \to 0$ then $\boldsymbol{B}\boldsymbol{\Upsilon}^{-1/2}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*_A) \xrightarrow{D} N(0, \boldsymbol{G})$, where $\boldsymbol{\Upsilon} = (\mathbf{X}_A^T\mathbf{\Omega}\mathbf{X}_A)^{-1}$ and $\boldsymbol{B}$ is a sequence defined in Theorem 1.

Proof of the theorem

At the first stage of Algorithm 1, a regularized estimator, or MLE (if feasible by sample size) for $\boldsymbol{\beta}$, namely $\widehat{\boldsymbol{\beta}}^{(1)}$ is found under working independence assumption. According to Theorems 1, under conditions (C1)-(C8), this estimator is $\sqrt{p/n}$ consistent. Then, an estimator of $\boldsymbol{\Omega}$, namely $\hat{\boldsymbol{\Omega}}^{(1)}$ is found by minimizing $Q(\boldsymbol{\Omega}|\widehat{\boldsymbol{\beta}}^{(1)})$. According to Lemma 1, if conditions (C9)-(C12) also hold, this estimator is $\sqrt{(m+p)m/n}$ consistent. Next, an estimator $\widehat{\boldsymbol{\beta}}^{(2)}$ is found be minimizing $Q(\boldsymbol{\beta}|\hat{\boldsymbol{\Omega}}^{(1)})$, and $\hat{\boldsymbol{\Omega}}^{(2)}$ is estimated by minimizing $Q(\boldsymbol{\Omega}|\boldsymbol{\beta}^{(2)})$.

(i) If conditions (C1)-(C8) hold, according to Theorem 1 $\widehat{\boldsymbol{\beta}}^{(2)}$ is consistent, sparsistent, and asymptotically normal.

(ii) If in addition conditions (C9)-(C11) hold, then according to Theorem 1 and Lemma 4, this estimator is consistent, sparsistent, asymptotically normally distributed and efficient. By Lemma 1 $\hat{\boldsymbol{\Omega}}^{(2)}$ is consistent and sparsistent. ∎

S2 Consistency of the BIC

Before proceeding, note that a penalized estimator of $\boldsymbol{\beta}$ with working/estimated precision matrix $\widehat{\boldsymbol{\Omega}}$ and with non-zero entries on the set $\hat{A}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\hat{A}} = (\mathbf{X}_{\hat{A}}^T\widehat{\boldsymbol{\Omega}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^T\widehat{\boldsymbol{\Omega}}\mathbf{Y} - n(\mathbf{X}_{\hat{A}}^T\widehat{\boldsymbol{\Omega}}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}})$$

S2.1 Proof of Theorem 3

part (a) (fixed precision matrix)

We first show that the probability of selection of any overfitting model (a model with at least one "false positive" selection) tends to 0. $\underline{A \subset \hat{A}}$:

Consider

$$\begin{aligned}\mathrm{BIC}(\hat{\boldsymbol{\beta}}_A) - \mathrm{BIC}(\hat{\boldsymbol{\beta}}_{\hat{A}}) &= \frac{1}{n}\Big[(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_A)^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_A) - (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A}})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A}}) \\ &\quad + k_n(s-\hat{s})\Big]\end{aligned}$$

the penalized estimator satisfies $\widehat{\boldsymbol{\beta}}_{\hat{A}} = (\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{Y} + n(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}})$. Therefore by using the decomposition $\mathbf{Y} = \mathbf{X}_A\boldsymbol{\beta}_A + \boldsymbol{\epsilon}$ and since

$$\{\mathbf{I} - \boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}^{1/2}\}$$

is a projection matrix on the null space of $\mathbf{X}_{\hat{A}}$, it is easy to verify that

$$\begin{aligned}&(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A}})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A}}) = \\ &= \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{1/2}\{\mathbf{I} - \boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}^{1/2}\}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon} + n^2P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}})^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}}) \\ &= (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML}) + n^2P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}})^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda}(\widehat{\boldsymbol{\beta}}_{\hat{A}})\end{aligned}$$

$\underline{\text{Claim: the terms involving the penalty are negligible.}}$ Using conditions (C2) and Lemma 5 for bounding the eigenvalues of $\left(n(\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}\right)$ and $\left(n(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\right)$, condition (C7) to bound the penalty on term belonging to the true model, and condition (C8) for false positive terms, we get that for the true model:

$$\begin{aligned}n^2P'_{\lambda_A}(\widehat{\boldsymbol{\beta}}_A)^T(\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}P'_{\lambda_A}(\widehat{\boldsymbol{\beta}}_A) &\leq n||P'_{\lambda_A}(\widehat{\boldsymbol{\beta}}_A)||^2\lambda_1\left(n(\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}\right) \\ &\leq Rnso_p(\frac{1}{np}) = o_p(1)\end{aligned}$$

while for the overfitted model:

$$n^2 P'_{\lambda_{\hat{A}}}(\widehat{\boldsymbol{\beta}}_{\hat{A}})^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda_{\hat{A}}}(\widehat{\boldsymbol{\beta}}_{\hat{A}}) \geq n||P'_{\lambda_{\hat{A}}}(\widehat{\boldsymbol{\beta}}_{\hat{A}})||^2\lambda_{\min}\left(n(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\right)$$
$$\geq R^{-1}n\Theta\left(\sqrt{\frac{\max(p,k_n)}{n}}\right) = \Theta\left(\sqrt{n\max(p,k_n)}\right).$$

Therefore, asymptotically, the difference between the BIC's satisfies:

$$n\Big(\mathrm{BIC}(\hat{\boldsymbol{\beta}}_A) - \mathrm{BIC}(\hat{\boldsymbol{\beta}}_{\hat{A}})\Big) \leq (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML})$$
$$-(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML}) + k_n(s-\hat{s})$$

Denote the projection of a vector $\boldsymbol{x}$ on the column space of a matrix $\mathbf{A}$ by $P_{\mathbf{A}}(\boldsymbol{x})$, and $P_{\mathbf{A}}^{\perp}(\boldsymbol{x})$ the projection on the sparse orthogonal to the column space of $\mathbf{A}$. One can see that

$$\begin{aligned}
(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML}) &= \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{1/2}\{\mathbf{I}-\boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}^{1/2}\}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon} \\
&= ||P^{\perp}_{\left[\boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\hat{A}}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2, \\
(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML}) &= \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{1/2}\{\mathbf{I}-\boldsymbol{\Lambda}^{1/2}\mathbf{X}_{A}(\mathbf{X}_{A}^T\boldsymbol{\Lambda}\mathbf{X}_{A})^{-1}\mathbf{X}_{A}^T\boldsymbol{\Lambda}^{1/2}\}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon} \\
&= ||P^{\perp}_{\left[\boldsymbol{\Lambda}^{1/2}\mathbf{X}_{A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2 \\
&= ||P^{\perp}_{\left[\boldsymbol{\Lambda}^{1/2}\mathbf{X}_{\hat{A}}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2 + ||P_{\left[\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2
\end{aligned}$$

Where $\tilde{\mathbf{X}}_{\hat{A}\setminus A} = P^{\perp}_{[\Lambda^{1/2}\mathbf{X}_A]}(\mathbf{X}_{\hat{A}})$, a matrix of rank $\hat{s}-s$. Therefore,

$$(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML}) - (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})$$
$$= ||P_{\left[\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2 \qquad \text{(S2.1)}$$

where

$$\begin{aligned}
||P_{\left[\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2 &= \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}^{1/2}\{\boldsymbol{\Lambda}^{1/2}\tilde{\mathbf{X}}_{\hat{A}\setminus A}(\tilde{\mathbf{X}}_{\hat{A}\setminus A}^T\boldsymbol{\Lambda}\tilde{\mathbf{X}}_{\hat{A}\setminus A})^{-1}\tilde{\mathbf{X}}_{\hat{A}\setminus A}^T\boldsymbol{\Lambda}^{1/2}\}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon} \\
&= n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\left(n^{-1}\tilde{\mathbf{X}}_{\hat{A}\setminus A}^T\boldsymbol{\Lambda}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right)^{-1}\tilde{\mathbf{X}}_{\hat{A}\setminus A}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon} \\
&\leq Rn||\frac{1}{n}\tilde{\mathbf{X}}_{\hat{A}\setminus A}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}||^2
\end{aligned}$$

Where we used condition (C2) and Lemma 5 for the inequality. According to Lemma 3, we have the uniform bound:

$$P\left(\sup_{A\subset\hat{A}}\frac{1}{\hat{s}-s}||P_{\left[\boldsymbol{\Lambda}^{1/2}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2 \geq k_n\right) \to 0$$

Therefore, using (S2.1) and (S2.2):

$$\begin{aligned} & P\left(\sup_{A\subset\hat{A}}\frac{n}{\hat{s}-s}\left(\mathrm{BIC}(\hat{\boldsymbol{\beta}}_A)-\mathrm{BIC}(\hat{\boldsymbol{\beta}}_{\hat{A}})\right)>0\right) \\ \leq\ & P\Bigg(\sup_{A\subset\hat{A}}\frac{n}{\hat{s}-s}\Big((\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{A,ML}) \\ & \qquad -(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\hat{A},ML})+k_n\Big)>0\Bigg) \\ & \leq P\left(\sup_{A\subset\hat{A}}\frac{1}{\hat{s}-s}||P_{\left[\boldsymbol{\Lambda}^{1/2}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\right]}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2\geq k_n\right)\to 0 \end{aligned}$$

The probability of selection of an underfitting model tends to 0 (Here "underfitting" model is a model with at least one true parameter set as zero, i.e. false negative). $A\setminus\hat{A}\neq\emptyset$: First, as before, we claim that the penalty terms are negligible. It was shown in the previous part,

$$n^2P'_{\lambda_A}(\widehat{\boldsymbol{\beta}}_A)^T(\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}P'_{\lambda_A}(\widehat{\boldsymbol{\beta}}_A)=o_p(1) \tag{S2.2}$$

and therefore could be seen to be dominated by other terms in the BIC. Also:

$$n^2P'_{\lambda_{\hat{A}}}(\widehat{\boldsymbol{\beta}}_{\hat{A}})^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_{\lambda_{\hat{A}}}(\widehat{\boldsymbol{\beta}}_{\hat{A}})>0. \tag{S2.3}$$

Denote the sub-vector of a vector $\boldsymbol{u}$ corresponding to the indices in a set B by $\boldsymbol{u}_{[B]}$. Write $\mathbf{Y}=\mathbf{X}_A\boldsymbol{\beta}^*_A+\boldsymbol{\epsilon}=\mathbf{X}_{A\cap\hat{A}}\boldsymbol{\beta}^*_{A[A\cap\hat{A}]}+\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}+\boldsymbol{\epsilon}=\mathbf{X}_{\hat{A}}\boldsymbol{\beta}^*_{A[\hat{A}]}+\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}+\boldsymbol{\epsilon}$ (since the β^*_j for $j\in\hat{A}\setminus A$ is zero). Then:

$$\begin{aligned} & (\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML})= \\ & =(\mathbf{X}_A\boldsymbol{\beta}^*_A+\boldsymbol{\epsilon}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{X}_A\boldsymbol{\beta}^*_A+\boldsymbol{\epsilon}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML}) \\ & \quad = \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}+(\boldsymbol{\beta}^*_A-\widehat{\boldsymbol{\beta}}_{A,ML})^T\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A(\boldsymbol{\beta}^*_A-\widehat{\boldsymbol{\beta}}_{A,ML})+2(\boldsymbol{\beta}^*_A-\widehat{\boldsymbol{\beta}}_{A,ML})^T\mathbf{X}_A^T\boldsymbol{\Lambda}\boldsymbol{\epsilon} \end{aligned}$$

Similarly:

$$\begin{aligned} & (\mathbf{X}_{\hat{A}}\boldsymbol{\beta}^*_{A[\hat{A}]}+\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}+\boldsymbol{\epsilon}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{X}_{\hat{A}}\boldsymbol{\beta}^*_{A[\hat{A}]}+\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}+\boldsymbol{\epsilon}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML}) \\ & \quad = \boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}+(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML}) \\ & \quad +\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}_{A\setminus\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}+2(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]} \\ & \quad +2(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}+2\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}_{A\setminus\hat{A}}^T\boldsymbol{\Lambda}\boldsymbol{\epsilon} \end{aligned}$$

from the consistency and identifiability assumptions, condition (C2) and lemma 5, the following hold:

$$n(\boldsymbol{\beta}^*_A-\widehat{\boldsymbol{\beta}}_{A,ML})^T(n^{-1}\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A)(\boldsymbol{\beta}^*_A-\widehat{\boldsymbol{\beta}}_{A,ML}) \quad = \quad O_p(np/n)=O_p(p)$$

By Lemma 2 part 2, $2(\boldsymbol{\beta}^*_A - \widehat{\boldsymbol{\beta}}_{A,ML})^T \mathbf{X}_A^T \boldsymbol{\Lambda}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^T \boldsymbol{\Lambda}\mathbf{X}_A \left(\mathbf{X}_A^T \boldsymbol{\Lambda}\mathbf{X}_A\right)^{-1} \mathbf{X}_A^T \boldsymbol{\Lambda}\boldsymbol{\epsilon} = O_p(p)$.

$$\inf_{\hat{A}\subset A}\left(n\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}(n^{-1}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}})\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}\right)$$

$$\geq \inf_{\hat{A}\subset A}\left(n||\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}]||^2\lambda_{min}\left(n^{-1}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\right)\right) \tag{S2.4}$$

$$\geq nR^2|A\setminus\hat{A}|\min_{j\in A\setminus\hat{A}}\beta_j^2 = R^2|A\setminus\hat{A}|\Theta(\max(p,k_n)) \tag{S2.5}$$

We will now show that the "cross product" terms are dominated by the other terms, uniformly over all models with at least one false negative variable. We first show it for $2(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}$, than for $2(\boldsymbol{\beta}^*_{A[\hat{A}]}-\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}+2\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}$.

From Lemma 6, $2(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}$ is dominated by

$$\max\left(n\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}(n^{-1}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}})\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]},\ n(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T(n^{-1}\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})\right).$$

From Lemma 7, we have that

$$2(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon} + 2\boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon} \leq$$
$$p^{-1/4}\Big[(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})$$
$$+2(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]} + \boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}\Big]$$

with probability tending to 1 uniformly over all sets $\hat{A}$ such that $A\setminus\hat{A}\neq\emptyset$.

Finally, we use the results on the penalty terms (S2.2) and (S2.3) in combining the above

results:

$$\begin{aligned}
\sup_{\hat{A}\subset A}\Big(BIC(\widehat{\boldsymbol{\beta}}_A) - BIC(\widehat{\boldsymbol{\beta}}_{\hat{A}})\Big) &= \sup_{\hat{A}\subset A}\Big(\frac{1}{n}\{(\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML}) \\
&\quad +n^2P'_\lambda(\widehat{\boldsymbol{\beta}}_A)^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_A)^{-1}P'_\lambda(\widehat{\boldsymbol{\beta}}_A) \\
&\quad -(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML}) \\
&\quad -n^2P'_\lambda(\widehat{\boldsymbol{\beta}}_{\hat{A}})^T(\mathbf{X}_{\hat{A}}^T\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}})^{-1}P'_\lambda(\widehat{\boldsymbol{\beta}}_{\hat{A}}) + k_n(s-\hat{s})\}\Big) \\
&\leq \sup_{\hat{A}\subset A}\Big(\frac{1}{n}\{(\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_A\widehat{\boldsymbol{\beta}}_{A,ML}) + o_p(1) \\
&\quad -(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\boldsymbol{\Lambda}(\mathbf{Y}-\mathbf{X}_{\hat{A}}\widehat{\boldsymbol{\beta}}_{\hat{A},ML}) + k_n|A\backslash\hat{A}|\}\Big) \\
&= \frac{1}{n}\Big\{O_p(p) + \inf_{\hat{A}\subset A}\Big(k_n|A\backslash\hat{A}| - n\boldsymbol{\beta}^{*T}_{A\backslash\hat{A}}(n^{-1}\mathbf{X}^T_{A\backslash\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\backslash\hat{A}})\boldsymbol{\beta}^*_{A\backslash\hat{A}}\Big)\Big\} \\
&= \frac{1}{n}\Big\{O_p(p) + |A\backslash\hat{A}|\Big[k_n - R^2\Theta(\max(p,k_n))\Big]\Big\}
\end{aligned}$$

thus, as $n\to\infty$ $BIC(\widehat{\boldsymbol{\beta}}_A) - BIC(\widehat{\boldsymbol{\beta}}_{\hat{A}}) < 0$ uniformly for every underfitted model $\hat{A}$. ∎

part (b) (estimated precision matrix)

Let $\boldsymbol{\Lambda}$ be an estimator of $\boldsymbol{\Omega}$, based on some initial $\sqrt{p/n}$-consistent estimator of $\boldsymbol{\beta}$. Then, by Lemma 1, we have that $||\boldsymbol{\Lambda}-\boldsymbol{\Omega}||_F = O_p(\sqrt{m(m+p)/n})$.

Let $\boldsymbol{\Lambda} = \widehat{\boldsymbol{\Omega}}$ be a $\sqrt{m(m+p)/n}$-consistent estimator of $\boldsymbol{\Omega}$. We want to show that with probability tending to 1 the eigenvalues of $\boldsymbol{\Lambda}$ are uniformly bounded. It is given that $||\boldsymbol{\Lambda}-\boldsymbol{\Omega}||_F = ||\boldsymbol{\Delta}||_F = O_p(\sqrt{m(m+p)/n})$. Since the operator norm of a matrix is bounded by its Frobenius norm, we have that $\lambda_1(\boldsymbol{\Delta}) = O_p(\sqrt{m(m+p)/n})$. Invoking the result from matrix perturbation theory (Stewart and Sun (1990)): if $\mathbf{A}$ is an $m\times m$ squared positive definite matrix and $\mathbf{E}$ is a perturbation matrix, then for all $k=1,\ldots,m$

$$\lambda_k(\mathbf{A}) + \lambda_1(\mathbf{E}) \geq \lambda_k(\mathbf{A}+\mathbf{E}) \geq \lambda_k(\mathbf{A}) + \lambda_m(\mathbf{E}). \tag{S2.6}$$

and take $\boldsymbol{\Omega}$ to be the positive definite matrix, and $\boldsymbol{\Delta}$ to be the perturbation matrix. With probability tending to 1, there exists an $N\in\mathbb{N}$ so that for all $n>N$, $\sqrt{m(m+p)/n} < R^{-1}/2$. Then, for $n>N$:

$$\begin{aligned}
\lambda_1(\boldsymbol{\Lambda}) \leq \lambda_1(\boldsymbol{\Omega}) + \lambda_1(\boldsymbol{\Delta}) &= R + O_p(\sqrt{m(m+p)/n}) \leq R + R^{-1}/2 \text{ and,} \\
\lambda_m(\boldsymbol{\Lambda}) \geq \lambda_m(\boldsymbol{\Omega}) - \lambda_1(\boldsymbol{\Delta}) &= R - O_p(\sqrt{m(m+p)/n}) \geq R^{-1} - R^{-1}/2 = R^{-1}/2
\end{aligned}$$

With probability tending to 1. Therefore, the BIC is valid and consistent for model selection with the consistent estimator $\boldsymbol{\Lambda}$. ∎

S3 Large p main results

S3.1 Modified technical conditions

(C1') Let $p > n$ at a sub-exponential rate, so that $\log(p)/n \to 0$. The number of outcomes is still smaller than the sample size $m < n$. In addition, the true model size $s = |A|$ satisfies $s < n$ and $\frac{s}{\log(p)} \to 0$.

(C2') The eigenvalues of the positive definite matrices $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, $\frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A$ satisfy

$$0 < R^{-1} < \lambda_{\min}\left(\frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A\right), \lambda_{\min}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right) < \lambda_{\max}\left(\frac{1}{n}\mathbf{X}_A^T\mathbf{X}_A\right), \lambda_{\max}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right) < R < \infty$$

(C3') Identifiability condition: $\min_{j:\beta_j \neq 0} \frac{\beta_j}{\sqrt{\log(p)/n}} \to \infty$

(C7') If r_n is such that $\lim_n \frac{r_n}{\sqrt{\log(p)/n}} = \infty$, then $n\sqrt{\log(p)/n}P'_\lambda(r_n) = o(1)$

(C8') If r_n is such that $\lim_n \frac{r_n}{\sqrt{\log(p)/n}} \leq c$, then $P'_\lambda(r_n)/m \to \infty$

(C10') The errors are normally distributed, i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$

S3.2 Proof of Theorem 4

Part (a) (consistency)

As in the proof of theorem 1, we decompose the difference between the penalized likelihood functions to $D(\boldsymbol{u}) = J_1 + J_2 + J_3$, where $||\boldsymbol{u}|| = M$ and $J_1(\boldsymbol{u}) = -2\sqrt{\log(p)/n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$, $J_2(\boldsymbol{u}) = (\log(p)/n)\boldsymbol{u}^T\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\boldsymbol{u}$ and $J_3(\boldsymbol{u}) = 2n\left[\sum_{j=1}^p P_\lambda(|\beta_j^* + \sqrt{\log(p)/n}u_j|) - P_\lambda(|\beta_j^*|)\right]$. Consider the presentation $\boldsymbol{u} = (\boldsymbol{u}_A^T, \boldsymbol{u}_{A^c}^T)^T, \mathbf{X} = (\mathbf{X}_A|\mathbf{X}_{A^c})$. Then:

$$\begin{aligned}
|J_1| &= 2\sqrt{\log(p)/n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}_A\boldsymbol{u}_A + 2\sqrt{\log(p)/n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}_{A^c}\boldsymbol{u}_{A^c} = J_{1A} + J_{1A^c} \\
J_2 &= \log(p)/n\Big(\boldsymbol{u}_A^T\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_A\boldsymbol{u}_A + \boldsymbol{u}_{A^c}^T\mathbf{X}_{A^c}^T\boldsymbol{\Lambda}\mathbf{X}_{A^c}\boldsymbol{u}_{A^c} \\
&\quad + \boldsymbol{u}_A^T\mathbf{X}_A^T\boldsymbol{\Lambda}\mathbf{X}_{A^c}\boldsymbol{u}_{A^c} + \boldsymbol{u}_{A^c}^T\mathbf{X}_{A^c}^T\boldsymbol{\Lambda}\mathbf{X}_A\boldsymbol{u}_A\Big) \\
J_3 &= 2n\left[\sum_{j\in A} P_\lambda(|\beta_j^* + \sqrt{\log(p)/n}u_j|) - P_\lambda(|\beta_j^*|)\right] + 2n\left[\sum_{j\in A^c} P_\lambda(|\sqrt{\log(p)/n}u_j|)\right] \\
&= J_{3A} + J_{3A^c}
\end{aligned}$$

Note that since $M = ||\boldsymbol{u}|| \leq ||\boldsymbol{u}_A|| + ||\boldsymbol{u}_{A^c}||$, at least one of $||\boldsymbol{u}_A||, ||\boldsymbol{u}_{A^c}|| \not\to 0$. Therefore, if $||\boldsymbol{u}_{A^c}|| \to 0$, there exists a bound B and N_B such that for $n > N_B$, $M \geq ||\boldsymbol{u}_A|| > B > 0$.

We first show that if $||\boldsymbol{u}_{A^c}||$ is bounded from below, then J_{3A^c} dominates the rest of the terms. We then show that if $||\boldsymbol{u}_{A^c}|| \to 0$, then J_2 dominates the sum $J_1 + J_2 + J_3$.

Suppose first that $||\boldsymbol{u}_{A^c}||$ is bounded. Let $x_n/\sqrt{\log(p)/n} \to \infty$, $\lim_{n\to\infty} |y_n/\sqrt{\log(p)/n}| \leq c$. Then

$$\begin{aligned}
|J_1| &\leq 2\sqrt{\log(p)/n}n(\frac{1}{N}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon})^{1/2}(\frac{1}{n}\boldsymbol{u}^T\mathbf{X}^T\boldsymbol{\Lambda}^2\mathbf{X}\boldsymbol{u})^{1/2} \leq O_p(\sqrt{mn\log(p)})||\boldsymbol{u}||R^3 \\
J_2 &\geq n\log(p)/n||\boldsymbol{u}||^2\lambda_{\min}(n^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}) \geq 0 \\
J_{3A} &= 2n\sqrt{\log(p)/n}\sum_{j\in A} u_j P'_\lambda(\beta^*_j + t_j\sqrt{\log(p)/n}u_j) \\
&\leq 2n\sqrt{\log(p)/n}\sqrt{s}||\boldsymbol{u}||_2 P'_\lambda(x_n) = 2\sqrt{s}Mo_p(1)
\end{aligned}$$

$$\begin{aligned}
J_{3A^c} &= 2n\sqrt{\log(p)/n}\sum_{j\notin A} |u_j|P'_\lambda(|t_j\sqrt{\log(p)/n}u_j|) \\
&\geq 2n\sqrt{\log(p)/n}\sum_{j\notin A} |u_j|P'_\lambda(|y_n|) = 2(\sum_{j\notin A}|u_j|)n\sqrt{\log(p)/n}\Theta_p(m) \\
&= 2||\boldsymbol{u}_{A^c}||_1\Theta_p(m\sqrt{n\log(p)})
\end{aligned}$$

and we can see that indeed the sum $J_1 + J_2 + J_3$ is positive with probability tending to 1. Suppose now that $||\boldsymbol{u}_{A^c}|| \to 0$ as $n \to \infty$. Then $||\boldsymbol{u}_A||$ is bounded. We use Lemma 2 and get:

$$\begin{aligned}
J_{1A} &\leq 2\sqrt{n\log(p)}||\frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\Lambda}\mathbf{X}_A||||\boldsymbol{u}_A|| \leq 2M\sqrt{n\log(p)}O_p(\sqrt{s/n}) = O_p(\sqrt{\log(p)s}) \\
J_{1A^c} &\leq 2\sqrt{n\log(p)}O_p(m)||\boldsymbol{u}_{A^c}||(\frac{1}{n}\boldsymbol{u}^T_{A^c}\mathbf{X}^T_{A^c}\boldsymbol{\Lambda}^2\mathbf{X}_{A^c}\boldsymbol{u}_{A^c})^{1/2} \\
&= O_p(m\sqrt{n\log(p)}||\boldsymbol{u}_{A^c}||_2)
\end{aligned}$$

Therefore, if $||\boldsymbol{u}_{A^c}|| \to 0$, but $||\boldsymbol{u}_{A^c}|| \neq 0$, we have that $J_{1A^c} = o_p(J_{3A^c})$. Further, if $||\boldsymbol{u}_{Ac}|| = 0$, then

$$J_2 \geq (\log(p)/n)\boldsymbol{u}^T_A\mathbf{X}^T_A\boldsymbol{\Lambda}\mathbf{X}_A\boldsymbol{u}_A \geq \log(p)||\boldsymbol{u}_A||^2\lambda_{\min}(\frac{1}{n}\mathbf{X}^T_A\boldsymbol{\Lambda}\mathbf{X}_A) = \Theta(\log(p))$$

and $J_1 = o_p(J_2)$. That proves part (a).

Part (b) (Uniform sparsity)

We proceed as in theorem 1 part (b), and show that if $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*|| < M\sqrt{\log(p)/n}$. then for each j such that $\beta^*_j = 0$, the derivative of the penalized likelihood is maximized at 0.

As in Theorem 1, we see that the derivative by β_j, where the true $\beta^* = 0$, uniformly for all $j = 1, \ldots, p$, satisfies:

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}Q(\boldsymbol{\beta}) &= -2\mathbf{X}^T_j\boldsymbol{\Lambda}\boldsymbol{\epsilon} + 2\mathbf{X}^T_j\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + 2n\text{sign}(\beta_j)P'_\lambda(|\beta_j|) \\
&\leq I_1 + I_2 + I_3
\end{aligned}$$

Consider I_1. Let $\boldsymbol{\epsilon}_i^{\Lambda} = \boldsymbol{\Lambda}\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda})$. Since $\mathbf{X}_i = \mathbf{I} \otimes \boldsymbol{x}_i$, there is some $k = 1, \ldots, m$ such that

$$|I_1| = 2\sqrt{n}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\epsilon_{ik}^{\Lambda}\right) \leq \sqrt{n}\max_{j=1,\ldots,p_0,k=1,\ldots,m}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\epsilon_{ik}^{\Lambda}\right)$$

According to Lyapunov CLT, for each pair of indices j, k we have that (using Lemma 5)

$$s_{jk}^2 = \sum_{i=1}^{n} x_{ij}^2 \mathrm{var}(\epsilon_{ik}^{\Lambda}) = \sigma_k^2 \sum_{i=1}^{n} x_{ij}^2 = n\sigma_k^2$$

then

$$\frac{1}{s_{jk}}\sum_{i=1}^{n} x_{ij}\epsilon_{ik}^{\Lambda} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\epsilon_{ik}^{\Lambda} \sim \mathcal{N}(0, \sigma_k^2).$$

Therefore, let $\max \sigma_1^2, \ldots, \sigma_m^2 < B$ (this bound exists, since we assume least $2+\delta$ bounded moments of the errors, and $\boldsymbol{\Lambda}$ has bounded eigenvalues), and denote by $u_{j,k}$ the j, k normal random variable with mean 0 and variance B. then a uniform bound for all $j = 1, \ldots, p$ is given by:

$$\sqrt{n}\max_{j=1,\ldots,p_0,k=1,\ldots,m}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\epsilon_{ik}^{\Lambda}\right) \leq \sqrt{n}\max_{j=1,\ldots,p_0,k=1,\ldots,m} u_{j,k} \leq B\sqrt{n}O_p(\sqrt{\log(p)})$$

The bound for the maximum of normal variables with diminishing correlation is given in Arias-Castro et al (2011), and note that correlations between the errors may only decrease this bound. Therefore, $I_1 < 2B\sqrt{n}O_p(\log(p))$ uniformly for all coordinates $j = 1, \ldots p$ of $\boldsymbol{\beta}$. Therefore:

$$\begin{aligned}
|I_1| &\leq \max_{j=1,\ldots p}\left(-2\mathbf{X}_j^T\boldsymbol{\Lambda}\boldsymbol{\epsilon}\right) \leq B\sqrt{n}O_p(\sqrt{\log(p)}) \\
I_2 &\leq \max_{j=1,\ldots p}\left(2\mathbf{X}_j^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right) \leq 2||\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)|| \\
&\leq 2n\lambda_1\left(\frac{1}{n}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\right)||\boldsymbol{\beta}-\boldsymbol{\beta}^*|| = 2nRO(\sqrt{\log(p)/n}) \\
I_3 &= 2n\mathrm{sign}(\beta_j)P_\lambda'(|\beta_j|) = 2n\Theta(m\log(p)/\sqrt{n})\mathrm{sign}(\beta_j) = 2\Theta(m\log(p)\sqrt{n})\mathrm{sign}(\beta_j)
\end{aligned}$$

It is easy to see that $\hat{\beta}_j = 0$ maximizes the penalized likelihood for all $j \neq A$.

S3.3 Proof of corollary

Similar to Theorem 2.

S4 Secondary lemmas

Lemma 2: Let conditions (C1)-(C10) hold. Then

1. $||\frac{1}{n}\mathbf{X}^T\mathbf{\Lambda}\boldsymbol{\epsilon}|| = O_p(\sqrt{p/n})$.
2. $\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\boldsymbol{\epsilon} = O_p(p)$ for an indices set $A \subset \{1,\ldots,p\}$.

Proof: 1. Let $Q = \boldsymbol{\epsilon}^T\mathbf{\Lambda}(n^{-1}\mathbf{X}\mathbf{X}^T)\mathbf{\Lambda}\boldsymbol{\epsilon}$ be a quadratic form with the positive semidefinite matrix of rank p $\mathbf{\Lambda}(n^{-1}\mathbf{X}\mathbf{X}^T)\mathbf{\Lambda}$, and denote $\boldsymbol{\epsilon} \sim F(\mathbf{0},\mathbf{\Sigma})$ for some distribution F and a block diagonal covariance matrix $\mathbf{\Sigma}$. Then

$$Q = \boldsymbol{\epsilon}^T\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Lambda}(n^{-1}\mathbf{X}\mathbf{X}^T)\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon} = \boldsymbol{z}^T\mathbf{A}\boldsymbol{z} = \boldsymbol{z}^T\mathbf{U}\mathbf{D}\mathbf{U}^T\boldsymbol{z} = (\mathbf{U}\boldsymbol{z})^T\mathbf{D}(\mathbf{U}\boldsymbol{z})$$

where $\boldsymbol{z} \sim F(\mathbf{0},\mathbf{I})$ and $\mathbf{U}^T\mathbf{D}\mathbf{U}$ is the spectral decomposition of the positive semidefinite matrix $\mathbf{A}$. $\mathbf{U}\boldsymbol{z} \sim F(\mathbf{0},\mathbf{I})$ since $\mathbf{U}$ is orthogonal. Denote $\mathbf{U}\boldsymbol{z} = (u_1,\ldots,u_N)^T$, where $u_i \sim F(0,1)$ for $i = 1,\ldots,N$. Then $Q = \sum_{i=1}^N \lambda_i u_i^2$, where $\lambda_1,\ldots,\lambda_N$ are the eigenvalues of $\mathbf{\Sigma}^{1/2}\mathbf{\Lambda}(n^{-1}\mathbf{X}\mathbf{X}^T)\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}$ and only p of them are different than zero and are bounded positive numbers from (C2). Fix M and let $\epsilon = \sqrt{p/n}$. Since the fourth moments of $\boldsymbol{\epsilon}$ are bounded, there exists a B such that $E((u_i-1)^4) < B, i = 1,\ldots,N$. By Chebyshev's inequality:

$$P\left(\left|\frac{1}{n}Q - \frac{1}{n}\sum_{i=1}^N \lambda_i\right| \geq M\epsilon\right) \leq \frac{n^2\sum_{i=1}^N \lambda_i^2 E(u_i^2-1)}{p^2n^2M^2} \leq \frac{pR^2B}{p^2M^2} \to 0 \text{ as } n\to\infty$$

Therefore $\frac{1}{n}Q$ is of the same order of $\frac{1}{n}\sum_{i=1}^N \lambda_i = O_p(p/n)$. We conclude that there exists M large enough and $||\frac{1}{n}\mathbf{X}^T\mathbf{\Lambda}\boldsymbol{\epsilon}|| = \sqrt{Q/n} = O_p(\sqrt{p/n})$.
For 2., notice that

$$\begin{aligned}
\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}_A\left(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\mathbf{\Lambda}\boldsymbol{\epsilon} &\leq ||\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}_A||^2\lambda_1\left([\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A]^{-1}\right)\\
&= \frac{1}{n}||\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}_A||^2\lambda_1\left([\frac{1}{n}\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A]^{-1}\right)\\
&\leq n||\frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}_A||^2\lambda_1\left([\frac{1}{n}\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}]^{-1}\right)\\
&\leq nR^2||\frac{1}{n}\boldsymbol{\epsilon}^T\mathbf{\Lambda}\mathbf{X}||^2 = O_p(p) \qquad \blacksquare
\end{aligned}$$

Lemma 3 Let conditions (C1)-(C10) hold, and let k_n be such that $p/k_n^{\frac{2+\delta}{2}} \to 0$.

1. Let $\hat{A}$ be a set of indices such that $\hat{A} \supset A$, and let $\tilde{\mathbf{X}}_{\hat{A}\setminus A} = P^{\perp}_{[\mathbf{\Lambda}^{1/2}\mathbf{X}_A]}(\mathbf{X}_{\hat{A}})$, where $P^{\perp}_{[\mathbf{\Lambda}^{1/2}\mathbf{X}_A]}(\mathbf{X}_{\hat{A}})$ is the projection of $\hat{\mathbf{X}}_{\hat{A}}$ on $\mathrm{col}([\mathbf{\Lambda}^{1/2}\mathbf{X}_A])^{\perp}$. This is a matrix of rank

$(\hat{s}-s)$, where $\hat{s}=|\hat{A}|$ and $s=|A|$. Then

$$P\left(\sup_{A\subset\hat{A}}\frac{1}{\hat{s}-s}||P_{[\tilde{\mathbf{X}}_{\hat{A}\setminus A}]}(\mathbf{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2\leq k_n\right)\to 1$$

2.

$$P\left(\sup_{j=1,\ldots,p}\left|\frac{1}{\sqrt{n}}\boldsymbol{x}_j^T\mathbf{\Lambda}\boldsymbol{\epsilon}\right|<\sqrt{k_n}\right)\to 1$$

Proof: First, we claim that $||\tilde{\boldsymbol{x}}_j||_2=||\boldsymbol{x}_j-P_M(\boldsymbol{x}_j)||_2\leq||\boldsymbol{x}_j||_2$. This is straight forward from the Pythagorean theorem. Next, we bound

$$\begin{aligned}\frac{1}{n}||\mathbf{\Sigma}^{1/2}\mathbf{\Lambda}\tilde{\boldsymbol{x}}_j||_2^2 &= \frac{1}{n}\tilde{\boldsymbol{x}}_j^T\mathbf{\Lambda}\mathbf{\Sigma}\mathbf{\Lambda}\tilde{\boldsymbol{x}}_j\leq\frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{x}_j\lambda_1\left(\mathbf{\Lambda}^2\right)\lambda_1(\mathbf{\Sigma})\leq R^3\frac{1}{n}\mathrm{tr}(\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ &= R^3\lambda_1\left(\frac{1}{n}\boldsymbol{x}_j\boldsymbol{x}_j^T\right)\leq R^3\lambda_1\left(\frac{1}{n}\mathbf{X}\mathbf{X}^T\right)\leq R^4\end{aligned}$$

where for we used the eigenvalue bound from condition (C2) a few times, and Lemma 6 for the fifth transition. Presented differently, we have that

$$||\frac{1}{\sqrt{n}}\tilde{\boldsymbol{x}}_j^T\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}||^2\leq R^4.$$

In the following steps we will use Lemma 3 in Dicker et al (2012), that says that there exists a constant K such that $\sup_{||\boldsymbol{u}||=1}E\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right|^{2+\delta}<K$ for $\tilde{\boldsymbol{\epsilon}}$ iid with bounded $2+\delta$ moments. Let $\tilde{\boldsymbol{\epsilon}}=\mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$, so that $\mathrm{cov}(\tilde{\boldsymbol{\epsilon}})=\mathbf{I}$.

$$\begin{aligned}P\left(\sup_{A\subset\hat{A}}\frac{1}{\hat{s}-s}||P_{[\tilde{\mathbf{X}}_{\hat{A}\setminus A}]}(\mathbf{\Lambda}^{1/2}\boldsymbol{\epsilon})||^2\geq k_n\right) &\leq P\left(\sup_{A\subset\hat{A}}\frac{Rn}{\hat{s}-s}||\frac{1}{n}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\mathbf{\Lambda}\boldsymbol{\epsilon}||^2\geq k_n\right)\\ &= P\left(\sup_{A\subset\hat{A}}\frac{R}{\hat{s}-s}||\frac{1}{\sqrt{n}}\tilde{\mathbf{X}}_{\hat{A}\setminus A}\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}\tilde{\boldsymbol{\epsilon}}||^2\geq k_n\right)\\ &\leq P\left((\hat{s}-s)\sup_{j\in1,\ldots,p}\frac{R}{\hat{s}-s}\left|\frac{1}{\sqrt{n}}\tilde{\boldsymbol{x}}_j^T\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}\tilde{\boldsymbol{\epsilon}}\right|^2\geq k_n\right)\\ &\leq \sum_{j=1,\ldots p}P\left(\sup_{j\in1,\ldots,p}R\left|\frac{1}{\sqrt{n}}\tilde{\boldsymbol{x}}_j^T\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}\tilde{\boldsymbol{\epsilon}}\right|^2\geq k_n\right)\leq pP\left(R^5\sup_{\boldsymbol{u}}|\boldsymbol{u}\tilde{\boldsymbol{\epsilon}}|^2\geq k_n\right)\\ &= pP\left(R^{5(2+\delta)/2}\sup_{\boldsymbol{u}}|\boldsymbol{u}\tilde{\boldsymbol{\epsilon}}|^{2+\delta}\geq k_n^{\frac{2+\delta}{2}}\right)\leq pR^{5(2+\delta)/2}\frac{E\left(|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}|\right)}{k_n^{\frac{2+\delta}{2}}}\leq\frac{pR^{5(2+\delta)/2}K}{k_n^{\frac{2+\delta}{2}}}\to 0\end{aligned}$$

which proves part 1. Part 2 follows similarly:

$$P\left(\sup_{j=1,\ldots p}\left|\frac{1}{\sqrt{n}}\boldsymbol{x}_j^T\mathbf{\Lambda}\mathbf{\Sigma}^{1/2}\tilde{\boldsymbol{\epsilon}}\right|>\sqrt{k_n}\right)\leq pR^2P\left(\left|\boldsymbol{u}^T\boldsymbol{\epsilon}_k\right|>\sqrt{k_n}\right)=pR^2P\left(\left(\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right)^{2+\delta}>k_n^{\frac{2+\delta}{2}}\right)\to 0$$

∎

Lemma 4 Suppose that $||\mathbf{\Lambda}_m - \mathbf{\Omega}_m|| \to 0$ holds. Then

$$\boldsymbol{B}(\check{\mathbf{\Upsilon}}^{-1/2} - \mathbf{\Upsilon}^{-1/2})\check{\mathbf{\Upsilon}}(\check{\mathbf{\Upsilon}}^{-1/2} - \mathbf{\Upsilon})^{-1/2})\boldsymbol{B}^T \to 0.$$

Proof: From the above expression, it is clear that, in fact, it suffices to show

$$\frac{1}{\sqrt{n}}||\check{\mathbf{\Upsilon}}^{-1/2} - \mathbf{\Upsilon}^{-1/2}|| \to 0. \tag{S4.1}$$

Now, since $||\mathbf{\Lambda} - \mathbf{\Omega}|| \to 0$, (C2) and some basic matrix manipulations imply that

$$\left|\left|\frac{1}{n}\check{\mathbf{\Upsilon}}^{-1} - \frac{1}{n}\mathbf{\Upsilon}^{-1}\right|\right| \to 0.$$

The convergence (S4.1) and part (ii) of the theorem follow from Theorem VII.2.12 of (Bhatia, 1997). ∎

Lemma 5. Let $\mathbf{X}$ be a $n \times p$ matrix, let A be a subset of the indices $\{1, \ldots, p\}$, and let $\mathbf{X}_A$ be the matrix whose columns are the subset of columns of $\mathbf{X}$ corresponding to A. Then:

(i) $\sigma_{min}(\mathbf{X}) \leq \sigma_{min}(\mathbf{X}_A) \leq \sigma_{max}(\mathbf{X}_A) \leq \sigma_{max}(\mathbf{X})$

(ii) $\lambda_{min}(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}) \leq \lambda_{min}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A) \leq \lambda_{max}(\mathbf{X}_A^T\mathbf{\Lambda}\mathbf{X}_A) \leq \lambda_{max}(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X})$ where $\mathbf{\Lambda}$ is a symmetric positive definite matrix.

Where $\sigma_{max}, \sigma_{min}, \lambda_{max}, \lambda_{min}$ are the largest and smallest singular values of a matrix, and the largest and smallest eigenvalues of a matrix.

Proof.

1. By definition, $\sigma_{max}(\mathbf{X}_A) = \max_{||\boldsymbol{u}_A||=1} ||\mathbf{X}_A\boldsymbol{u}_A||_2$. Let $\boldsymbol{u}_A^*$ be the vector with the indices corresponding to A equal to $\boldsymbol{u}_A$ and the rest are zero. Then $\sigma_{max}(\mathbf{X}) = \max_{||\boldsymbol{u}||=1} ||\mathbf{X}\boldsymbol{u}||_2 \geq \max_{||\boldsymbol{u}_A^*||=1} ||\mathbf{X}\boldsymbol{u}_A^*||_2$. Similarly it can be shown for the minimum singular value.
2. From the Cholesky decomposition $\mathbf{\Lambda} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is a lower diagonal matrix. Then $\lambda_{max}(\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}) = \sigma_{max}^2(\boldsymbol{L}\mathbf{X})$. Since $(\boldsymbol{L}\mathbf{X}_A) = (\boldsymbol{L}\mathbf{X})_A$, the inequalities follow from part (i).

∎

Lemma 6. Let $\mathbf{\Lambda}$ be a positive definite symmetric matrix, $\boldsymbol{u}, \boldsymbol{v}$ vectors. Then

$$|\boldsymbol{u}^T\mathbf{\Lambda}\boldsymbol{v}| \leq \sqrt{\boldsymbol{u}^T\mathbf{\Lambda}\boldsymbol{u}}\sqrt{\boldsymbol{v}^T\mathbf{\Lambda}\boldsymbol{v}} \leq \frac{1}{2}\boldsymbol{u}^T\mathbf{\Lambda}\boldsymbol{u} + \frac{1}{2}\boldsymbol{v}^T\mathbf{\Lambda}\boldsymbol{v}$$

Proof. Notice that $f(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{y}$ is an inner product of $\boldsymbol{x}$ and $\boldsymbol{y}$. Therefore we can use the Cauchy-Schwartz inequality and the averages inequality to get:

$$|\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{v}| \leq \sqrt{\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{u}}\sqrt{\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}} = \sqrt{\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{u}\cdot\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v}} \leq \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Lambda}\boldsymbol{u} + \frac{1}{2}\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v} \quad \blacksquare$$

Lemma 7: Under the regularity conditions (C1)-(C10), and the identifiability condition in Theorem 3, the following hold:

$$P\Bigg(\sup_{\hat{A}:A\setminus\hat{A}\neq\emptyset}\Big|(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon} + \boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}\Big| > p^{-1/4}\Big[(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A},ML}) + 2(\boldsymbol{\beta}^*_{A[\hat{A}]} - \widehat{\boldsymbol{\beta}}_{\hat{A}})^T\mathbf{X}^T_{\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]} + \boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]}\mathbf{X}^T_{A\setminus\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}\Big]\Bigg) \to 0$$

Proof: To simplify, note that we want to show that for some vectors $\boldsymbol{u}_1, \boldsymbol{u}_2$ we have that $|\boldsymbol{u}_1^T\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon} + \boldsymbol{u}_2^T\boldsymbol{\Lambda}^{1/2}\boldsymbol{\epsilon}| > n^{1/4}||\boldsymbol{u}_1 + \boldsymbol{u}_2||^2$. We are going to bring the expression to a form easier to work with. First, let us stack the two terms together, by having $\mathbf{X}_A\boldsymbol{\beta}^*_{A[A\cup\hat{A}]}$ instead of the two $\mathbf{X}_{A\setminus\hat{A}}\boldsymbol{\beta}^*_{A[A\setminus\hat{A}]}, \mathbf{X}_{\hat{A}}\boldsymbol{\beta}^*_{A[\hat{A}]}$. We can do it by stacking $\mathbf{X}_{\hat{A}}$ on $\mathbf{X}_{A\setminus\hat{A}}$ as $\mathbf{X}_{\hat{A}\cup A}$, and similarly setting $\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} = (\boldsymbol{\beta}^{*T}_{A[\hat{A}]}, \boldsymbol{\beta}^{*T}_{A[A\setminus\hat{A}]})^T$. Denote $\widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0} = (\widehat{\boldsymbol{\beta}}^T_{\hat{A},ML}, \mathbf{0}^T_{|A\setminus\hat{A}|})^T$ the stacked vector of $\widehat{\boldsymbol{\beta}}_{\hat{A},ML}$ and zeros corresponding to $A\setminus\hat{A}$.

We want to show:

$$P\left(\sup_{\hat{A}:A\setminus\hat{A}\neq\emptyset}\Big|(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}\Big| > p^{-1/4}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})\right) \to 0.$$

First we bound the right term from below. We use Lemma 5 and condition (C2):

$$\begin{aligned}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}) &\geq \\ &\geq n||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\lambda_m\left(\frac{1}{n}\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}\right) \\ &\geq n||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\lambda_m\left(\frac{1}{n}\mathbf{X}^T\boldsymbol{\Lambda}\mathbf{X}\right) \\ &\geq nR^{-2}||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\end{aligned}$$

and note that, from the identifiability condition in Theorem 3 $\left(\inf_{j\in A}|\beta_j^*| = \Theta(\sqrt{\max(k_n,p)/n}\right)$, we have that

$$\left(\inf_{\hat{A}:A\setminus\hat{A}\neq\emptyset}||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||\right) = \Theta(\sqrt{\max(k_n,p)/n}).$$

Using similar arguments to before:

$$\begin{aligned}
||(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{1/2}||^2 \quad &= \\
= (\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}) & \\
\leq n||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\lambda_1\left(\frac{1}{n}\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}\right) & \\
\leq n||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\lambda_1\left(\frac{1}{n}\mathbf{X}^T\boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\mathbf{X}\right) & \\
\leq nR^4||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2 &
\end{aligned}$$

Consider the required probability. Let $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ be a vector of iid random variables with bounded $2+\delta$ moments. Let $\boldsymbol{u}$ be the vector of norm 1:

$$\boldsymbol{u} = (\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{1/2}/||(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{1/2}||.$$

Then:

$$\begin{aligned}
& P\left(\sup_{\hat{A}:A\setminus\hat{A}\neq\emptyset}\left|(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}\right| > \right. \\
& \qquad\qquad \left. p^{-1/4}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_A\boldsymbol{\Lambda}\mathbf{X}_A(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})\right) \\
\leq \; & P\left(\sup_{\hat{A}:A\setminus\hat{A}\neq\emptyset} R^2||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right| > np^{-1/4}R^{-2}||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||^2\right) \\
\leq \; & P\left(\sup_{\boldsymbol{u}:||\boldsymbol{u}||=1}\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right| > \sqrt{n}p^{-1/4}R^{-4}\inf_{\hat{A}:A\setminus\hat{A}\neq\emptyset}||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||\right) \\
= \; & P\left(\sup_{\boldsymbol{u}:||\boldsymbol{u}||=1}\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right|^{2+\delta} > \left[\sqrt{n}p^{-1/4}R^{-4}\inf_{\hat{A}:A\setminus\hat{A}\neq\emptyset}||\boldsymbol{\beta}^*_{A[\hat{A}\cup A]} - \widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0}||\right]^{2+\delta}\right) \\
\leq \; & P\left(\sup_{\boldsymbol{u}:||\boldsymbol{u}||=1}\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right|^{2+\delta} > \left[\sqrt{n}p^{-1/4}R^{-4}\sqrt{\max(k_n,p)/n}\right]^{2+\delta}\right) \\
\leq \; & \frac{\sup_{\boldsymbol{u}:||\boldsymbol{u}||=1}E\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right|^{2+\delta}}{\left[R^{-4}p^{-1/4}\sqrt{\max(k_n,p)}\right]^{2+\delta}}
\end{aligned}$$

From Lemma 3 in Dicker et al (2012) we have that $\sup_{\boldsymbol{u}:||\boldsymbol{u}||=1} E\left|\boldsymbol{u}^T\tilde{\boldsymbol{\epsilon}}\right|^{2+\delta} < K$ for some constant K. Therefore:

$$P\Bigg(\sup_{\hat{A}:A\backslash\hat{A}\neq\emptyset}\left|(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]}-\widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\boldsymbol{\epsilon}\right| > p^{-1/4}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]}-\widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})^T\mathbf{X}^T_{A\cup\hat{A}}\boldsymbol{\Lambda}\mathbf{X}_{A\cup\hat{A}}(\boldsymbol{\beta}^*_{A[\hat{A}\cup A]}-\widehat{\boldsymbol{\beta}}_{\hat{A}:ML,0})\Bigg)\to 0$$

uniformly over all $\hat{A}$ such that $A\backslash\hat{A}\neq\emptyset$. ∎

S5 Additional small-p simulations results

The following tables summarize three simulation studies. In all simulation studies, the settings are exactly the same as in the paper. However, we consider two alternatives to the estimation procedure:

1. Working correlation structure - identity matrix. We consider using the identity matrix as the working correlation structure. We use both BIC and data validation for tuning parameter selection. Table 1 provides variable selection results and Table 2 provides the estimated and empirical standard errors of the regression parameter estimates.
2. Iterative algorithm in which the regression parameters are estimated and a penalty parameter is selected using either BIC or data validation, then precision matrix is estimated and penalty parameter is selected, then regression parameters are estimated, etc. Table 3 provides variable selection results and Table 4 provides the estimated and empirical standard errors of the regression parameter estimators.

In addition, Table 5 provides variable selection results in the (unrealistic) scenario when the true correlation structure is known, and the precision matrix is estimated parametrically accordingly.

Simulation	Method	size	TM	FP	FN	ME	PE
AR cov, $n = 50$	LASSO - BIC	4.95	0.17	2.10	0.15	1.60	10.57
	Adaptive LASSO - BIC	4.94	0.08	2.32	0.38	1.62	10.36
	SCAD - BIC	4.75	0.16	1.99	0.24	1.53	10.34
	SELO - BIC	3.89	0.33	1.17	0.28	1.42	10.56
	LASSO - validation	10.44	0.01	7.47	0.04	1.31	10.49
	Adaptive LASSO - validation	8.10	0.03	5.29	0.19	1.33	10.39
	SCAD - validation	7.77	0.06	4.86	0.10	1.15	10.38
	SELO - validation	3.75	0.46	0.92	0.18	0.97	10.45
AR cov, $n = 100$	LASSO - BIC	4.92	0.19	1.94	0.03	0.91	10.49
	Adaptive LASSO - BIC	4.95	0.12	2.01	0.06	0.66	10.36
	SCAD- BIC	4.30	0.32	1.34	0.04	0.60	10.37
	SELO - BIC	3.59	0.62	0.62	0.04	0.43	10.46
	LASSO - validation	11.08	0.01	8.09	0.01	0.71	10.46
	Adaptive LASSO - validation	8.01	0.03	5.03	0.02	0.54	10.39
	SCAD - validation	7.29	0.09	4.30	0.01	0.43	10.40
	SELO - validation	3.84	0.56	0.85	0.01	0.34	10.43
EX cov, $n = 50$	LASSO - BIC	5.02	0.15	2.17	0.15	1.55	10.22
	Adaptive LASSO - BIC	4.93	0.11	2.30	0.37	1.61	10.19
	SCAD - BIC	4.95	0.19	2.19	0.23	1.66	10.22
	SELO - BIC	3.94	0.37	1.20	0.26	1.39	10.35
	LASSO - validation	10.86	0.01	7.89	0.03	1.30	10.26
	Adaptive LASSO - validation	8.16	0.03	5.34	0.18	1.30	10.24
	SCAD - validation	7.82	0.06	4.92	0.10	1.13	10.22
	SELO - validation	3.64	0.51	0.83	0.19	0.91	10.23
EX cov, $n = 100$	LASSO - BIC	4.86	0.21	1.88	0.02	0.92	10.19
	Adaptive LASSO - BIC	4.77	0.16	1.84	0.07	0.70	10.17
	SCAD - BIC	4.08	0.42	1.13	0.06	0.63	10.16
	SELO - BIC	3.49	0.67	0.54	0.05	0.44	10.24
	LASSO - validation	11.48	0.01	8.49	0.01	0.74	10.22
	Adaptive LASSO - validation	8.28	0.04	5.30	0.02	0.58	10.21
	SCAD - validation	7.40	0.12	4.40	0.01	0.45	10.21
	SELO - validation	3.87	0.60	0.89	0.02	0.36	10.19

Table 1: Small-p simulation results for the 4 scenarios, by outcome correlation matrix and sample size, when using the two stage algorithm with the identity as the working correlation matrix.

Simulation	Method	$\beta_1 = 3$ se_est(se_emp)	$\beta_2 = 1.5$, es_est(se_emp)	$\beta_3 = 2$, es_est(se_emp)
AR cov, $n = 50$	LASSO - BIC	0.26(0.29)	0.14(0.27)	0.17(0.34)
	Adaptive LASSO - BIC	0.41(0.3)	0.06(0.47)	0.11(0.43)
	SCAD - BIC	0.31(0.3)	0.17(0.45)	0.21(0.58)
	SELO - BIC	0.42(0.3)	0.42(0.34)	0.41(0.47)
	LASSO - validation	0.3(0.3)	0.2(0.28)	0.24(0.35)
	Adaptive LASSO - validation	0.41(0.3)	0.1(0.34)	0.13(0.4)
	SCAD - validation	0.34(0.31)	0.19(0.38)	0.24(0.45)
	SELO - validation	0.42(0.3)	0.42(0.28)	0.41(0.43)
AR cov, $n = 100$	LASSO - BIC	0.19(0.21)	0.11(0.22)	0.13(0.23)
	Adaptive LASSO - BIC	0.29(0.21)	0.06(0.29)	0.07(0.24)
	SCAD - BIC	0.22(0.2)	0.13(0.33)	0.16(0.33)
	SELO - BIC	0.3(0.21)	0.3(0.25)	0.3(0.22)
	LASSO - validation	0.22(0.19)	0.16(0.22)	0.18(0.25)
	Adaptive LASSO - validation	0.3(0.21)	0.1(0.22)	0.1(0.24)
	SCAD - validation	0.24(0.21)	0.15(0.23)	0.18(0.23)
	SELO - validation	0.3(0.21)	0.3(0.23)	0.3(0.23)
EX cov, $n = 50$	LASSO - BIC	0.26(0.28)	0.15(0.29)	0.18(0.34)
	Adaptive LASSO - BIC	0.41(0.28)	0.07(0.47)	0.11(0.44)
	SCAD - BIC	0.32(0.31)	0.17(0.51)	0.22(0.58)
	SELO - BIC	0.41(0.28)	0.42(0.34)	0.41(0.46)
	LASSO - validation	0.31(0.3)	0.21(0.29)	0.24(0.33)
	Adaptive LASSO - validation	0.4(0.28)	0.11(0.39)	0.13(0.39)
	SCAD - validation	0.34(0.31)	0.2(0.31)	0.24(0.46)
	SELO - validation	0.42(0.3)	0.42(0.28)	0.41(0.43)
EX cov, $n = 100$	LASSO - BIC	0.19(0.23)	0.11(0.22)	0.13(0.24)
	Adaptive LASSO - BIC	0.29(0.23)	0.06(0.28)	0.07(0.28)
	SCAD - BIC	0.21(0.23)	0.12(0.31)	0.16(0.33)
	SELO - BIC	0.3(0.23)	0.3(0.22)	0.3(0.23)
	LASSO - validation	0.22(0.2)	0.16(0.19)	0.18(0.26)
	Adaptive LASSO - validation	0.29(0.23)	0.09(0.21)	0.1(0.25)
	SCAD - validation	0.24(0.23)	0.16(0.23)	0.18(0.24)
	SELO - VALIDATION	0.3(0.23)	0.3(0.21)	0.3(0.26)

Table 2: Small-p simulation results. Estimated and empirical standard errors for the regression parameter estimators provided in Table 1.

Simulation	Method	size	TM	FP	FN	ME	PE
AR cov, $n = 50$	LASSO - BIC	4.48	0.24	1.57	0.08	1.88	9.23
	Adaptive LASSO - BIC	3.49	0.31	0.90	0.41	1.61	9.10
	SCAD - BIC	3.53	0.15	1.15	0.61	1.90	9.09
	SELO - BIC	3.31	0.36	0.78	0.47	1.52	9.04
	LASSO - validation	9.93	0.05	6.98	0.05	1.63	9.15
	Adaptive LASSO - validation	7.10	0.09	4.35	0.24	1.64	9.11
	SCAD - validation	6.69	0.07	4.04	0.34	1.63	9.05
	SELO - validation	5.10	0.31	2.49	0.40	1.49	9.04
AR cov, $n = 100$	LASSO - BIC	4.22	0.35	1.22	0.00	0.84	9.02
	Adaptive LASSO - BIC	3.36	0.59	0.50	0.13	0.57	8.89
	SCAD - BIC	3.39	0.41	0.66	0.26	0.65	8.86
	SELO - BIC	3.05	0.70	0.24	0.19	0.49	8.85
	LASSO - validation	9.14	0.05	6.14	0.00	0.68	8.95
	Adaptive LASSO - validation	6.30	0.20	3.35	0.06	0.58	8.89
	SCAD - validation	6.14	0.22	3.19	0.06	0.58	8.86
	SELO - validation	5.35	0.47	2.42	0.07	0.55	8.87
EX cov, $n = 50$	LASSO - BIC	4.67	0.21	1.73	0.06	1.81	9.25
	Adaptive LASSO - BIC	3.65	0.28	1.02	0.37	1.46	9.28
	SCAD - BIC	3.62	0.12	1.23	0.61	1.73	9.37
	SELO - BIC	3.24	0.32	0.73	0.49	1.28	9.28
	LASSO - validation	10.24	0.06	7.27	0.03	1.60	9.26
	Adaptive LASSO - validation	7.20	0.12	4.43	0.23	1.54	9.28
	SCAD - validation	6.79	0.09	4.06	0.27	1.43	9.31
	SELO - validation	5.03	0.24	2.46	0.43	1.35	9.29
EX cov, $n = 100$	LASSO - BIC	4.28	0.33	1.28	0.00	0.82	9.12
	Adaptive LASSO - BIC	3.37	0.61	0.47	0.09	0.52	9.14
	SCAD - BIC	3.47	0.41	0.68	0.21	0.55	9.17
	SELO - BIC	3.20	0.66	0.34	0.14	0.40	9.12
	LASSO - validation	9.80	0.05	6.80	0.00	0.68	9.13
	Adaptive LASSO - validation	6.67	0.21	3.73	0.06	0.58	9.16
	SCAD - validation	6.16	0.23	3.22	0.05	0.52	9.16
	SELO - validation	5.79	0.37	2.88	0.09	0.52	9.16

Table 3: Small-p simulations results for the 4 scenarios, by outcome correlation matrix and sample size, when using an iterative algorithm in which the regression parameters and the precision matrix are alternately estimated, and tuning parameters are selected at each iteration.

Simulation	Method	$\beta_1 = 3$ se_est(se_emp)	$\beta_2 = 1.5$, es_est(se_emp)	$\beta_3 = 2$, es_est(se_emp)
AR cov, $n = 50$	LASSO - BIC	0.35(0.39)	0.29(0.43)	0.27(0.3)
	Adaptive LASSO - BIC	0.39(0.49)	0.22(0.9)	0.3(0.34)
	SCAD - BIC	0.36(0.58)	0(0.77)	0.28(0.48)
	SELO - BIC	0.38(0.54)	0.33(1.07)	0.33(0.3)
	LASSO - validation	0.38(0.35)	0.32(0.46)	0.31(0.29)
	Adaptive LASSO - validation	0.41(0.44)	0.32(0.64)	0.37(0.38)
	SCAD - validation	0.38(0.57)	0.28(0.93)	0.3(0.36)
	SELO - validation	0.41(0.52)	0.35(0.94)	0.34(0.32)
AR cov, $n = 100$	LASSO - BIC	0.27(0.3)	0.23(0.27)	0.22(0.19)
	Adaptive LASSO - BIC	0.28(0.38)	0.24(0.41)	0.23(0.2)
	SCAD - BIC	0.27(0.51)	0.23(0.68)	0.22(0.2)
	SELO - BIC	0.28(0.39)	0.26(0.4)	0.24(0.17)
	LASSO - validation	0.29(0.3)	0.25(0.27)	0.24(0.2)
	Adaptive LASSO - validation	0.3(0.35)	0.3(0.38)	0.25(0.2)
	SCAD - validation	0.28(0.39)	0.24(0.43)	0.23(0.19)
	SELO - validation	0.3(0.34)	0.27(0.32)	0.25(0.19)
EX cov, $n = 50$	LASSO - BIC	0.31(0.34)	0.26(0.37)	0.27(0.3)
	Adaptive LASSO - BIC	0.33(0.47)	0.22(0.75)	0.3(0.36)
	SCAD - BIC	0.32(0.61)	0(0.73)	0.28(0.44)
	SELO - BIC	0.33(0.53)	0.27(0.89)	0.32(0.28)
	LASSO - validation	0.33(0.31)	0.3(0.41)	0.31(0.28)
	Adaptive LASSO - validation	0.35(0.42)	0.34(0.59)	0.36(0.3)
	SCAD - validation	0.32(0.51)	0.26(0.84)	0.3(0.29)
	SELO - validation	0.34(0.48)	0.3(0.93)	0.33(0.26)
EX cov, $n = 100$	LASSO - BIC	0.23(0.2)	0.21(0.27)	0.21(0.2)
	Adaptive LASSO - BIC	0.24(0.23)	0.21(0.31)	0.23(0.22)
	SCAD - BIC	0.23(0.36)	0.21(0.57)	0.22(0.19)
	SELO - BIC	0.24(0.25)	0.23(0.31)	0.23(0.19)
	LASSO - validation	0.25(0.26)	0.24(0.24)	0.24(0.2)
	Adaptive LASSO - validation	0.25(0.33)	0.28(0.3)	0.25(0.19)
	SCAD - validation	0.24(0.32)	0.22(0.33)	0.23(0.18)
	SELO - validation	0.25(0.32)	0.24(0.32)	0.24(0.19)

Table 4: Small-p simulations results. Estimated and empirical standard errors for the regression parameter estimators provided in Table 3.

Simulation	Method	size	True model	false pos	false neg	ME	pred err
AR cov, $n = 50$	LASSO - BIC	5.22	0.08	2.29	0.07	1.64	9.52
	Adaptive LASSO - BIC	3.82	0.29	1.17	0.34	1.46	9.57
	SCAD - BIC	4.08	0.20	1.53	0.45	1.79	9.67
	SELO - BIC	3.72	0.40	1.00	0.28	1.52	9.62
	LASSO - validation	9.51	0.02	6.54	0.03	1.49	9.52
	Adaptive LASSO - validation	7.04	0.10	4.25	0.21	1.46	9.58
	SCAD - validation	5.82	0.10	3.21	0.39	1.69	9.65
	SELO - validation	4.52	0.41	1.84	0.32	1.52	9.61
AR cov, $n = 100$	LASSO - BIC	5.21	0.14	2.21	0.00	0.88	9.42
	Adaptive LASSO - BIC	3.73	0.46	0.88	0.14	0.61	9.44
	SCAD - BIC	3.71	0.51	0.90	0.19	0.63	9.47
	SELO - BIC	3.60	0.56	0.68	0.08	0.54	9.45
	LASSO - validation	10.62	0.02	7.62	0.00	0.80	9.43
	Adaptive LASSO - validation	6.82	0.16	3.87	0.04	0.63	9.45
	SCAD - validation	5.66	0.28	2.78	0.12	0.63	9.47
	SELO - validation	4.30	0.59	1.40	0.10	0.53	9.45
EX cov, $n = 50$	LASSO - BIC	5.32	0.14	2.38	0.06	1.46	9.35
	Adaptive LASSO - BIC	4.08	0.29	1.36	0.28	1.26	9.26
	SCAD - BIC	3.84	0.34	1.20	0.36	1.34	9.30
	SELO - BIC	3.94	0.44	1.16	0.22	1.35	9.24
	LASSO - validation	9.81	0.02	6.82	0.00	1.32	9.31
	Adaptive LASSO - validation	6.92	0.12	4.09	0.16	1.28	9.27
	SCAD - validation	4.93	0.22	2.23	0.30	1.17	9.29
	SELO - validation	3.92	0.48	1.19	0.27	1.13	9.25
EX cov, $n = 100$	LASSO - BIC	4.99	0.18	1.99	0.00	0.80	9.10
	Adaptive LASSO - BIC	3.84	0.49	0.91	0.08	0.50	9.12
	SCAD - BIC	3.44	0.66	0.52	0.08	0.39	9.11
	SELO - BIC	3.57	0.62	0.60	0.03	0.40	9.10
	LASSO - validation	10.39	0.02	7.39	0.00	0.67	9.11
	Adaptive LASSO - validation	6.76	0.16	3.80	0.03	0.50	9.11
	SCAD - validation	5.04	0.39	2.08	0.04	0.40	9.11
	SELO - validation	4.23	0.72	1.28	0.05	0.38	9.11

Table 5: Small-p simulation results for the 4 scenarios, by outcome correlation matrix and sample size, when using the two stage algorithm with the true correlation structures as the working correlation matrix.

S6 Additional large-p simulations results

In the following simulations, we provide additional simulation results, in which we apply the two-stage procedure on the same scenarios as in the large-p simulations section in the main manuscript, but use different estimators of the precision matrix. Table 6 provides results when the working correlation matrix is the true outcome correlation matrix, and Table 7 provides results when the working correlation matrix is the identity. In all simulations we had $n = 50$ subjects.

S7 Data set and code: description and instructions

The data set that was used in the data analysis section in the manuscript is supplied, as well as code for analysis, in the journal website. Three files are given:

1. *diabetes_data_set.txt* is a text file with the data set. For conveniency, only the probes used in the data analysis are supplied (i.e. non expressed probes that were filtered are not supplied) and the expression values are already log transformed.
2. *TSofer_sparse_mult_reg_code.R* in as R code supplying implementations of all the algorithm and functions needed in order to analysis the data set.
3. *TSofer_analysis_code.R* is the R code that reads the data, calls the algorithms, and performs the analysis.

penalty	mean T	mean FP	mean FN	mean ME
AR cov				
		$p_0 = 5,\ m = 5$		
LASSO	0.00	9.57	0	0.61
Adaptive LASSO	0.08	4.46	0	0.42
SCAD	0.18	3.56	0	0.26
SELO	0.68	0.74	0	0.20
		$p_0 = 20,\ m = 5$		
LASSO	0.00	15.37	0.00	0.82
Adaptive LASSO	0.01	6.06	0.00	0.51
SCAD	0.02	7.17	0.00	0.29
SELO	0.68	0.64	0.00	0.24
		$p_0 = 5,\ m = 20$		
LASSO	0.00	29.19	0.00	1.22
Adaptive LASSO	0.00	10.12	0.00	0.70
SCAD	0.03	9.06	0.00	0.29
SELO	0.39	2.62	0.00	0.30
		$p_0 = 20,\ m = 20$		
LASSO	0.00	29.19	0.00	1.22
Adaptive LASSO	0.00	13.55	0.00	0.80
SCAD	0.00	19.41	0.00	0.42
SELO	0.44	2.64	0.00	0.38
EX cov				
		$p_0 = 5,\ m = 5$		
LASSO	0.00	10.07	0	0.57
Adaptive LASSO	0.06	4.72	0	0.39
SCAD	0.20	3.93	0	0.27
SELO	0.70	0.66	0	0.18
		$p_0 = 20,\ m = 5$		
LASSO	0.00	15.70	0.00	0.72
Adaptive LASSO	0.01	7.76	0.02	1.10
SCAD	0.03	6.90	0.00	0.25
SELO	0.68	0.70	0.00	0.22
		$p_0 = 5,\ m = 20$		
LASSO	0.00	19.25	0.00	0.72
Adaptive LASSO	0.00	12.06	0.00	1.31
SCAD	0.08	7.38	0.00	0.22
SELO	0.40	2.77	0.00	0.24
		$p_0 = 20,\ m = 20$		
LASSO	0.00	29.09	0.00	0.89
Adaptive LASSO	0.00	16.61	0.00	1.50
SCAD	0.00	18.34	0.00	0.35
SELO	0.45	2.27	0.01	0.49

Table 6: Large-p simulation results when the working correlation matrix is the true outcome correlation matrix.

penalty	mean T	mean FP	mean FN	mean ME
AR cov				
		$p_0 = 5,\ m = 5$		
LASSO	0.00	9.32	0	0.72
Adaptive LASSO	0.05	4.81	0	0.55
SCAD	0.06	4.55	0	0.34
SELO	0.72	0.61	0	0.27
		$p_0 = 20,\ m = 5$		
LASSO	0.00	15.08	0.00	1.06
Adaptive LASSO	0.00	6.37	0.00	0.71
SCAD	0.00	9.24	0.00	0.49
SELO	0.74	0.58	0.00	0.34
		$p_0 = 5,\ m = 20$		
LASSO	0.00	17.76	0.00	1.18
Adaptive LASSO	0.00	10.96	0.00	1.13
SCAD	0.01	12.00	0.00	0.56
SELO	0.26	3.33	0.00	0.52
		$p_0 = 20,\ m = 20$		
LASSO	0.00	27.85	0.00	1.59
Adaptive LASSO	0.00	14.43	0.00	1.25
SCAD	0.00	26.89	0.00	0.92
SELO	0.48	2.27	0.00	0.55
EX cov				
		$p_0 = 5,\ m = 5$		
LASSO	0.00	9.50	0	0.72
Adaptive LASSO	0.05	4.79	0	0.54
SCAD	0.09	4.71	0	0.35
SELO	0.69	0.76	0	0.28
		$p_0 = 20,\ m = 5$		
LASSO	0.00	15.36	0.00	1.07
Adaptive LASSO	0.00	6.57	0.00	0.73
SCAD	0.01	9.30	0.00	0.49
SELO	0.73	0.57	0.00	0.33
		$p_0 = 5,\ m = 20$		
LASSO	0.00	21.74	0.00	1.26
Adaptive LASSO	0.00	12.40	0.00	1.20
SCAD	0.02	14.20	0.00	0.63
SELO	0.25	3.54	0.00	0.53
		$p_0 = 20,\ m = 20$		
LASSO	0.00	31.31	0.00	1.67
Adaptive LASSO	0.00	16.61	0.00	1.30
SCAD	0.00	28.72	0.00	1.00
SELO	0.42	4.07	0.01	0.64

Table 7: Large-p simulation results when the working correlation matrix is the identity matrix.