SUPPLEMENTARY MATERIAL TO "INNOVATED SCALABLE EFFICIENT ESTIMATION IN ULTRA-LARGE GAUSSIAN GRAPHICAL MODELS"

By Yingying Fan and Jinchi Lv

University of Southern California

This Supplementary Material contains the proofs of Theorem 3, Proposition 1, and additional technical details, as well as an extension of ISEE by incorporating the idea of feature screening.

APPENDIX B: ULTRA-LARGE GRAPH SCREENING

B.1. SIS-assisted ISEE. When the scale of the number of nodes p is ultra large, we can exploit the sure independence screening (SIS) in [16] to reduce the computational cost for each scaled Lasso regression. For each node j in the index set A_l with $1 \le l \le L$, the SIS ranks the components of the vector

$$\mathbf{w} = (w_k)_{k \in A_l^c} = \mathbf{X}_{A_l^c}^T \mathbf{X}_j \tag{A.1}$$

obtained by componentwise regression and for any given $\zeta \in (0,1)$, defines a submodel

$$\mathcal{M}_{jl,\zeta} = \{k \in A_l^c : \ |w_k| \text{ is among the first } [\zeta n] \text{ largest of all}\}, \tag{A.2}$$

where $[\zeta n]$ denotes the integer part of ζn. Here for simplicity, each node random variable X_j is assumed to have standard deviation one as in [16].

Following [16], based on the reduced model $\mathcal{M}_{jl,\zeta}$ obtained by the SIS one can construct the SIS-SLasso estimator $\widehat{\boldsymbol{\beta}}_{j,l}^*$, which is the scaled Lasso estimator $\widehat{\boldsymbol{\beta}}_{j,l}$ as defined in (12) with zero components outside the index set $\mathcal{M}_{jl,\zeta}$ for $\boldsymbol{\beta}$. Similarly as in (16), we define the initial ISEE estimator $\widehat{\boldsymbol{\Omega}}_{\text{ISEE,ini}}^*$ as the sample covariance matrix

$$\widehat{\boldsymbol{\Omega}}_{\text{ISEE,ini}}^* = n^{-1}(\widehat{\mathbf{X}}^*)^T \widehat{\mathbf{X}}^*, \tag{A.3}$$

where the estimator $\widehat{\mathbf{X}}^*$ for the oracle empirical matrix $\widetilde{\mathbf{X}}$ is constructed as in (15) using the SIS-SLasso estimator $\widehat{\boldsymbol{\beta}}_{j,l}^*$. Then we can construct the ISEE estimator for the graph $\widehat{\boldsymbol{\Omega}}_{\text{ISEE},g}$ and the ISEE estimator with refinement

$\widehat{\mathbf{\Omega}}_{\mathrm{ISEE}}$ based on the SIS-assisted initial ISEE estimator $\widehat{\mathbf{\Omega}}^*_{\mathrm{ISEE,ini}}$ in (A.3) as described in Section 2.3. Similarly the iterative SIS (ISIS) in [16] can also be applied to improve over the SIS in ultra-large scale problems.

B.2. Technical conditions.

CONDITION 3. *It holds that* $p > n$ *and* $\log p = O(n^{\gamma})$ *for some constant* $0 < \gamma < 1 - 2\kappa$ *with* κ *defined in Condition 4.*

CONDITION 4. *There exist some constants* $0 \le \kappa < 1/2$ *and* $c_1, c_2, c_3 > 0$ *such that for each* $j \in A_l$ *with* $1 \le l \le L$, *the support of the regression coefficient vector* $\boldsymbol{\beta}_{j,l} = (\beta_{jlk})_{k \in A_l^c}$ *in (11) admits a decomposition* $\mathrm{supp}(\boldsymbol{\beta}_{j,l}) = S_{jl0} \cup S_{jl1}$, *where for each* $k \in S_{jl0}$, $|\beta_{jlk}| \ge c_1 n^{-\kappa}$ *and* $|\mathrm{cov}(\beta_{jlk}^{-1} X_j, X_k)| \ge c_2$, *and for each* $k \in A_l^c$, $|\mathrm{cov}(\sum_{m \in S_{jl1}} \beta_{jlm} X_m, X_k)| \le c_3 \lambda$. *Moreover, it holds that*

$$\max_{j \in A_l,\, 1 \le l \le L} \max \Big\{ \sum_{m \in S_{jl1}} |\beta_{jlm}|, \lambda^{-1} \sum_{m \in S_{jl1}} \beta_{jlm}^2 \Big\} = O(K\lambda). \tag{A.4}$$

Conditions 3 and 4 are additional assumptions that facilitate the analysis of the SIS-assisted ISEE approach and ensure the sure screening property of the SIS procedure as in [16]. In particular, Condition 3 allows the dimensionality p to increase exponentially with sample size n. Condition 4 is imposed to ensure that the SIS-assisted ISEE estimate can enjoy nice asymptotic properties.

B.3. Theoretical properties. As introduced in Section B.1, to reduce the computational cost we can apply ISEE along with SIS or ISIS in the initial step for ultra-large graph screening. The computational cost can be further reduced if we also apply SIS or ISIS in the refinement step of estimating the link strength. In the refinement step, for each identified link (j, k) we can fit model (11) instead on the union of the supports of the jth and kth rows of $\widehat{\mathbf{\Omega}}_{\mathrm{ISEE},g}$, with nodes j and k excluded; see (60) in the proof of Theorem 2 for more details.

The following two theorems characterize the performance of the SIS-assisted ISEE estimators in both the initial step and the refinement step.

THEOREM 4. *Assume that the conditions of Theorem 1 and Conditions 3–4 hold and* ζ *in* (A.2) *is at least of order* $n^{-\gamma_0}$ *with some constant* $0 < \gamma_0 < 1 - 2\kappa$. *Then the SIS-assisted initial ISEE estimator* $\widehat{\mathbf{\Omega}}^*_{ISEE,ini}$ *in (A.3) satisfies the same properties as in Theorem 1.*

THEOREM 5. *Under the conditions of Theorems 2 and 4, the ISEE estimators $\widehat{\mathbf{\Omega}}_{ISEE,g}$ and $\widehat{\mathbf{\Omega}}_{ISEE}$ based on $\widehat{\mathbf{\Omega}}^{*}_{ISEE,ini}$ in (A.3) satisfy the same properties as in Theorem 2.*

APPENDIX C: PROOFS OF ADDITIONAL MAIN RESULTS

C.1. Proof of Theorem 3. By (16), (35), and the definition of the bias corrected initial ISEE estimator $\widehat{\mathbf{\Omega}}_{\text{ISEE,cini}}$ in (23) and (24), it suffices to consider the off-block-diagonal entries of the initial ISEE estimator $\widehat{\mathbf{\Omega}}_{\text{ISEE,ini}}$, that is, the submatrices $(\widehat{\mathbf{\Omega}}_{\text{ISEE,ini}})_{A_l,A_m}$ with $1 \le l \ne m \le L$. The bias of the initial ISEE estimator $\widehat{\mathbf{\Omega}}_{\text{ISEE,ini}}$ comes from these entries. Note that for each $l \ne m$, $(\widehat{\mathbf{\Omega}}_{\text{ISEE,ini}})_{A_l,A_m}$ admits the representation in (38). By (54) and (55), we see that the aforementioned bias is incurred by the second and third terms $\boldsymbol{\eta}_2$ and $\boldsymbol{\eta}_3$ in (38).

Due to the symmetry, we focus only on the term $\boldsymbol{\eta}_2$. Examining Part 1 of the proof of Theorem 1, we see that the bias in the term $\boldsymbol{\eta}_2$ is caused only by the additive component

$$\widetilde{\mathbf{F}}_2 = -\mathbf{\Omega}_{A_l}^T \mathbf{F}_2 \widehat{\mathbf{\Omega}}_{A_m}, \tag{A.5}$$

where $\mathbf{F}_2$ defined in (43) is given by $(n^{-1}\mathbf{X}_{A_l}^T\mathbf{E}_{A_l})^T(\widehat{\mathbf{C}}_{A_m}^{A_l} - \mathbf{C}_{A_m}^{A_l})$, and $\widehat{\mathbf{C}}_{A_m}^{A_l}$ and $\mathbf{C}_{A_m}^{A_l}$ denote submatrices of $\widehat{\mathbf{C}}_{A_m}$ and $\mathbf{C}_{A_m}$ consisting of rows with indices in A_l, respectively. We now add a bias correction term $\widehat{\mathbf{C}}_{A_m}^{A_l}\widehat{\mathbf{\Omega}}_{A_m}$ as specified in (24) to $(\widehat{\mathbf{\Omega}}_{\text{ISEE,ini}})_{A_l,A_m}$, and subsequently to $\widetilde{\mathbf{F}}_2$ given in (A.5). Let us consider the resulting new term

$$\widetilde{\mathbf{F}}_2^{*} = \widetilde{\mathbf{F}}_2 + \widehat{\mathbf{C}}_{A_m}^{A_l}\widehat{\mathbf{\Omega}}_{A_m} = \mathbf{F}_4 + \mathbf{F}_5 + \mathbf{C}_{A_m}^{A_l}\mathbf{\Omega}_{A_m}, \tag{A.6}$$

where $\mathbf{F}_4 = -[\mathbf{\Omega}_{A_l}^T(n^{-1}\mathbf{X}_{A_l}^T\mathbf{E}_{A_l})^T - I_{|A_l|}](\widehat{\mathbf{C}}_{A_m}^{A_l} - \mathbf{C}_{A_m}^{A_l})\widehat{\mathbf{\Omega}}_{A_m}$ and $\mathbf{F}_5 = \mathbf{C}_{A_m}^{A_l}(\widehat{\mathbf{\Omega}}_{A_m} - \mathbf{\Omega}_{A_m})$. We study these two terms $\mathbf{F}_4$ and $\mathbf{F}_5$ separately.

As in Part 1 of the proof of Theorem 1, we condition on the event $\mathcal{E} \cap (\cap_{3\le i\le 5}\mathcal{E}_i)$ hereafter. Note that $\widehat{\mathbf{C}}_{A_m}^{A_l} - \mathbf{C}_{A_m}^{A_l}$ is exactly the matrix $\mathbf{F}_3$ introduced therein. In light of the definitions of $\mathcal{E}$, $\mathcal{E}_4$, and $\mathcal{E}_5$ in (A.40) and (45)–(46), by the facts of $\|\mathbf{\Omega}_{A_l}\|_\infty = O(1)$ and $\widehat{\mathbf{\Omega}}_{A_m} = O(1)$ it holds uniformly over $1 \le l \ne m \le L$ that

$$\begin{aligned} \|\mathbf{F}_4\|_\infty &\le O(1)\|n^{-1}\mathbf{X}_{A_l}^T\mathbf{E}_{A_l} - \mathbf{\Omega}_{A_l}^{-1}\|_\infty \|\mathbf{F}_3\|_\infty O(1) \\ &\le O(\lambda)\cdot O(K^\alpha\lambda) = O(K^\alpha\lambda^2). \end{aligned} \tag{A.7}$$

Using similar arguments to those in the proof of Lemma 2, we can show that $\|\mathbf{C}_{A_m}^{A_l}\|_\infty = \|-\mathbf{\Omega}_{A_l,A_m}\mathbf{\Omega}_{A_m}^{-1}\|_\infty = O(1)$, which along with (A.40) entails

$$\|\mathbf{F}_5\|_\infty = O\left\{\max(K\lambda^2, \lambda)\right\}. \tag{A.8}$$

Since $\alpha \le 1/2$ by Condition 2, it follows from (A.7) and (A.8) that

$$\|\mathbf{F}_4 + \mathbf{F}_5\|_\infty \le O\left\{\max(K\lambda^2, \lambda)\right\}. \tag{A.9}$$

Observe that $\mathbf{C}_{A_m}^{A_l}\mathbf{\Omega}_{A_m} = -\mathbf{\Omega}_{A_l,A_m}\mathbf{\Omega}_{A_m}^{-1}\mathbf{\Omega}_{A_m} = -\mathbf{\Omega}_{A_l,A_m}$. Therefore, combining (A.6) and (A.9) proves the desired bound for the bias corrected initial ISEE estimator $\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,cini}}$ with off-block-diagonal entries

$$\left(\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,cini}}\right)_{A_l,A_m} = -\left[\left(\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,ini}}\right)_{A_l,A_m} + \widehat{\mathbf{C}}_{A_l}^{A_m}\widehat{\mathbf{\Omega}}_{A_l} + \widehat{\mathbf{C}}_{A_m}^{A_l}\widehat{\mathbf{\Omega}}_{A_m}\right];$$

that is, with the same probability bound as in Theorem 1 it holds that

$$\left\|\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,cini}} - \mathbf{\Omega}\right\|_\infty = O\left\{\max(K\lambda^2, \lambda)\right\},$$

which order is in fact $O(\lambda)$ as explained in the proof of Theorem 2.

The second part of Theorem 3, which is graph recovery consistency of the bias corrected initial ISEE estimator $\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,cini}}$, can be proved using similar arguments to those in the proof for part a of Theorem 2, by noting that $\widehat{\mathbf{\Omega}}_{\mathrm{ISEE},g} = T_\tau(\widehat{\mathbf{\Omega}}_{\mathrm{ISEE,cini}})$ and $\omega_0^* = C\lambda$ with $C > 0$ some sufficiently large constant.

C.2. Proof of Theorem 4. We first show that the two events $\mathcal{H}_1$ and $\mathcal{H}_2$ defined as

$$\mathcal{H}_1 = \bigcap\nolimits_{j\in A_l, 1\le l\le L}\{S_{jl0} \subset \mathcal{M}_{jl,\zeta}\} \tag{A.10}$$

and

$$\mathcal{H}_2 = \left\{\max\nolimits_{j\in A_l,1\le l\le L}\left\|n^{-1}\mathbf{X}_{A_l^c}^T\sum\nolimits_{m\in S_{jl1}}\beta_{jlm}\mathbf{X}_m\right\|_\infty \le O(\lambda)\right\} \tag{A.11}$$

have large probabilities. The event $\mathcal{H}_1$ in (A.10) characterizes the sure screening property of the SIS associated with the sets of indices S_{jl0}. It is easy to check that Conditions 1–4 in [16] are entailed by our Conditions 1 and 3–4, with $\mathcal{M}_*$ replaced by S_{jl0}. In particular, they verified the property C (a concentration property) for Gaussian distributions.

A key observation is that the proof of Theorem 1 in [16] applies equally well to the case where the set of desired effects S_{jl0} plays the role of $\mathcal{M}_*$

and the set of additional effects $S_{jl1} = \mathrm{supp}(\boldsymbol{\beta}_{j,l}) \setminus S_{jl0}$ may not be empty. Thus an application of the same arguments leads to a similar conclusion to that in Theorem 1 of [16]; that is, for ζ at least in the order of $n^{-\gamma_0}$ with some positive constant $\gamma_0 < 1 - 2\kappa$, we have

$$P\{S_{jl0} \subset \mathcal{M}_{jl,\zeta}\} = 1 - O\left\{\exp[-Cn^{1-2\kappa}/(\log n)]\right\}, \tag{A.12}$$

where C is some positive constant. Since $\log p = O(n^{\gamma})$ with constant $0 < \gamma < 1-2\kappa$ by Condition 3, we see immediately from (A.12) and Bonferroni's inequality over all nodes j in the index sets A_l that

$$P(\mathcal{H}_1) \geq 1 - p \cdot o\left\{p^{-(\delta-1)}\right\} = 1 - o\left\{p^{-(\delta-2)}\right\}. \tag{A.13}$$

Note that for each $k \in A_l^c$, the expectation of $n^{-1}\mathbf{X}_k^T \sum_{m\in S_{jl1}} \beta_{jlm}\mathbf{X}_m$ is equal to $\mathrm{cov}(\sum_{m\in S_{jl1}} \beta_{jlm}X_m, X_k)$. Thus in view of the assumption of $\max_{k\in A_l^c} |\mathrm{cov}(\sum_{m\in S_{jl1}} \beta_{jlm}X_m, X_k)| \leq c_3\lambda$ by Condition 4, using similar arguments to those for proving (A.35) with t chosen to be $[(\delta+1)(\log p)/(cn)]^{1/2}$ leads to

$$P(\mathcal{H}_2) \geq 1 - p(p-1) \cdot O\left\{p^{-(\delta+1)}\right\} = 1 - o\left\{p^{-(\delta-2)}\right\}. \tag{A.14}$$

Combining (A.13) and (A.14) yields the desired probability bound

$$P(\mathcal{H}_1 \cap \mathcal{H}_2) \geq 1 - o\left\{p^{-(\delta-2)}\right\}. \tag{A.15}$$

From now on we condition on the event $\mathcal{H}_1 \cap \mathcal{H}_2$. On this event, for each node j in the index set A_l, the submodel $\mathcal{M}_{jl,\zeta}$ given by the SIS contains the set of desired effects S_{jl0}. In light of (A.11), we can treat the component $\sum_{m\in S_{jl1}} \beta_{jlm}\mathbf{X}_m$ of the mean vector $\mathbf{X}_{A_l^c}\boldsymbol{\beta}_{j,l}$ in the univariate linear regression model (11) as part of the error vector in the technical analysis for the scaled Lasso. A key observation is that all the error bounds and probability bounds used in the arguments for proving Lemma 1 hold uniformly over the submodels $\mathcal{M}_{jl,\zeta}$. Thus an application of the proof of Lemma 1 shows that with probability $1 - o\{p^{-(\delta-2)}\}$ tending to one, it holds uniformly over all nodes j in the index sets A_l with $1 \leq l \leq L$ and all submodels $\mathcal{M}_{jl,\zeta}$ that

$$\left\|\widehat{\boldsymbol{\beta}}^*_{j,l,\mathcal{M}_{jl,\zeta}} - \boldsymbol{\beta}_{j,l,\mathcal{M}_{jl,\zeta}}\right\|_1 = O(K\lambda), \tag{A.16}$$

$$n^{-1}\left\|\mathbf{X}_{\mathcal{M}_{jl,\zeta}}(\widehat{\boldsymbol{\beta}}^*_{j,l,\mathcal{M}_{jl,\zeta}} - \boldsymbol{\beta}_{j,l,\mathcal{M}_{jl,\zeta}})\right\|_2^2 = O(K\lambda^2), \tag{A.17}$$

where $\widehat{\boldsymbol{\beta}}^*_{j,l}$ denotes the SIS-SLasso estimator, which is the scaled Lasso estimator $\widehat{\boldsymbol{\beta}}_{j,l}$ as defined in (12) with zero components for $\boldsymbol{\beta}$ outside the reduced

index set $\mathcal{M}_{jl,\zeta}$ obtained by the SIS, and $\mathcal{M}_{jl,\zeta}$ in the subscripts indicates the corresponding subvectors or submatrices.

In view of (A.15), the intersection of the event $\mathcal{H}_1 \cap \mathcal{H}_2$ and the one given in (A.16)–(A.17) still has large probability $1 - o\{p^{-(\delta-2)}\}$. On such an event, it follows immediately from the sure screening property of $S_{jl0} \subset \mathcal{M}_{jl,\zeta}$, (A.16), and (A.4) that

$$\left\| \widehat{\boldsymbol{\beta}}_{j,l}^{*} - \boldsymbol{\beta}_{j,l} \right\|_1 = O(K\lambda). \tag{A.18}$$

Note that the proof of Theorem 2 in [33] applies equally well for the largest singular value to show that

$$P\left\{ \max_{|\Lambda| \le \widetilde{K}} \lambda_{\max}(n^{-1}\mathbf{X}_{\Lambda}^{T}\mathbf{X}_{\Lambda}) \le O(1) \right\} \le p^{\widetilde{K}} e^{-Cn}, \tag{A.19}$$

where $\widetilde{K}$ is as defined in the proof of Lemma 1 and C is some positive constant. Since $\widetilde{K} \le \widetilde{c}_0 n/(\log p)$ for some sufficiently small positive constant $\widetilde{c}_0$, it is easy to derive that (A.19) entails

$$P\left\{ \max_{|\Lambda| \le \widetilde{K}} \lambda_{\max}(n^{-1}\mathbf{X}_{\Lambda}^{T}\mathbf{X}_{\Lambda}) \le O(1) \right\} = 1 - o\left\{ p^{-(\delta-2)} \right\}. \tag{A.20}$$

Thus conditioning on this additional event does not change our asymptotic probability bound $1 - o\{p^{-(\delta-2)}\}$.

Denote by $\Lambda_0 = \operatorname{supp}(\boldsymbol{\beta}_{j,l}) \setminus \mathcal{M}_{jl,\zeta}$. Since $\|\boldsymbol{\beta}_{j,l}\|_0 \le \widetilde{K}$ as shown in the proof of Lemma 1 which implies $|\Lambda_0| \le \widetilde{K}$, by (A.20), (A.4) in Condition 4, and $S_{jl0} \subset \mathcal{M}_{jl,\zeta}$ we have

$$\begin{aligned} n^{-1}\|\mathbf{X}_{\Lambda_0}\boldsymbol{\beta}_{j,l,\Lambda_0}\|_2^2 &\le \lambda_{\max}(n^{-1}\mathbf{X}_{\Lambda_0}^{T}\mathbf{X}_{\Lambda_0})\|\boldsymbol{\beta}_{j,l,\Lambda_0}\|_2^2 \\ &\le \lambda_{\max}(n^{-1}\mathbf{X}_{\Lambda_0}^{T}\mathbf{X}_{\Lambda_0})\|\boldsymbol{\beta}_{j,l,S_{jl1}}\|_2^2 \\ &\le O(1) \cdot O(K\lambda^2) = O(K\lambda^2). \end{aligned} \tag{A.21}$$

Combining (A.17) and (A.21) leads to

$$n^{-1} \left\| \mathbf{X}_{A_l^c}(\widehat{\boldsymbol{\beta}}_{j,l}^{*} - \boldsymbol{\beta}_{j,l}) \right\|_2^2 = O(K\lambda^2). \tag{A.22}$$

In light of (A.18) and (A.22), we have shown that with probability $1 - o\{p^{-(\delta-2)}\}$ tending to one, it holds uniformly over all nodes j in the index sets A_l with $1 \le l \le L$ that the same bounds as (A.23)–(A.24) in Lemma 1 are also valid for the SIS-SLasso estimator. Therefore, the same arguments as in the proof of Theorem 1 carry through.

C.3. Proof of Theorem 5. Theorem 5 holds immediately as a consequence of Theorems 2 and 4.

APPENDIX D: PROOFS OF TECHNICAL RESULTS

D.1. Lemma 1 and its proof.

LEMMA 1. *Under Condition 1, with probability* $1 - o\{p^{-(\delta-2)}\}$ *tending to one it holds uniformly over all nodes* j *in the index sets* A_l *with* $1 \le l \le L$ *and simultaneously that*

$$\|\widehat{\boldsymbol{\beta}}_{j,l} - \boldsymbol{\beta}_{j,l}\|_1 = O(K\lambda), \tag{A.23}$$

$$n^{-1}\|\mathbf{X}_{A_l^c}(\widehat{\boldsymbol{\beta}}_{j,l} - \boldsymbol{\beta}_{j,l})\|_2^2 = O(K\lambda^2), \tag{A.24}$$

$$\|n^{-1}\mathbf{X}_{A_l^c}^T\mathbf{E}_{j,l}\|_\infty = O(\lambda), \tag{A.25}$$

where $\widehat{\theta}_{j,l} = n^{-1}\widehat{\mathbf{E}}_{j,l}^T\widehat{\mathbf{E}}_{j,l}$, $\widetilde{\theta}_{j,l} = n^{-1}\mathbf{E}_{j,l}^T\mathbf{E}_{j,l}$, *and the additional subscript* l *indicates the same scalars and vectors as defined previously with the index set* A *replaced by* A_l.

Proof of Lemma 1. Let us first make a few observations. First, for each index set A_l, the random error vector $\boldsymbol{\eta}_{A_l}$ in the scalar form of the multivariate linear regression model (7) with index set $A = A_l$ is Gaussian with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}_{A_l}^{-1}$ and independent of $\mathbf{x}_{A_l^c}$. Since by Condition 1, the spectrum of the precision matrix $\boldsymbol{\Omega}$ is bounded between M^{-1} and M. We see immediately that the spectrum of its principal submatrix $\boldsymbol{\Omega}_{A_l}$ is also bounded between M^{-1} and M, so is that of its inverse $\boldsymbol{\Omega}_{A_l}^{-1}$. This shows that for each corresponding univariate linear regression model (11), its error vector $\mathbf{E}_{j,l}$ is $N(\mathbf{0}, \theta_{j,l}I_n)$ with marginal variance $\theta_{j,l}$ bounded between M^{-1} and M, where the additional subscript l indicates the same scalars and vectors as defined previously with the index set A replaced by A_l.

Second, by Condition 1, the precision matrix $\boldsymbol{\Omega}$ is K-sparse, that is, each of its row or column has at most K nonzero off-diagonal entries. Since $\max_l |A_l| = O(1)$, it follows that the total number of nonzero entries $\widetilde{K}$ in the submatrix $\boldsymbol{\Omega}_{A_l^c,A_l}$ is bounded from above by $K|A_l| = O(K)$. In view of $K \le c_0 n/(\log p)$ for some sufficiently small positive constant c_0, we have $\widetilde{K} \le \widetilde{c}_0 n/(\log p)$ with $\widetilde{c}_0 = O(c_0)$ still some sufficiently small positive constant. Thus for each index set A_l, the regression coefficient matrix $\mathbf{C}_{A_l} = -\boldsymbol{\Omega}_{A_l^c,A_l}\boldsymbol{\Omega}_{A_l}^{-1}$ in the matrix form of the multivariate linear regression model (9) with index set $A = A_l$ satisfies that each column vector has at

most $\widetilde{K}$ nonzero components. This shows that for each corresponding univariate linear regression model (11), its regression coefficient vector $\boldsymbol{\beta}_{j,l}$ has sparsity $\|\boldsymbol{\beta}_{j,l}\|_0 \le \widetilde{K} = O(K) \le \widetilde{c}_0 n/(\log p)$ uniformly over all nodes j and index sets A_l.

Third, for each index set A_l, the corresponding univariate linear regression model (11) is a linear regression model with Gaussian design matrix $\mathbf{X}_{A_l^c}$ and Gaussian error vector $\mathbf{E}_{j,l}$ that is independent of $\mathbf{X}_{A_l^c}$. Note that in light of $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ and (1), $\mathbf{X}_{A_l^c} \sim N(\mathbf{0}, I_n \otimes \boldsymbol{\Sigma}_{A_l^c})$, where $\boldsymbol{\Sigma}_{A_l^c}$ denotes the principal submatrix of $\boldsymbol{\Sigma}$ given by the index set A_l^c. Since $\boldsymbol{\Omega}$ has spectrum bounded between M^{-1} and M, the spectrum of $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$ is also bounded between M^{-1} and M and so is that of its principal submatrix $\boldsymbol{\Sigma}_{A_l^c}$.

Denote by $\mathcal{E}_{j,l}$ the event that the bounds (A.23)–(A.25) hold simultaneously for node j in the index set A_l. With the above three observations, an application of the proof of Lemma 2 in [39] shows that

$$P(\mathcal{E}_{j,l}) = 1 - o\left\{p^{-(\delta-1)}\right\}. \tag{A.26}$$

Thus applying Bonferroni's inequality over all nodes j in the index sets A_l along with (A.26) yields the uniform bounds (A.23)–(A.25) satisfied with probability

$$P(\mathcal{E}_1) \ge 1 - p \cdot o\left\{p^{-(\delta-1)}\right\} = 1 - o\left\{p^{-(\delta-2)}\right\} \tag{A.27}$$

which converges to one since $\delta \ge 2$, where the event $\mathcal{E}_1$ is defined as

$$\mathcal{E}_1 = \bigcap\nolimits_{j \in A_l, 1 \le l \le L} \mathcal{E}_{j,l}. \tag{A.28}$$

In view of $\widehat{\mathbf{E}}_{j,l} = \mathbf{X}_j - \mathbf{X}_{A_l^c}\widehat{\boldsymbol{\beta}}_{j,l}$, the fact that $\widehat{\theta}_{j,l} = n^{-1}\widehat{\mathbf{E}}_{j,l}^T\widehat{\mathbf{E}}_{j,l}$ follows easily from the definition of the minimizer $(\widehat{\boldsymbol{\beta}}_{j,l}, \widehat{\theta}_{j,l}^{1/2})$ of the scaled Lasso problem (12).

D.2. Lemma 2 and its proof.

LEMMA 2. *Under Condition 1, with probability* $1 - o\{p^{-(\delta-2)}\}$ *tending to one it holds uniformly over* $1 \le l \le L$ *that*

$$\|\widehat{\boldsymbol{\Omega}}_{A_l} - \boldsymbol{\Omega}_{A_l}\|_\infty = O\left\{\max\left(K\lambda^2, \lambda\right)\right\}, \tag{A.29}$$

where $\|\cdot\|_\infty$ *denotes the entrywise* L_∞*-norm of a given matrix.*

Proof of Lemma 2. Note that by (13) and (9), we have the following decomposition of the residual matrix

$$\widehat{\mathbf{E}}_{A_l} = \mathbf{X}_{A_l} - \mathbf{X}_{A_l^c}\widehat{\mathbf{C}}_{A_l} = \mathbf{E}_{A_l} - \mathbf{X}_{A_l^c}(\widehat{\mathbf{C}}_{A_l} - \mathbf{C}_{A_l}), \tag{A.30}$$

where $\widehat{\mathbf{C}}_{A_l} = (\widehat{\boldsymbol{\beta}}_{j,l})_{j\in A_l}$ is a $(p - |A_l|) \times |A_l|$ matrix of estimated regression coefficients. Combining (14) and (A.30) yields

$$\widehat{\boldsymbol{\Omega}}_{A_l}^{-1} - \boldsymbol{\Omega}_{A_l}^{-1} = n^{-1}\widehat{\mathbf{E}}_{A_l}^T\widehat{\mathbf{E}}_{A_l} - \boldsymbol{\Omega}_{A_l}^{-1} = \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \boldsymbol{\xi}_3, \tag{A.31}$$

where $\boldsymbol{\xi}_1 = n^{-1}\mathbf{E}_{A_l}^T\mathbf{E}_{A_l} - \boldsymbol{\Omega}_{A_l}^{-1}$, $\boldsymbol{\xi}_2 = -2n^{-1}\mathbf{E}_{A_l}^T\mathbf{X}_{A_l^c}(\widehat{\mathbf{C}}_{A_l} - \mathbf{C}_{A_l})$, and $\boldsymbol{\xi}_3 = n^{-1}(\widehat{\mathbf{C}}_{A_l} - \mathbf{C}_{A_l})^T\mathbf{X}_{A_l^c}^T\mathbf{X}_{A_l^c}(\widehat{\mathbf{C}}_{A_l} - \mathbf{C}_{A_l})$. Let us first consider the last two terms $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ conditional on the event $\mathcal{E}_1$ defined in (A.28). On the event $\mathcal{E}_1$, bounds (A.25) and (A.23) control the maximum rowwise L_∞-norm of matrix $n^{-1}\mathbf{E}_{A_l}^T\mathbf{X}_{A_l^c}$ and maximum columnwise L_1-norm of matrix $\widehat{\mathbf{C}}_{A_l} - \mathbf{C}_{A_l}$, respectively, which lead to

$$\|\boldsymbol{\xi}_2\|_\infty = O(K\lambda^2), \tag{A.32}$$

where $\|\cdot\|_\infty$ denotes the entrywise L_∞-norm of a given matrix. An application of the Cauchy-Schwarz inequality along with bound (A.24) results in

$$\|\boldsymbol{\xi}_3\|_\infty = O(K\lambda^2). \tag{A.33}$$

Note that bounds (A.32) and (A.33) are uniform over $1 \le l \le L$. It remains to consider the first term $\boldsymbol{\xi}_1$.

As mentioned in the proof of Lemma 1, the spectrum of $\boldsymbol{\Omega}_{A_l}^{-1}$ is bounded between M^{-1} and M. In view of (9) and (7), $n^{-1}\mathbf{E}_{A_l}^T\mathbf{E}_{A_l}$ is the oracle sample covariance matrix estimator for $\boldsymbol{\Omega}_{A_l}^{-1}$. Thus the concentration bounds in [41] and [2], together with Bonferroni's inequality and $\max_l |A_l| = O(1)$, yield for any $t \le \alpha$,

$$P\{\|\boldsymbol{\xi}_1\|_\infty \le t\} = 1 - O(e^{-cnt^2}), \tag{A.34}$$

where c and α are some positive constants. Taking $t = [\delta(\log p)/(cn)]^{1/2}$ in (A.34) and applying Bonferroni's inequality over $1 \le l \le L$ lead to

$$P(\mathcal{E}_2) \ge 1 - p \cdot O(e^{-cnt^2}) = 1 - O\left\{p^{-(\delta-1)}\right\} = 1 - o\left\{p^{-(\delta-2)}\right\}, \tag{A.35}$$

where the event $\mathcal{E}_2$ is defined as

$$\mathcal{E}_2 = \left\{\max_{1\le l\le L}\left\|n^{-1}\mathbf{E}_{A_l}^T\mathbf{E}_{A_l} - \boldsymbol{\Omega}_{A_l}^{-1}\right\|_\infty \le t = O(\lambda)\right\}. \tag{A.36}$$

Therefore, combining (A.31)–(A.34) and (A.35) leads to

$$P\left\{\max_{1\le l\le L}\left\|\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\right\|_\infty = O\left\{\max\left(K\lambda^2,\lambda\right)\right\}\right\} = 1-o\left\{p^{-(\delta-2)}\right\}. \tag{A.37}$$

We still need to derive the bounds for the matrices $\widehat{\mathbf{\Omega}}_{A_l}$.

Let us work with the bound $\|\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\|_\infty = O\{\max(K\lambda^2,\lambda)\}$. Since $|A_l| = O(1)$, the Frobenius norm $\|\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\|_F = O\{\max(K\lambda^2,\lambda)\}$. In light of Condition 1, the quantity $O\{\max(K\lambda^2,\lambda)$ is bounded above by some sufficiently small positive constant. Then it follows from the matrix perturbation theory (Corollary 6.3.8 of [27]) that

$$\begin{aligned}\lambda_{\min}(\widehat{\mathbf{\Omega}}_{A_l}^{-1}) &\ge \lambda_{\min}(\mathbf{\Omega}_{A_l}^{-1}) - \|\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\|_F \\ &\ge M^{-1} - O\left\{\max\left(K\lambda^2,\lambda\right)\right\} \ge (2M)^{-1}\end{aligned}$$

for large enough n. The above spectral inequality leads to $\lambda_{\max}(\widehat{\mathbf{\Omega}}_{A_l}) = \lambda_{\min}^{-1}(\widehat{\mathbf{\Omega}}_{A_l}^{-1}) = O(1)$. Similarly, we can show that $\lambda_{\min}(\widehat{\mathbf{\Omega}}_{A_l})$ is also bounded away from zero.

Note a fact that the entrywise L_∞-norm of any symmetric positive definite matrix is bounded above by its largest eigenvalue. This claim follows from the facts that each diagonal entry is positive and no larger than the largest eigenvalue and that the 2×2 principal submatrix corresponding to each off-diagonal entry is necessarily nonsingular. Since both $\mathbf{\Omega}_{A_l}$ and $\widehat{\mathbf{\Omega}}_{A_l}$ have spectra bounded away from 0 and ∞, we see that $\|\mathbf{\Omega}_{A_l}\|_\infty = O(1)$ and $\|\widehat{\mathbf{\Omega}}_{A_l}\|_\infty = O(1)$, which along with $\max_{1\le l\le L}\|\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\|_\infty = O\{\max(K\lambda^2,\lambda)\}$ and $\max_l|A_l| = O(1)$ entails

$$\left\|\widehat{\mathbf{\Omega}}_{A_l}-\mathbf{\Omega}_{A_l}\right\|_\infty = \left\|\mathbf{\Omega}_{A_l}\left(\widehat{\mathbf{\Omega}}_{A_l}^{-1}-\mathbf{\Omega}_{A_l}^{-1}\right)\widehat{\mathbf{\Omega}}_{A_l}\right\|_\infty = O\left\{\max(K\lambda^2,\lambda)\right\}. \tag{A.38}$$

Therefore, combining (A.27), (A.35), and (A.37)–(A.38) yields

$$P(\mathcal{E}) = 1-o\left\{p^{-(\delta-2)}\right\}, \tag{A.39}$$

where the event $\mathcal{E}$ is defined as

$$\mathcal{E} = \mathcal{E}_1\cap\mathcal{E}_2\cap\left\{\max_{1\le l\le L}\left\|\widehat{\mathbf{\Omega}}_{A_l}-\mathbf{\Omega}_{A_l}\right\|_\infty = O\left\{\max(K\lambda^2,\lambda)\right\}\right\}. \tag{A.40}$$

Hereafter we condition on the event $\mathcal{E}$.

D.3. Proof of Proposition 1. For any $\mathbf{\Omega} \in \mathcal{G}(M, K)$, we know that each row of $\mathbf{\Omega}$ has at most $K+1$ nonzero components and the spectrum of $\mathbf{\Omega}$ is bounded between M^{-1} and M. Thus it follows easily that for $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ and any $\mathbf{u} \neq \mathbf{0}$,

$$\|\mathbf{u}\|_\infty = \|\mathbf{\Omega}\mathbf{\Sigma}\mathbf{u}\|_\infty \leq \|\mathbf{\Omega}\|_{\infty,\infty}\|\mathbf{\Sigma}\mathbf{u}\|_\infty, \tag{A.41}$$

where $\|\cdot\|_{\infty,\infty}$ denotes the operator norm of a matrix induced by the L_∞-norm. Note that $\|\mathbf{\Omega}\|_{\infty,\infty}$ is the maximum rowwise L_1-norm of $\mathbf{\Omega}$, which is bounded above by $(K+1)^{1/2}$ multiplied by the maximum rowwise L_2-norm of $\mathbf{\Omega}$, thanks to the Cauchy-Schwarz inequality and the fact that each row of $\mathbf{\Omega}$ has L_0-norm bounded above by $K+1$. By the definition of the spectral norm, the maximum rowwise L_2-norm of $\mathbf{\Omega}$ is further bounded above by $\lambda_{\max}(\mathbf{\Omega}) \leq M$, which entails

$$\|\mathbf{\Omega}\|_{\infty,\infty} \leq (K+1)^{1/2} M. \tag{A.42}$$

Combining (A.41)–(A.42) yields the desired bound $\inf\{\|\mathbf{\Sigma}\mathbf{u}\|_\infty / \|\mathbf{u}\|_\infty : \mathbf{u} \neq \mathbf{0}\} \geq (K+1)^{-1/2} M^{-1}$.

D.4. Lemma 3 and its proof.

LEMMA 3. *Assume that Conditions 1–2 hold and $K^{1+\alpha}\lambda = o(1)$. Then with probability $1 - o\{p^{-(\delta-2)}\}$ tending to one it holds uniformly over all nodes j in the index sets A_l with $1 \leq l \leq L$ that the L_∞-norm cone invertibility factor*

$$F_{\infty,j,l} = \inf\left\{ \frac{\|\widehat{\mathbf{R}}_{j,l}\mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty} : \|\mathbf{u}_{S_{j,l}^c}\|_1 \leq \xi \|\mathbf{u}_{S_{j,l}}\|_1 \neq 0 \right\} \tag{A.43}$$

satisfies $F_{\infty,j,l} \geq c_1 F_\infty$, where $c_1 < 1$ is some positive constant, $S_{j,l}$ denotes the support $\mathrm{supp}(\boldsymbol{\beta}_{j,l})$, and $\widehat{\mathbf{R}}_{j,l} = n^{-1}\mathbf{Y}_{A_l^c}^T \mathbf{Y}_{A_l^c}$ with $\mathbf{Y}_{A_l^c}$ the design matrix $\mathbf{X}_{A_l^c}$ rescaled columnwise to have L_2-norm $n^{1/2}$ for each column.

Proof of Lemma 3. Let $\mathbf{R}$ be the correlation matrix corresponding to the covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})$. Since the spectrum of $\mathbf{\Sigma}$ is bounded between M^{-1} and M thanks to the same property of $\mathbf{\Omega}$, all diagonal entries σ_{jj} of $\mathbf{\Sigma}$ are also bounded between M^{-1} and M and so are all their reciprocals σ_{jj}^{-1}. Thus the L_1-norms and L_∞-norms induced by both linear transformations corresponding to matrices $\mathbf{S} = \mathrm{diag}\{\sigma_{11}^{1/2}, \cdots, \sigma_{pp}^{1/2}\}$ and

$\mathbf{S}^{-1} = \mathrm{diag}\{\sigma_{11}^{-1/2}, \cdots, \sigma_{pp}^{-1/2}\}$ are equivalent to the original ones. Thus it follows from the identity

$$\mathbf{R} = \mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1} \tag{A.44}$$

that the L_∞-norm cone invertibility factor F'_∞ with $\mathbf{\Sigma}$ replaced by $\mathbf{R}$ in (20) and the original one F_∞ defined for $\mathbf{\Sigma}$ are within a constant factor of each other. To simplify the notation, we still write F'_∞ as F_∞ which is implicitly understood as the L_∞-norm cone invertibility factor defined for $\mathbf{R}$ hereafter.

For each node j in the index set A_l, define the population version of the L_∞-norm cone invertibility factor in (A.43) as

$$\widetilde{F}_{\infty,j,l} = \inf\left\{\frac{\|\mathbf{R}_{A_l^c}\mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty} : \|\mathbf{u}_{S_{j,l}^c}\|_1 \le \xi\|\mathbf{u}_{S_{j,l}}\|_1 \ne 0\right\}, \tag{A.45}$$

where $\mathbf{R}_{A_l^c}$ denotes the principal submatrix of $\mathbf{R}$ given by the index set A_l^c. As mentioned in the proof of Lemma 1, $|S_{j,l}| = \|\boldsymbol{\beta}_{j,l}\|_0 \le \widetilde{K} = O(K)$, which together with (20) defined for $\mathbf{R}$ and (A.45) leads to

$$\widetilde{F}_{\infty,j,l} \ge F_\infty. \tag{A.46}$$

We will show that the empirical version of the L_∞-norm cone invertibility factor $F_{\infty,j,l}$ in (A.43) concentrates around its population counterpart $\widetilde{F}_{\infty,j,l}$ in (A.45) with overwhelming probability.

Using similar arguments to those for proving (A.35) with t chosen to be $[(\delta+1)(\log p)/(cn)]^{1/2}$, we can show that

$$P(\mathcal{F}) = 1 - o\left\{p^{-(\delta-2)}\right\}, \tag{A.47}$$

where $\mathcal{F} = \{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \le O(\lambda)\}$ with $\widehat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^T\mathbf{X}$ and $\|\cdot\|_\infty$ denoting the entrywise L_∞-norm of a given matrix. Note that $\widehat{\mathbf{R}}_{j,l} = n^{-1}\mathbf{Y}_{A_l^c}^T\mathbf{Y}_{A_l^c}$ is simply the principal submatrix $\widehat{\mathbf{R}}_{A_l^c}$ of the sample correlation matrix

$$\widehat{\mathbf{R}} = \left(\mathrm{diag}\{\widehat{\mathbf{\Sigma}}\}\right)^{-1/2}\widehat{\mathbf{\Sigma}}\left(\mathrm{diag}\{\widehat{\mathbf{\Sigma}}\}\right)^{-1/2} \tag{A.48}$$

given by the index set A_l^c. By some standard calculations, we can show that on the event $\mathcal{F}$, it also holds that $\|\widehat{\mathbf{R}} - \mathbf{R}\|_\infty \le O(\lambda)$. This result together with (A.47) yields

$$P(\mathcal{F}_1) \ge 1 - o\left\{p^{-(\delta-2)}\right\}, \tag{A.49}$$

where the event $\mathcal{F}_1$ is defined as the intersection of events $\mathcal{F}$ and $\{\|\widehat{\mathbf{R}} - \mathbf{R}\|_\infty \le O(\lambda)\}$.

Finally let us do some algebraic calculations conditional on event $\mathcal{F}_1$. On this event, for each $\mathbf{u} \in \mathbb{R}^{p-|A_l|}$ satisfying $\|\mathbf{u}_{S_{j,l}^c}\|_1 \le \xi \|\mathbf{u}_{S_{j,l}}\|_1 \ne 0$ we have

$$
\begin{aligned}
\|\widehat{\mathbf{R}}_{j,l}\mathbf{u}\|_\infty = \|\widehat{\mathbf{R}}_{A_l^c}\mathbf{u}\|_\infty &\ge \|\mathbf{R}_{A_l^c}\mathbf{u}\|_\infty - \left\|\left(\widehat{\mathbf{R}}_{A_l^c} - \mathbf{R}_{A_l^c}\right)\mathbf{u}\right\|_\infty \\
&\ge \widetilde{F}_{\infty,j,l}\|\mathbf{u}\|_\infty - \|\widehat{\mathbf{R}} - \mathbf{R}\|_\infty \|\mathbf{u}\|_1 \\
&\ge \widetilde{F}_{\infty,j,l}\|\mathbf{u}\|_\infty - O(\lambda)(1+\xi)\|\mathbf{u}_{S_{j,l}}\|_1 \\
&\ge \widetilde{F}_{\infty,j,l}\|\mathbf{u}\|_\infty - O(\lambda)(1+\xi)|S_{j,l}|\|\mathbf{u}_{S_{j,l}}\|_\infty \\
&\ge \left[\widetilde{F}_{\infty,j,l} - O(K\lambda)\right]\|\mathbf{u}\|_\infty,
\end{aligned} \tag{A.50}
$$

since $|S_{j,l}| \le \widetilde{K} = O(K)$. Therefore, combining (A.46), (A.49)–(A.50), and the assumption of $K^{1+\alpha}\lambda = o(1)$ yields $F_{\infty,j,l} \ge c_1 F_\infty$ for some positive constant $c_1 < 1$, uniformly over all nodes j in the index sets A_l with $1 \le l \le L$.

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