Submitted to the Annals of Applied Statistics
June 22, 2011 1
arXiv: math.PR/0000000

SUPPLEMENTAL MATERIALS FOR "A SPARSE CONDITIONAL GAUSSIAN GRAPHICAL MODEL FOR ANALYSIS OF GENETICAL GENOMICS DATA"

By Jianxin Yin and Hongzhe Li

University of Pennsylvania School of Medicine

1. Summary of the Supplemental Materials. This supplemental material includes:

1. Tables of the standard errors of simulation results presented in Table 1 and Table 2.
2. Propositions on Hessian matrix and the convergence of the algorithm.
3. Asymptotic results when p and q are fixed when $n \to \infty$.
4. Rates of convergence of the estimates and sparsistency as p_n and q_n diverge.
5. Proofs of Lemmas and Theorems.

2. Tables of Standard Errors of Simulation Results.

TABLE 1

Comparison of the performances of the cGGM, adaptive cGGM (acGGM), graphical Lasso (glasso), adaptive graphical Lasso (aglasso) and a modified neighborhood selection procedure using multiple Lasso (mLasso) for Models 1 - 3 when $p < n$ based on 50 replications. For each measurement, standard deviation is given based on 50 replications.

	Estimation of Θ					Neighbor Selection			
Method	LOSS	$\|\Delta\|_\infty$	$\|\|\|\Delta\|\|\|_\infty$	$\|\Delta\|$	$\|\Delta\|_F$	DIST	SPE	SEN	MCC
				Model 1					
cGGM	0.5758	0.0253	0.0562	0.0164	0.0479	18.0398	0.0016	0.0371	0.0308
a*cGGM*	0.5903	0.0216	0.0716	0.0192	0.0609	19.5074	0.0018	0.0333	0.0315
glasso	0.2457	0.0155	0.0523	0.0111	0.0358	24.5960	0.0025	0.0309	0.0269
aglasso	0.2635	0.0182	0.0660	0.0134	0.0433	20.0107	0.0018	0.0327	0.0274
mLasso	-	-	-	-	-	15.3984	0.0013	0.0296	0.0289
				Model 2					
cGGM	0.4046	0.0375	0.0906	0.0208	0.0625	13.5234	0.0051	0.0521	0.0420
a*cGGM*	0.4594	0.0288	0.0835	0.0269	0.0784	10.6138	0.0032	0.0404	0.0376
glasso	0.2139	0.0111	0.0631	0.0089	0.0225	15.7731	0.0063	0.0456	0.0315
aglasso	0.2463	0.0133	0.0744	0.0094	0.0360	14.8530	0.0063	0.0504	0.0317
mLasso	-	-	-	-	-	13.1962	0.0053	0.0331	0.0286
				Model 3					
cGGM	0.2937	0.0343	0.0737	0.0309	0.0620	9.2159	0.0190	0.0681	0.0483
a*cGGM*	0.2782	0.0347	0.0841	0.0357	0.0652	8.4340	0.0138	0.0627	0.0528
glasso	0.1335	0.0211	0.0543	0.0173	0.0274	13.7787	0.0253	0.0516	0.0423
aglasso	0.1420	0.0222	0.0611	0.0183	0.0311	12.8987	0.0235	0.0559	0.0433
mLasso	-	-	-	-	-	8.3911	0.0147	0.0485	0.0401

TABLE 2

Comparison of the performances of the cGGM, the graphical Lasso (glasso) and a modified neighbor selection procedure using multiple Lasso (mLasso) for Model 4∼ Model 6 when $p > n$ based on 50 replications. For each measurement, standard deviation is given based on 50 replications.

	Estimation of Θ					Neighbor Selection			
Method	LOSS	$\\|\Delta\\|_\infty$	$\|\|\|\Delta\|\|\|_\infty$	$\\|\Delta\\|$	$\\|\Delta\\|_F$	DIST	SPE	SEN	MCC
				Model 4					
cGGM	2.0425	0.0368	0.0958	0.0398	0.0445	65.4080	1e-04	0.0098	0.0149
glasso	0.6635	0.0092	0.0803	0.0066	0.0237	168.0979	2e-04	0.0055	0.0021
mLasso	-	-	-	-	-	39.5748	4e-05	0.0050	0.0065
				Model 5					
cGGM	1.7833	0.0437	0.2562	0.0570	0.0639	118.0631	2e-04	0.0093	0.0231
glasso	0.6239	0.0087	0.0689	0.0088	0.0312	183.4719	3e-04	0.0056	0.0020
mLasso	-	-	-	-	-	160.3108	3e-04	0.0094	0.0030
				Model 6					
cGGM	0.7571	0.0436	0.0425	0.0268	0.0369	14.2789	4e-05	0.0063	0.0148
glasso	0.4139	0.0106	0.0901	0.008	0.0285	77.9837	5e-04	0.0043	0.0023
mLasso	-	-	-	-	-	35.9707	2e-04	0.0063	0.0082

3. Propositions on Hessian Matrix and Convergence of the Algorithm. We first study the log-likelihood function without the penalty term as a function of the matrix parameter $\Xi = (\Theta, \Gamma)$. Denote

$$l(\Xi) = -loglik = -\log\det\Theta + tr\Big[\mathbf{C}_Y\Theta - \mathbf{C}_{YX}\Gamma'\Theta - \Gamma\mathbf{C}'_{YX}\Theta + \Gamma\mathbf{C}_X\Gamma'\Theta\Big].$$

PROPOSITION 1. *The Hessian matrix of the negative log-likelihood function $l(\Xi)$ is*

$$Hl(\Xi) = \begin{pmatrix} \Theta^{-1}\otimes\Theta^{-1} & -2\mathbf{C}_{YX}\otimes I_p + 2(\Gamma\mathbf{C}_X)\otimes I_p \\ -2\mathbf{C}'_{YX}\otimes I_p + 2(\mathbf{C}_X\Gamma')\otimes I_p & 2\mathbf{C}_X\otimes\Theta \end{pmatrix}.$$

In addition, $l(\Xi)$ is a bi-convex function of Γ and Θ. In words, this means that for any fixed Θ, $l(\Xi)$ is a convex function of Γ, and for any Γ, $l(\Xi)$ is a convex function of Θ.

Proof: It is easy to check that the first and the second derivatives of $l(\Xi)$ are

$$\begin{aligned} dl &= -\frac{1}{\det\Theta}\det\Theta tr(\Theta^{-1}d\Theta) + tr(\mathbf{C}_Y d\Theta) - 2tr(\mathbf{C}'_{YX}\Theta d\Gamma) \\ &\quad -2tr(\mathbf{C}_{YX}\Gamma' d\Theta) + 2tr(\mathbf{C}_X\Gamma'\Theta d\Gamma) + tr(\Gamma\mathbf{C}_X\Gamma' d\Theta) \\ &= -tr(\Theta^{-1}d\Theta) + tr(\mathbf{C}_Y d\Theta) - 2tr(\mathbf{C}'_{YX}\Theta d\Gamma) - 2tr(\mathbf{C}_{YX}\Gamma' d\Theta) \\ &\quad +2tr(\mathbf{C}_X\Gamma'\Theta d\Gamma) + tr(\Gamma\mathbf{C}_X\Gamma' d\Theta), \end{aligned}$$

and

$$\begin{aligned} d^2 l &= tr(\Theta^{-1}d\Theta\Theta^{-1}d\Theta) - 4tr(\mathbf{C}'_{YX}d\Theta d\Gamma) + 2tr(\mathbf{C}_X d\Gamma'\Theta d\Gamma) \\ &\quad +3tr(\mathbf{C}_X\Gamma' d\Theta d\Gamma) + tr(\Gamma\mathbf{C}_X d\Gamma' d\Theta). \end{aligned} \tag{1}$$

Denote

$$vec(\Xi) = \begin{pmatrix} vec(\Theta) \\ vec(\Gamma) \end{pmatrix},$$

based on the fact that $trABCD = (vec\ B')'(A'\otimes C)vecD$, we have

$$\begin{aligned} tr(\Theta^{-1}d\Theta\Theta^{-1}d\Theta) &= (dvec\Theta)'(\Theta^{-1}\otimes\Theta^{-1})dvec\Theta, \\ tr(\mathbf{C}'_{YX}d\Theta d\Gamma) &= (dvec\Theta)'(\mathbf{C}_{YX}\otimes I_p)dvec\Gamma, \\ tr(\mathbf{C}_X d\Gamma'\Theta d\Gamma) &= (dvec\Gamma)'(\mathbf{C}_X\otimes\Theta)dvec\Gamma \\ tr(\mathbf{C}_X\Gamma' d\Theta d\Gamma) &= (dvec\Theta)'(\Gamma\mathbf{C}_X\otimes I_p)dvec\Gamma, \\ tr(\Gamma\mathbf{C}_X d\Gamma' d\Theta) &= (dvec\Gamma)'(\mathbf{C}_X\Gamma'\otimes I_p)dvec\Theta. \end{aligned}$$

From the last two equalities, we have $tr(\mathbf{C}_X\Gamma' d\Theta d\Gamma) = tr(\Gamma\mathbf{C}_X d\Gamma' d\Theta)$. Substituting these equations into (1), we obtain

$$\begin{aligned} d^2 l &= (dvec\Theta)'(\Theta^{-1}\otimes\Theta^{-1})dvec\Theta - 4(dvec\Theta)'(\mathbf{C}_{YX}\otimes I_p)dvec\Gamma \\ &\quad +2(dvec\Gamma)'(\mathbf{C}_X\otimes\Theta)dvec\Gamma + 4(dvec\Theta)'(\Gamma\mathbf{C}_X\otimes I_p)dvec\Gamma \\ &= (dvec\Theta)'(\Theta^{-1}\otimes\Theta^{-1})dvec\Theta - 2(dvec\Theta)'(\mathbf{C}_{YX}\otimes I_p)dvec\Gamma \\ &\quad +2(dvec\Theta)'(\Gamma\mathbf{C}_X\otimes I_p)dvec\Gamma - 2(dvec\Gamma)'(\mathbf{C}'_{YX}\otimes I_p)dvec\Theta \\ &\quad +2(dvec\Gamma)'(\mathbf{C}_X\Gamma'\otimes\Theta)dvec\Theta + 2(dvec\Gamma)'(\mathbf{C}_X\otimes\Theta)dvec\Gamma \\ &= ((dvec\Theta)', (dvec\Gamma)')\begin{pmatrix} \Theta^{-1}\otimes\Theta^{-1} & -2\mathbf{C}_{YX}\otimes I_p + 2(\Gamma\mathbf{C}_X)\otimes I_p \\ -2\mathbf{C}'_{YX}\otimes I_p + 2(\mathbf{C}_X\Gamma')\otimes I_p & 2\mathbf{C}_X\otimes\Theta \end{pmatrix} \\ &\quad \times\begin{pmatrix} dvec\Theta \\ dvec\Gamma \end{pmatrix} \\ &= (dvec\Xi)'\begin{pmatrix} \Theta^{-1}\otimes\Theta^{-1} & -2\mathbf{C}_{YX}\otimes I_p + 2(\Gamma\mathbf{C}_X)\otimes I_p \\ -2\mathbf{C}'_{YX}\otimes I_p + 2(\mathbf{C}_X\Gamma')\otimes I_p & 2\mathbf{C}_X\otimes\Theta \end{pmatrix} dvec\Xi, \end{aligned}$$

from which the result follows.

The proof of bi-convexity is straight-forward. □

PROPOSITION 2. *The coordinate descent algorithm minimizing $pl(\Xi)$ with respective to Γ and Θ converges to a stationary point of $pl(\Xi)$.*

Proof: Notice that the $pl(\Xi)$ is differentiable and continuous in Ξ. Then, since $pl(\Xi)$ is convex in Γ with Θ fixed and in Θ with Γ fixed, minimization with respect to each gradient gives a unique coordinate-wise minimum. Thus, we have satisfied the conditions of Tseng (2001), Theorem 4.1 (c), and block coordinate-wise minimization converges to a stationary point of $pl(\Xi)$. □

4. Asymptotic Results When p and q Are Fixed. We derive the asymptotic properties of the proposed estimators that are analogous to those for the Lasso (Knight and Fu, 2000) and to those for the Gaussian graphical model (Yuan and Lin, 2007; Fan, Feng and Wu, 2009). Also, we assume that p and q are held fixed as the sample size $n \to \infty$. We assume the design matrix $\mathbf{C}_X = 1/n\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'$ is well behaved and meets the regularity conditions as in Knight and Fu (2000). Specifically, we assume $\bar{\mathbf{x}}$ exists and is finite.

THEOREM 1. *If $\sqrt{n}\rho \to \rho_0 \geq 0$, $\sqrt{n}\lambda \to \lambda_0 \geq 0$ as $n \to \infty$, the Lasso-penalized estimator for the cGGM is such that*

$$\sqrt{n}\big\{(\hat{\Theta}, \hat{\Gamma}) - (\Theta, \Gamma)\big\} \longrightarrow \text{argmin}_{U=U',V}\{\phi(U,V)\}$$

in distribution, where

$$\begin{aligned}\phi(U,V) &= tr(U\Sigma U\Sigma) + tr(\Theta V \mathbf{C}_X V') + tr\{U(A - B\Gamma' - \Gamma B')\} \\ &\quad - tr\{\Theta(VB' + BV')\} \\ &\quad + \rho_0 \sum_{i,j}\{u_{ij} sgn(\theta_{ij}) I(\theta_{ij} \neq 0) + |u_{ij}| I(\theta_{ij} = 0)\} \\ &\quad + \lambda_0 \sum_{i,j}\{v_{ij} sgn(\gamma_{ij}) I(\gamma_{ij} \neq 0) + |v_{ij}| I(\gamma_{ij} = 0)\},\end{aligned}$$

in which A is a random symmetric $p \times p$ matrix such that $vec(A) \sim \mathcal{N}(0, \Lambda_Q)$ and Λ_Q is such that

$$cov(a_{ij}, a_{i',j'}) = cov(\mathbf{y}_i\mathbf{y}_j, \mathbf{y}_{i'}\mathbf{y}_{j'}|\mathbf{X});$$

and B is a random $p \times q$ matrix such that $vec(B) \sim \mathcal{N}(0, \Lambda_P)$ and Λ_P is such that

$$cov(b_{ij}, b_{i',j'}) = \Sigma_{i,i'}(\mathbf{C}_X)_{j,j'}.$$

Furthermore,

$$cov(a_{ij}, b_{i',j'}) = cov(\mathbf{y}_i\mathbf{y}_j, \mathbf{y}_{i'}|\mathbf{X})\bar{\mathbf{x}}_{j'}.$$

Proof: Define

$$\begin{aligned}\phi_n(U,V) &= -\log\det(\Theta + \frac{U}{\sqrt{n}}) \\ &\quad + tr\Big\{(\Theta + \frac{U}{\sqrt{n}})\big[\mathbf{C}_Y - \mathbf{C}_{YX}(\Gamma + \frac{V}{\sqrt{n}})' - (\Gamma + \frac{V}{\sqrt{n}})\mathbf{C}'_{YX} \\ &\quad + (\Gamma + \frac{V}{\sqrt{n}})\mathbf{C}_X(\Gamma + \frac{V}{\sqrt{n}})'\big]\Big\} \\ &\quad + \rho \sum_{i,j} |\theta_{ij} + \frac{u_{ij}}{\sqrt{n}}| + \lambda \sum_{i,j} |\gamma_{ij} + \frac{v_{ij}}{\sqrt{n}}| \\ &\quad + \log\det(\Theta) - tr\Big\{\Theta\big[\mathbf{C}_Y - \mathbf{C}_{YX}\Gamma' - \Gamma\mathbf{C}'_{YX} + \Gamma\mathbf{C}_X\Gamma'\big]\Big\} \\ &\quad - \rho\sum_{i,j}|\theta_{ij}| - \lambda\sum_{i,j}|\gamma_{ij}|.\end{aligned}$$

Using the same argument as in Yuan and Lin (2007), one can show that

$$\log\det(\Theta+\frac{U}{\sqrt{n}})-\log\det(\Theta)=\frac{1}{\sqrt{n}}tr(U\Sigma)-\frac{1}{n}tr(U\Sigma U\Sigma)+o(\frac{1}{n}).$$

On the other hand,

$$\begin{aligned}
&tr\Big\{(\Theta+\frac{U}{\sqrt{n}})[\mathbf{C}_Y-\mathbf{C}_{YX}(\Gamma+\frac{V}{\sqrt{n}})'-(\Gamma+\frac{V}{\sqrt{n}})\mathbf{C}'_{YX}\\
&+(\Gamma+\frac{V}{\sqrt{n}})\mathbf{C}_X(\Gamma+\frac{V}{\sqrt{n}})']\Big\}-tr\Big\{\Theta(\mathbf{C}_Y-\mathbf{C}_{YX}\Gamma'-\Gamma\mathbf{C}'_{YX}+\Gamma\mathbf{C}_X\Gamma')\Big\}\\
(2)\ &=\frac{1}{\sqrt{n}}tr\Big\{\Theta[V\mathbf{C}_X\Gamma'-V\mathbf{C}'_{YX}+\Gamma\mathbf{C}_XV'-\mathbf{C}_{YX}V']\Big\}\\
&+\frac{1}{\sqrt{n}}tr\Big\{U[\mathbf{C}_Y-\mathbf{C}_{YX}\Gamma'-\Gamma\mathbf{C}'_{YX}+\Gamma\mathbf{C}_X\Gamma']\Big\}\\
&+\frac{1}{n}tr\{\Theta V\mathbf{C}_XV'\}+\frac{1}{n}tr\Big\{U[V\mathbf{C}_X\Gamma'-V\mathbf{C}'_{YX}+\Gamma\mathbf{C}_XV'-\mathbf{C}_{YX}V']\Big\}.\\
&+\frac{1}{n\sqrt{n}}tr\{UV\mathbf{C}_XV'\}
\end{aligned}$$

Denote $P_n=\mathbf{C}_{YX}-\Gamma\mathbf{C}_X$, then $B_n=\sqrt{n}P_n\to B$, where $B\sim\mathcal{N}(0,\Lambda_P)$ with $\mathrm{cov}(b_{ij},b_{i'j'})=\Sigma_{i,i'}(\mathbf{C}_X)_{j,j'}$. Similarly, denote $Q_n=\mathbf{C}_Y-(\Sigma+\Gamma\mathbf{C}_X\Gamma')$, then $A_n=\sqrt{n}Q_n\to A$, where $A\sim\mathcal{N}(0,\Lambda_Q)$ with

$$\mathrm{cov}(a_{ij},a_{i'j'})=\mathrm{cov}(Y^{(i)}Y^{(j)},Y^{(i')}Y^{(j')}|X).$$

It is easy to check that $\mathrm{cov}(a_{ij},b_{i'j'})=\mathrm{cov}(Y^{(i)}Y^{(j)},Y^{(i')}|X)\overline{x}_{j'}$. Then (2) can be rewritten as

$$\begin{aligned}
(2)\quad=\quad&-\frac{1}{\sqrt{n}}tr\{\Theta(VP'_n+P_nV')\}+\frac{1}{\sqrt{n}}trU\Sigma\\
&+\frac{1}{\sqrt{n}}tr\{U[Q_n-P_n\Gamma'-\Gamma P'_n]\}\\
&+\frac{1}{n}tr\{\Theta V\mathbf{C}_XV'\}-\frac{1}{n}tr\{U[VP'_n+P_nV']\}\\
&+\frac{1}{n\sqrt{n}}tr\{UV\mathbf{C}_XV'\}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\rho\sum_{ij}\big(|\theta_{ij}+\frac{u_{ij}}{\sqrt{n}}|-|\theta_{ij}|\big)=\frac{\rho}{\sqrt{n}}\sum_{ij}\{&u_{ij}sgn(\theta_{ij})I(\theta_{ij}\neq 0)\\
&+|u_{ij}|I(\theta_{ij}=0)\},
\end{aligned}$$

$$\lambda\sum_{ij}(|\gamma_{ij}+\frac{v_{ij}}{\sqrt{n}}|-|\gamma_{ij}|)=\frac{\lambda}{\sqrt{n}}\sum_{ij}\big\{v_{ij}sgn(\gamma_{ij})I(\gamma_{ij}\neq 0)$$
$$+|v_{ij}|I(\gamma_{ij}=0)\big\}.$$

So $n\phi_n(U,V)$ can be rewritten as

$$\begin{aligned} n\phi_n(U,V) &= tr\{U\Sigma U\Sigma\}+tr\{\Theta V\mathbf{C}_X V'\}+tr\big\{U[A_n-B_n\Gamma'-\Gamma B_n']\big\} \\ &\quad -tr\big\{\Theta[VB_n'+B_nV']\big\} \\ &\quad +\sqrt{n}\rho\sum_{ij}\big\{u_{ij}sgn(\theta_{ij})I(\theta_{ij}\neq 0)+|u_{ij}|I(\theta_{ij}=0)\big\} \\ &\quad +\sqrt{n}\lambda\sum_{ij}\big\{v_{ij}sgn(\gamma_{ij})I(\gamma_{ij}\neq 0)+|v_{ij}|I(\gamma_{ij}=0)\big\}+o(1), \end{aligned}$$

where $A_n=\sqrt{n}Q_n=\sqrt{n}(\mathbf{C}_Y-\Sigma-\Gamma\mathbf{C}_X\Gamma')$, $B_n=\sqrt{n}P_n=\sqrt{n}(\mathbf{C}_{YX}-\Gamma\mathbf{C}_X)$ and $A_n\to\mathcal{N}(0,\Lambda_Q)$, $B_n\to\mathcal{N}(0,\Lambda_P)$ as defined before. Therefore, $n\phi_n(U,V)\to\phi(U,V)$ in distribution. Since both $\phi(U,V)$ and $n\phi_n(U,V)$ are convex and $\phi(U,V)$ has a unique minimum, it follows that

$$\text{argmin } n\phi_n(U,V)=\sqrt{n}(\hat{\Theta}-\Theta,\hat{\Gamma}-\Gamma)\to\text{argmin } \phi(U,V).$$

□

The next theorem shows that when the adaptive Lasso penalty function is used for the means and also the concentration matrix and when the tuning parameters are chosen appropriately, the resulting estimates are consistent and have the oracle properties in the sense of Fan and Li (2001).

THEOREM 2. *Let $(\hat{\Theta},\hat{\Gamma})$ be the maximizer of (8) with initial consistent estimator $(\tilde{\Theta},\tilde{\Gamma})$ and adaptive penalty $pen(x)=|x|/|\tilde{x}|^{\gamma}$ for some $\gamma>0$. If $n^{(1+\gamma)/2}\rho\to\infty$, $n^{(1+\gamma)/2}\lambda\to\infty$, $\sqrt{n}\rho\to 0$ and $\sqrt{n}\lambda\to 0$ as $n\to\infty$, then $Pr(\hat{\theta}_{ij}=0)\to 1$ if $\theta_{ij}=0$, $Pr(\hat{\gamma}_{ij}=0)\to 1$ if $\gamma_{ij}=0$ and other elements of $\hat{\Theta}$ and $\hat{\Gamma}$ have the same limiting distribution as the maximum likelihood estimate based on the true means and the true graphical structure.*

Proof: We define $\phi_n(U,V)$ as

$$\begin{aligned}\phi_n(U,V) &= -\log\det(\Theta+\frac{U}{\sqrt{n}})+tr\big\{(\Theta+\frac{U}{\sqrt{n}})\big[\mathbf{C}_Y-\mathbf{C}_{YX}(\Gamma+\frac{V}{\sqrt{n}})'\\ &\quad -(\Gamma+\frac{V}{\sqrt{n}})\mathbf{C}_{YX}'+(\Gamma+\frac{V}{\sqrt{n}})\mathbf{C}_X(\Gamma+\frac{V}{\sqrt{n}})'\big]\big\}+\rho\sum_{ij}\frac{1}{|\tilde{\theta}_{ij}|^{\gamma}}|\theta_{ij}+\frac{u_{ij}}{\sqrt{n}}|\\ &\quad +\lambda\sum_{ij}\frac{1}{|\tilde{\Gamma}_{ij}|^{\gamma}}|\gamma_{ij}+\frac{v_{ij}}{\sqrt{n}}|+\log\det(\Theta)\\ &\quad -tr\big\{\Theta\big[\mathbf{C}_Y-\mathbf{C}_{YX}\Gamma'-\Gamma\mathbf{C}_{YX}'+\Gamma\mathbf{C}_X\Gamma'\big]\big\}\\ &\quad -\rho\sum_{ij}\frac{|\theta_{ij}|}{|\tilde{\theta}_{ij}|^{\gamma}}-\lambda\sum_{ij}\frac{|\gamma_{ij}|}{|\tilde{\Gamma}_{ij}|^{\gamma}}.\end{aligned}$$

Similar to proof of Theorem 1, we have

$$\begin{aligned}&n\phi_n(U,V)=tr\{U\Sigma U\Sigma\}+tr\{\Theta V\mathbf{C}_X V'\}+tr\{U(A_n-B_n\Gamma'-\Gamma B_n')\}\\ &-tr\{\Theta(VB_n'+B_nV')\}+n\rho\sum_{ij}\frac{1}{|\tilde{\theta}_{ij}|^{\gamma}}\big[|\theta_{ij}+\frac{u_{ij}}{\sqrt{n}}|-|\theta_{ij}|\big]\\ &+n\lambda\sum_{ij}\frac{1}{|\tilde{\Gamma}_{ij}|^{\gamma}}\big[|\gamma_{ij}+\frac{v_{ij}}{\sqrt{n}}|-|\gamma_{ij}|\big]+o(1).\end{aligned}$$

Note that $\tilde{\theta}_{ij}=O_p(n^{-1/2})$ if $\theta_{ij}=0$, $\tilde{\Gamma}_{ij}=O_p(n^{-1/2})$ if $\gamma_{ij}=0$, $\tilde{\theta}_{ij}\to\theta_{ij}$ in probability, $\tilde{\Gamma}_{ij}\to\gamma_{ij}$ in probability, $\sqrt{n}\rho\to 0$ and $\sqrt{n}\lambda\to 0$.

Therefore, if $\theta_{ij}\neq 0$,

$$\sqrt{n}\big(|\theta_{ij}+\frac{u_{ij}}{\sqrt{n}}|-|\theta_{ij}|\big)\to u_{ij}sgn(\theta_{ij}),$$

then by Slutsky's theorem we have $\sqrt{n}\rho|\tilde{\theta}_{ij}|^{-\gamma}\sqrt{n}(|\theta_{ij}+u_{ij}/\sqrt{n}|-|\theta_{ij}|)\to 0$. If $\theta_{ij}=0$,

$$\sqrt{n}(|\theta_{ij}+\frac{u_{ij}}{\sqrt{n}}|-|\theta_{ij}|)=|u_{ij}|,$$

and $\sqrt{n}\rho|\tilde{\theta}_{ij}|^{-\gamma}=\sqrt{n}\rho n^{\gamma/2}(\sqrt{n}|\tilde{\theta}_{ij}|)^{-\gamma}$, where $\sqrt{n}|\tilde{\theta}_{ij}|=O_p(1)$. By the Slutsky's theorem, we have $\sqrt{n}\rho|\tilde{\theta}_{ij}|^{-\gamma}\sqrt{n}(|\theta_{ij}+u_{ij}/\sqrt{n}|-|\theta_{ij}|)\to\infty$. Similar results hold for the corresponding term of γ_{ij}. By the Slutsky's theorem, we have that $n\phi_n(U,V)\to_d\phi(U,V)$ for every (U,V), where

$$\phi(U,V)=trU\Sigma U\Sigma+tr\Theta V\mathbf{C}_XV'+tr\{U(A-B\Gamma'-\Gamma B')\}-tr\{\Theta(VB'+BV')\}$$

if $U=U'$ such that $u_{ij}=0$ if $\theta_{ij}=0$, and $v_{ij}=0$, if $\gamma_{ij}=0$, and $\phi(U,V)=\infty$, otherwise. Since ϕ is convex, the minimizer of ϕ satisfies

$u_{ij} = 0$ if $\theta_{ij} = 0$, $v_{ij} = 0$ if $\gamma_{ij} = 0$. The proof is now completed if we note that the maximum likelihood estimator $(\hat{\Theta}, \hat{\Gamma})$ for the true graph and true mean model $(V, E = (\theta_{ij} \neq 0)), \Gamma = (\gamma_{ij} \neq 0)$ is such that

$$\sqrt{n}\big((\hat{\Theta}^{true}, \hat{\Gamma}^{true}) - (\Theta, \Gamma)\big) \longrightarrow \text{argmin}\Big\{tr\{U\Sigma U\Sigma\} + tr\{\Theta V \mathbf{C}_X V'\}$$

$$+tr\big\{U(A - B\Gamma' - \Gamma B')\big\} - tr\big\{\Theta(VB' + BV')\big\}\Big\}$$

in distribution, where the minimum is taken over all symmetric matrices U such that $u_{ij} = 0$ if $\theta_{ij} = 0$ and $p \times q$ matrices V such that $v_{ij} = 0$ if $\gamma_{ij} = 0$. □

5. Main Theoretical Results and Assumptions when p_n and q_n Diverge. We denote the support of the true concentration matrix $\Sigma_0^{-1} = \Theta_0 = (\theta_{ij})$ as $S = \{(i,j) : i \neq j,\ \theta_{ij} \neq 0\}$ and the support of the true regression coefficient matrix $\Gamma_0 = (\gamma_{ij})$ as $T = \{(i,j) : \gamma_{ij} \neq 0\}$. Let $s_n = |S|$ and $k_n = |T|$ be the cardinality of these two supports. Denote $\lambda_{min}(A)$ and $\lambda_{max}(A)$ as the minimum and maximum eigenvalues of a matrix A. Define $\|A\|_F = \sqrt{\text{tr}(A^T A)}$ as the Frobenius norm of a matrix A. In order to establish the asymptotic properties of the estimates, we assume the following regularity conditions: for all n,

(A) There exist constants τ_1 and τ_2 such that

$$0 < \tau_1 < \lambda_{min}(\Sigma_0) \leq \lambda_{max}(\Sigma_0) < \tau_2 < \infty;$$

(B) There exists a constant M_1 such that for all n

$$\lambda_{\max}(\mathbf{C}_X) \leq M_1;$$

(C) There exists a constant ξ_1 such that for all n

$$0 < \xi_1 < \lambda_{min}\{(\mathbf{C}_X \otimes \Theta_0)_{T,T}\};$$

(D) There exists a constant $C_0 > 0$, such that for all n,

$$\lambda_{\max}\Big((\mathbf{C}_X \otimes \Theta_0)_{T^c,T}\big\{(\mathbf{C}_X \otimes \Theta_0)_{T,T}\big\}^{-1}(\mathbf{C}_X \otimes \Theta_0)_{T,T^c}\Big) \leq C_0.$$

Condition (A) is the same as the condition (A) in Lam and Fan (2009), which bounds uniformly the eigenvalues of Σ_0. Condition (B) guarantees that the design matrix $\mathbf{X}$ is appropriately behaved. Since we essentially treat the Γ matrix as a vector, condition (C) is similar to the conditions (26b) and (26c) of Wainwright (2009). Condition (D) imposes a requirement on the partition of the Hessian matrix corresponding to Γ between the relevant and irrelevant covariates.

The following theorem establishes the rate of convergence of the penalized likelihood estimates $\hat{\Theta}$ and $\hat{\Gamma}$ in Frobenius norm.

THEOREM 3. (Rate of convergence) *Assume* $q_n/p_n \to c^* < 1$. *Under the regularity conditions (A)-(D), if* $\log p_n/n = O(\rho_n^2)$, $(\log p_n + \log q_n)/n = O(\lambda_n^2)$, $(p_n + s_n)(\log p_n)^k/n = O(1)$, $k_n(\log p_n + \log q_n)^l/n = O(1)$ *for some* $k > 1$ *and* $l > 1$, *and* $k_n = O(s_n + p_n)$, *there exists a local minimizer* $(\hat{\Theta}, \hat{\Gamma})$ *of (8) with the Lasso penalty such that* $\|\hat{\Theta} - \Theta\|_F^2 = O_P\{(p_n + s_n)\log p_n/n\}$ *and* $\|\hat{\Gamma} - \Gamma\|_F^2 = O_P\{k_n(\log p_n + \log q_n)/n\}$.

Note that the asymptotic bias for $\hat{\Theta}$ is at the same rate as Lam and Fan (2009) for sparse *GGM*s, which is $(p_n+s_n)/n$ multiplied by a logarithm factor $\log p_n$, and goes to zero as long as $(p_n + s_n)/n$ is at a rate of $O\{(\log p_n)^{-k}\}$ with some $k > 1$. The total square errors for $\hat{\Gamma}$ are at least of rate k_n/n since each of the k_n nonzero elements can be estimated with rate $n^{-1/2}$. The price we pay for high-dimensionality is a logarithmic factor $\log(p_n q_n)$. The estimate $\hat{\Gamma}$ is consistent as long as k_n/n is at a rate of $O\{(\log p_n + \log q_n)^{-l}\}$ with some $l > 1$. Here, we refer to the local minimizer as an interior point within a given closed set such that it minimizes the target function. If the Hessian matrix $Hl(\Xi)$ is positive definite, the local minimizer becomes the global minimizer.

To establish the sparsistency of the penalized estimators, we denote $\|A\| = \max\{||Ax||/||x||, x \in R^p, x \neq 0\}$ as the operator norm of a matrix A, $\|A\|_\infty$ as the element-wise l_∞ norm of a matrix A, and $|||A|||_\infty = \max_{1\le i\le p}\sum_{j=1}^q |a_{ij}|$ for $A = (a_{ij})_{p\times q}$ as the matrix l_∞ norm of a matrix A. By sparsistency, we mean the property that all parameters that are zero are actually estimated as zero with probability tending to one (Lam and Fan, 2009). We have the following theorem on sparsistency of our penalized estimates of Γ and Θ when the adaptive Lasso penalty functions are used.

THEOREM 4. (Sparsistency) *Under the conditions given in Theorem 3, and adaptive Lasso penalty* $pen_1(\gamma_{ij}) = |\gamma_{ij}|/|\tilde{\gamma}_{ij}|^{\eta_1}$, $pen_2(\theta_{kl}) = |\theta_{kl}|/|\tilde{\theta}_{kl}|^{\eta_2}$, *for some* $\eta_1 > 0$, $\eta_2 > 0$, *where* $\tilde{\Gamma} = (\tilde{\gamma}_{ij})$ *and* $\tilde{\Theta} = (\tilde{\theta}_{kl})$ *are any two* e_n*- and* f_n*-consistent estimator, i.e.,* $e_n\|\tilde{\Gamma} - \Gamma\|_\infty = O_P(1)$, $f_n\|\tilde{\Theta} - \Theta\|_\infty = O_P(1)$. *For any local maximizer of (8)* $(\hat{\Theta}, \hat{\Gamma})$ *satisfying* $\|\hat{\Theta} - \Theta\|_F^2 = O_P\{(p_n + s_n)\log p_n/n\}$, $\|\hat{\Theta} - \Theta\|^2 = O_P(a_n)$, $\|\hat{\Gamma} - \Gamma\|_F^2 = O_P\{k_n(\log p_n + \log q_n)/n\}$ *and* $|||\hat{\Gamma}-\Gamma|||_\infty^2 = O_P(c_n)$ *for sequences* $a_n \to 0$ *and* $c_n \to 0$, *if* $e_n^{-2\eta_1}k_n(\log p_n+$

$\log q_n)/n = O(\lambda_n^2)$ and $f_n^{-2\eta_2}\{\log p_n/n + a_n + 2c_n k_n(\log p_n + \log q_n)/n\} = O(\rho_n^2)$, then with probability tending to 1, $\hat{\theta}_{ij} = 0$ for all $(i,j) \in S^c$ and $\hat{\gamma}_{kl} = 0$ for all $(k,l) \in T^c$.

Note that to obtain the sparsistency results of the estimates, we require a certain rate of convergence for $\hat{\Theta}$ in both Frobenius and spectral norms and a certain rate for $\hat{\Gamma}$ in both Frobenius and matrix l_∞ norms. The lower bound for the tuning parameter λ_n for $\hat{\Gamma}$'s is related to e_n, η_1 and k_n, while the lower bound for tuning parameter ρ_n for $\hat{\Theta}$'s is related to not only f_n, η_2 and a_n, but also c_n and k_n. This reflects the fact that the sparsistency result on the concentration matrix can be affected by the estimates of the regression parameters.

6. Proofs of The Theorems.

6.1. *Proof of Lemmas.*

LEMMA 1. *Under regularity conditions (A) and (B), let* $\mathbf{W}$, $\mathbf{X}$ *be defined as in the main text. Then*

$$\max_{1\le i\le p_n, 1\le j\le q_n} |(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X})_{ij}| = O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right).$$

Proof: It is easy to check that $\overline{\mathbf{W}'} \sim \mathrm{N}(0, I_n \otimes \Theta^{-1})$. This together with the fact that $n^{-1}\overline{\Theta\mathbf{W}'\mathbf{X}} = n^{-1}(\mathbf{X}' \otimes \Theta)\overline{\mathbf{W}'}$ leads to

$$n^{-1}\overline{\Theta\mathbf{W}'\mathbf{X}} \sim N\big(0, n^{-2}(\mathbf{X}' \otimes \Theta)(I_n \otimes \Theta^{-1})(\mathbf{X} \otimes \Theta)\big),$$

which implies that

$$y := n^{-1}\overline{\Theta\mathbf{W}'\mathbf{X}} \sim N\big(0, n^{-1}\mathbf{C}_X \otimes \Theta\big).$$

Since the regularity conditions (A) and (B) are assumed to be held, applying the Chernoff bound, there exist constants C_1 and C_2 such that $P(|y_i| > v) < C_1 \exp\{-C_2 n v^2\}$ for all $i = 1, \cdots, p_n q_n$. Then $P(\max_{1\le i\le p_n q_n} |y_i| > v) < p_n q_n C_1 \exp\{-C_2 n v^2\}$. We choose

$$v = \sqrt{\frac{\log p_n + \log q_n}{C_2 n}} M,$$

for arbitrary M, then

$$P(\max_{1\le i\le p_n q_n} |y_i| > \sqrt{\frac{\log p_n + \log q_n}{C_2 n}} M) < \frac{C_1}{(p_n q_n)^{M-1}},$$

which implies

$$\max_{1\leq i\leq p_n, 1\leq j\leq q_n} |(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X})_{ij}| = O_P(\sqrt{\frac{\log p_n + \log q_n}{n}}).$$

□

LEMMA 2. *Under regularity condition (B), let* $\mathbf{W}$, $\mathbf{X}$ *be defined as in the main text. Then*

$$\max_{1\leq i\leq p_n, 1\leq j\leq q_n} |(\frac{1}{n}\mathbf{W}'\mathbf{X})_{ij}| = O_P\left(\sqrt{\frac{\log p_n + \log q_n}{n}}\right)$$

Proof: This lemma can be proved using the same argument as the proof of Lemma 1. □

LEMMA 3. *Let* $\|\cdot\|_\infty$ *and* $|||\cdot|||_\infty$ *be the element-wise and matrix* l_∞ *norm of a matrix, then for two matrices* A *and* B*, we have*

$$\|AB\|_\infty \leq |||A|||_\infty \|B\|_\infty.$$

Proof: Observe that $|(AB)_{ij}| = |a_i' b_j| \leq \|B\|_\infty \|a_i\|_1$, where a_i is the i-th row of A while b_j is the j-th column of B. □

LEMMA 4. *Suppose* $\mathbf{W}$ *is an* $n \times p_n$ *data matrix, with each row of* $\mathbf{W}$ *being independent identically distributed multivariate normal* $N(0, \Theta^{-1})$. *Assume* $\|\Theta\|_2 \leq \tau_1^{-1}$ *for some constant* $\tau_1 > 0$. $\mathbf{X}$ *is an* $n \times q_n$ *constant matrix with a bounded* l_2 *norm in the sense that there exists a constant* $C_1 > 0$*, such that* $\|n^{-1}\mathbf{X}'\mathbf{X}\|_2 \leq C_1$. *Assume* $p_n/n = o(1)$ *and there exists the limit* $q_n/p_n \to y$*, where* $y \in (0, 1)$. *Then*

$$\|\frac{1}{n^2}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\|_2 = o_P(1),$$

and

$$\|\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\|_2 = o_P(1).$$

Proof: Since $\Theta\Theta^{-1}\Theta = \Theta$, $\mathrm{tr}\Theta\Theta^{-1} = p_n$, then by Exercise 3.4.10 in Mardia et al. (1979), $\mathbf{W}\Theta\mathbf{W}' \sim W_n(I_n, p_n)$, where $W_n(I_n, p_n)$ denotes for the Wishart distribution. Then $n^{-1}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X} \sim W_{q_n}(n^{-1}\mathbf{X}'\mathbf{X}, p_n)$. Then the distribution of $(np_n)^{-1}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}$ is the same as a sample covariance matrix of p_n observations from the $q_n \times 1$ multivariate normal distribution $N_{q_n}(0, n^{-1}\mathbf{X}'\mathbf{X})$. Since $\|n^{-1}\mathbf{X}'\mathbf{X}\|_2 \leq C_1$, using the same argument as in the proof of Lemma 3 in Lam and Fan (2009) and Theorem 5.10 in Bai and Silverstein (2006), with the condition $q_n/p_n \to y$, where $y \in (0,1)$, we have

$$\Big\|\frac{1}{np_n}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X} - \frac{1}{n}\mathbf{X}'\mathbf{X}\Big\|_2 = O_P(1)$$

so $\|(np_n)^{-1}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\|_2 \leq C_1 + O_P(1) = O_P(1)$. Since $p_n/n = o(1)$, then $\big\|n^{-2}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big\|_2 = p_n/n\big\|(np_n)^{-1}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big\|_2 = o_P(1)$. For the second part, since

$$\begin{aligned}\Big\|\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\Big\|_2 &= \Big\|\Theta^{1/2}(\frac{1}{n}\Theta^{1/2}\mathbf{W}'\mathbf{X})\Big\|_2 \\ &\leq \big\|\Theta^{1/2}\big\|_2\big\|\frac{1}{n}\Theta^{1/2}\mathbf{W}'\mathbf{X}\big\|_2 \\ &\leq \tau_1^{-1/2}\sqrt{\|\frac{1}{n^2}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\|_2} = o_P(1).\end{aligned}$$

Thus we proved the lemma. □
Note that when the conditions in Theorem 3 are satisfied, then the conditions in Lemma 4 are satisfied and can be applied.

6.2. *Proof of Theorem 3.* Similar to Lam and Fan (2009), let U be a symmetric matrix of size p_n, D_U be its diagonal matrix and $R_U = U - D_U$ be its off-diagonal matrix. Let S be the support of Θ's off-diagonal entries $(\Theta)_{i\neq j}$ and $s_n = |S|$ be its cardinality. Let V be a p-by-q matrix and T be the support of the matrix Γ as defined before and T^c be its complement $T^c = \{(i,j) : 1 \leq i \leq p, 1 \leq j \leq q\} \setminus T$. Let $k_n = |T|$ be the cardinality of T. Set $\Delta_U = \alpha_n R_U + \beta_n D_U$, $\Delta_V = \gamma_n V_T + \delta_n V_{T^c}$. We will show that, for

$$\alpha_n = \sqrt{s_n\frac{\log(p_n)}{n}}, \beta_n = \sqrt{p_n\frac{\log(p_n)}{n}}, \gamma_n = \sqrt{k_n\frac{\log(p_n)+\log(q_n)}{n}},$$

and $\delta_n = o(\gamma_n)$, and for a set $\mathscr{A}$ defined as

$$\mathscr{A} = \{(U,V) : \|\Delta_U\|_F^2 = C_1^2\alpha_n^2 + C_2^2\beta_n^2, \|(\Delta_V)_T\|_F^2 = C_3^2\gamma_n^2, \|(\Delta_V)_{T^c}\|_F^2 = C_4^2\delta_n^2\},$$

$$P\Big(\inf_{(U,V)\in\mathscr{A}} q(\Theta+\Delta_U, \Gamma+\Delta_V) > q(\Theta,\Gamma)\Big) \to 1,$$

for sufficiently large constants C_1, C_2, C_3 and C_4. This implies that there is a local minimizer in $\{(\Theta+\Delta_U, \Gamma+\Delta_V) : \|\Delta_U\|_F^2 \leq C_1^2\alpha_n^2 + C_2^2\beta_n^2, \|(\Delta_V)_T\|_F^2 =$

$C_3^2\gamma_n^2, \|(\Delta_V)_{T^c}\|_F^2 = C_4^2\delta_n^2\}$ such that $\|\hat{\Theta}-\Theta\|_F = O_P(\alpha_n+\beta_n)$, $\|\hat{\Gamma}-\Gamma\|_F = O_P(\gamma_n)$ for sufficiently large n.

In the following argument, let $S^c = \{(i,j) : i \neq j\} \setminus S$. Consider $\Theta_1 = \Theta + \Delta_U$, $\Gamma_1 = \Gamma + \Delta_V$, the difference

$$q(\Theta_1, \Gamma_1) - q(\Theta, \Gamma) = I_1 + I_2 + I_3 + I_4 + I_5,$$

where

$$\begin{aligned}
I_1 &= \mathrm{tr}\{(\mathbf{C}_Y - \mathbf{C}_{YX}\Gamma_1' - \Gamma_1\mathbf{C}_{YX}' + \Gamma_1\mathbf{C}_X\Gamma_1')\Theta_1\} \\
&\quad -\mathrm{tr}\{(\mathbf{C}_Y - \mathbf{C}_{YX}\Gamma' - \Gamma\mathbf{C}_{YX}' + \Gamma\mathbf{C}_X\Gamma')\Theta\} \\
&\quad -\log\det(\Theta_1) + \log\det(\Theta), \\
I_2 &= \rho_n \sum_{(i,j)\in S^c} (|\theta_{ij}^1| - |\theta_{ij}|), \\
I_3 &= \rho_n \sum_{(i,j)\in S} (|\theta_{ij}^1| - |\theta_{ij}|), \\
I_4 &= \lambda_n \sum_{(i,j)\in T^c} (|\gamma_{ij}^1| - |\gamma_{ij}|), \\
I_5 &= \lambda_n \sum_{(i,j)\in T} (|\gamma_{ij}^1| - |\gamma_{ij}|).
\end{aligned}$$

It is sufficient to show that the difference is positive asymptotically with probability tending to 1. Let $\mathbf{y}_i = \Gamma\mathbf{x}_i + \mathbf{w}_i$, for $i = 1, \cdots, n$ with $\mathbf{w}_i \sim \mathrm{N}(0, \Theta^{-1})$. Denote $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)'$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$. Then

$$\begin{aligned}
&\mathbf{C}_Y - \mathbf{C}_{YX}\Gamma' - \Gamma\mathbf{C}_{YX}' + \Gamma\mathbf{C}_X\Gamma' \\
=& \frac{1}{n}\sum_{i=1}^n (\mathbf{y}_i - \Gamma\mathbf{x}_i)(\mathbf{y}_i - \Gamma\mathbf{x}_i)' \\
=& \frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i' \\
:=& S_W,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{C}_{YX} - \Gamma\mathbf{C}_X &= \frac{1}{n}\sum_{i=1}^n (\mathbf{y}_i - \Gamma\mathbf{x}_i)\mathbf{x}_i' \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{w}_i\mathbf{x}_i' \\
&= \frac{1}{n}\mathbf{W}'\mathbf{X}.
\end{aligned}$$

Next denote $\bar{A}$ as the vectorization of a matrix $A = (a_1, \cdots, a_n)$, that is $\bar{A} = (a_1', \cdots, a_n')'$. Then I_1 can be simplified $I_1 = K_1 + K_2 + K_3 + K_4$, where

$$\begin{aligned} K_1 &= \text{tr}\{[S_W - \Theta^{-1}]\Delta_U\}, \\ K_2 &= -2\text{tr}\{(\frac{1}{n}\mathbf{X}'\mathbf{W})\Theta\Delta_V\}, \\ K_3 &= [\bar{\Delta}_U', \bar{\Delta}_V']\{\int_0^1 dv(1-v)\Psi_v\} \begin{pmatrix} \bar{\Delta}_U \\ \bar{\Delta}_V \end{pmatrix}, \\ K_4 &= \text{tr}\{\Delta_V \mathbf{C}_X \Delta_V' \Delta_U\}, \end{aligned}$$

and

$$\Psi_v = \begin{bmatrix} \Theta_v^{-1} \otimes \Theta_v^{-1} & -(\frac{2}{n}\mathbf{W}'\mathbf{X}) \otimes I_{p_n} \\ -(\frac{2}{n}\mathbf{X}'\mathbf{W}) \otimes I_{p_n} & 2\mathbf{C}_X \otimes \Theta_v \end{bmatrix},$$

$\Theta_v = \Theta + v\Delta_U$ and I_{p_n} is a p_n dimensional identity matrix. We can rewrite Ψ_v as

$$\begin{aligned} \Psi_v &= \begin{pmatrix} I_{p_n^2} & 0 \\ -(\frac{2}{n}\mathbf{X}'\mathbf{W}\Theta_v) \otimes \Theta_v & I_{p_n q_n} \end{pmatrix} \begin{pmatrix} \Theta_v^{-1} \otimes \Theta_v^{-1} & 0 \\ 0 & \Omega_v \end{pmatrix} \\ &\times \begin{pmatrix} I_{p_n^2} & -(\frac{2}{n}\Theta_v \mathbf{W}'\mathbf{X}) \otimes \Theta_v \\ 0 & I_{p_n q_n} \end{pmatrix}, \end{aligned}$$

where $\Omega_v = 2\mathbf{C}_X \otimes \Theta_v - \frac{4}{n^2}(\mathbf{X}'\mathbf{W}\Theta_v\mathbf{W}'\mathbf{X}) \otimes \Theta_v$. Let

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} I_{p_n^2} & -(\frac{2}{n}\Theta_v \mathbf{W}'\mathbf{X}) \otimes \Theta_v \\ 0 & I_{p_n q_n} \end{pmatrix} \begin{pmatrix} \bar{\Delta}_U \\ \bar{\Delta}_V \end{pmatrix},$$

so

$$K_3 = \int_0^1 dv(1-v)\{(y_1', y_2') \begin{pmatrix} \Theta_v^{-1} \otimes \Theta_v^{-1} & 0 \\ 0 & \Omega_v \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\}.$$

Then

$$\begin{aligned} K_3 &\geq \frac{1}{2}\min_{0\leq v\leq 1}\{\lambda_{\max}^{-2}(\Theta_v)\|y_1\|^2\} + \frac{1}{2}\min_{0\leq v\leq 1} y_2'\Omega_v y_2 \\ &\geq \frac{1}{2}(\|\Theta\| + \|\Delta_U\|)^{-2}\min_{0\leq v\leq 1}\|y_1\|^2 + \frac{1}{2}\min_{0\leq v\leq 1} y_2'\Omega_v y_2, \end{aligned} \tag{3}$$

where $\|\Theta\| \leq \tau_1^{-1}$, $\|\Delta_U\| = o(1)$ and

$$\begin{aligned} y_1 &= \bar{\Delta}_U - [(\frac{2}{n}\Theta_v \mathbf{W}'\mathbf{X}) \otimes \Theta_v]\bar{\Delta}_V \\ &= \bar{\Delta}_U - [\frac{2}{n}(\Theta\mathbf{W}'\mathbf{X}) \otimes \Theta]\bar{\Delta}_V - [\frac{2v}{n}(\Delta_U\mathbf{W}'\mathbf{X}) \otimes \Theta]\bar{\Delta}_V \\ &\quad -[\frac{2v}{n}(\Theta\mathbf{W}'\mathbf{X}) \otimes \Delta_U]\bar{\Delta}_V - [\frac{2v^2}{n}(\Delta_U\mathbf{W}'\mathbf{X}) \otimes \Delta_U]\bar{\Delta}_V. \end{aligned}$$

We then have

$$\min_{0\le v\le 1}\|y_1\| \ge \|\bar{\Delta}_U\| - \|\frac{2}{n}(\Theta\mathbf{W}'\mathbf{X})\otimes\Theta\|\|\bar{\Delta}_V\|\big(1+o_P(1)\big).$$

From Lemma 4, we know $\|n^{-1}\Theta\mathbf{W}'\mathbf{X}\| = o_P(1)$, together with $\|\Theta\| \le \tau_1^{-1}$, we have $\|2n^{-1}(\Theta\mathbf{W}'\mathbf{X})\otimes\Theta\| = o_P(1)$. From the designed framework, $\|\bar{\Delta}_V\| = O(\sqrt{k_n\frac{\log p_n+\log q_n}{n}})$, $\|\bar{\Delta}_U\| \ge \min(C_1,C_2)\sqrt{\frac{(s_n+p_n)\log p_n}{n}}$. From the conditions in theorem, we know that $q_n/p_n \to y \in (0,1)$ and $k_n = O(s_n+p_n)$. Combining together, we have $\|\bar{\Delta}_V\| = O(\|\bar{\Delta}_U\|)$. So there exists a constant $c_1^* > 0$, such that

$$\min_{0\le v\le 1}\|y_1\| \ge c_1^*\|\bar{\Delta}_U\| = c_1^*\sqrt{C_1^2\alpha_n^2 + C_2^2\beta_n^2} \tag{4}$$

Next,

$$\begin{aligned}
\Omega_v &= 2\mathbf{C}_\mathbf{X}\otimes\Theta_v - \frac{4}{n^2}\big[\mathbf{X}'\mathbf{W}\Theta_v\mathbf{W}'\mathbf{X}\big]\otimes\Theta_v\\
&= 2\mathbf{C}_\mathbf{X}\otimes\Theta + 2v\mathbf{C}_\mathbf{X}\otimes\Delta_U - \frac{4}{n^2}\big[\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big]\otimes\Theta\\
&\quad -\frac{4v}{n^2}\big[\mathbf{X}'\mathbf{W}\Delta_U\mathbf{W}'\mathbf{X}\big]\otimes\Theta - \frac{4v}{n^2}\big[\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big]\otimes\Delta_U\\
&\quad -\frac{4v^2}{n^2}\big[\mathbf{X}'\mathbf{W}\Delta_U\mathbf{W}'\mathbf{X}\big]\otimes\Delta_U.
\end{aligned}$$

So

$$\begin{aligned}
y_2'\Omega_v y_2 &= 2y_2'\big(\mathbf{C}_\mathbf{X}\otimes\Theta\big)y_2 + 2vy_2'\big(\mathbf{C}_\mathbf{X}\otimes\Delta_U\big)y_2\\
&\quad -\frac{4}{n^2}y_2'\Big\{\big[\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big]\otimes\Theta\Big\}y_2 - \frac{4v}{n^2}y_2'\Big\{\big[\mathbf{X}'\mathbf{W}\Delta_U\mathbf{W}'\mathbf{X}\big]\otimes\Theta\Big\}y_2\\
&\quad -\frac{4v}{n^2}y_2'\Big\{\big[\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\big]\otimes\Delta_U\Big\}y_2 - \frac{4v^2}{n^2}y_2'\Big\{\big[\mathbf{X}'\mathbf{W}\Delta_U\mathbf{W}'\mathbf{X}\big]\otimes\Delta_U\Big\}y_2.
\end{aligned}$$

We know $\|\mathbf{C}_\mathbf{X}\otimes\Delta_U\| = o(1)$. From Lemma 4, we have $\|n^{-2}\mathbf{X}'\mathbf{W}\Theta\mathbf{W}'\mathbf{X}\| = o_P(1)$. Similarly, the l_2 norms of the other three matrices are $o_P(1)$. Hence,

$$y_2'\Omega_v y_2 = 2y_2'\big(\mathbf{C}_\mathbf{X}\otimes\Theta\big)y_2 + o_P(\|y_2\|^2). \tag{5}$$

According to the support T of Γ, we can decompose y_2 as $y_2 = \big((y_2)_T', (y_2)_{T^c}'\big)'$ and decompose the matrix $\mathbf{C}_\mathbf{X}\otimes\Theta$ correspondingly. Then

$$\begin{aligned}
y_2'(\mathbf{C}_\mathbf{X}\otimes\Theta)y_2 &= (y_2)_T'(\mathbf{C}_\mathbf{X}\otimes\Theta)_{T,T}(y_2)_T + (y_2)_{T^c}'(\mathbf{C}_\mathbf{X}\otimes\Theta)_{T^c,T^c}(y_2)_{T^c}\\
&\quad +2(y_2)_{T^c}'(\mathbf{C}_\mathbf{X}\otimes\Theta)_{T^c,T}(y_2)_T.
\end{aligned}$$

From Cauchy-Schwarz inequality,

$$|(y_2)'_{T^c}(\mathbf{C_X}\otimes\Theta)_{T^c,T}(y_2)_T| \le \sqrt{(y_2)'_T(\mathbf{C_X}\otimes\Theta)_{T,T}(y_2)_T}\times$$

$$\sqrt{(y_2)'_{T^c}(\mathbf{C_X}\otimes\Theta)_{T^c,T}\big[(\mathbf{C_X}\otimes\Theta)_{T,T}\big]^{-1}(\mathbf{C_X}\otimes\Theta)_{T,T^c}(y_2)_{T^c}}$$

By condition (A), (B) and (D), we have

$$|(y_2)'_{T^c}(\mathbf{C_X}\otimes\Theta)_{T^c,T}(y_2)_T| \le \sqrt{C_0M_1\tau_1^{-1}}\|(y_2)_{T^c}\|\|(y_2)_T\|.$$

But from the design of $\mathcal{A}$, $\|(y_2)_{T^c}\| = o(\|(y_2)_T\|)$, hence $|(y_2)'_{T^c}(\mathbf{C_X}\otimes\Theta)_{T^c,T}(y_2)_T| = o(\|(y_2)_T\|^2)$. Then we have

$$y_2'(\mathbf{C_X}\otimes\Theta)y_2 \ge \lambda_{\min}\Big((\mathbf{C_X}\otimes\Theta)_{T,T}\Big)\|(y_2)_T\|^2 + o_P(\|(y_2)_T\|^2).$$

By condition (C) we know there exists a constant $k_1>0$, such that

$$y_2'(\mathbf{C_X}\otimes\Theta)y_2 \ge k_1\|(y_2)_T\|^2 = k_1C_3^2\gamma_n^2. \tag{6}$$

From (5), (6) we know there exists a constant $k_2>0$, such that

$$\frac{1}{2}\min_{0\le v\le 1} y_2'\Omega_v y_2 \ge k_2C_3^2\gamma_n^2. \tag{7}$$

Combining (3), (4) and (7) we have

$$K_3 \ge \frac{1}{2}(\tau_1^{-1}+o(1))c_1^{*2}(C_1^2\alpha_n^2 + C_2^2\beta_n^2) + k_2C_3^2\gamma_n^2.$$

Using a similar technique as in Lam and Fan (2009), we have $|K_2| \le M_1 + M_2$, where

$$\begin{aligned} M_1 &= 2\sum_{(i,j)\in T}|\big(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\big)_{ij}(\Delta_V)_{ij}|,\\ M_2 &= 2\sum_{(i,j)\in T^c}|\big(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\big)_{ij}(\Delta_V)_{ij}|. \end{aligned}$$

Using Lemma 1, we have

$$\begin{aligned} M_1 &\le 2\sqrt{k_n}\max_{i,j}|(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X})_{i,j}|\|(\overline{\Delta_V})_T\|\\ &\le O_P\Big(\sqrt{k_n\frac{\log p_n+\log q_n}{n}}\Big)C_3\gamma_n\\ &= O_P(C_3\gamma_n^2), \end{aligned}$$

which is dominated by K_3 when C_3 is sufficiently large.

We now consider

$$\begin{aligned} I_4 - M_2 &\geq \sum_{(i,j)\in T^c} \{\lambda_n|\gamma^1_{ij}| - |2\big(\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\big)_{ij}||\gamma^1_{ij}|\} \\ &\geq \sum_{(i,j)\in T^c} [\lambda_n - O_P\big(\sqrt{\frac{\log p_n + \log q_n}{n}}\big)]|\gamma^1_{ij}|, \end{aligned}$$

from the assumption that $(\log p_n + \log q_n)/n = O(\lambda_n^2)$, we have $I_4 - M_2 \geq 0$.

We then consider I_5,

$$\begin{aligned} I_5 &= \lambda_n \sum_{(i,j)\in T} (|\gamma^1_{ij}| - |\gamma_{ij}|) \\ &\leq \lambda_n \sum_{(i,j)\in T} |\gamma^1_{ij} - \gamma_{ij}| \\ &\leq \lambda_n\sqrt{k_n}C_3\gamma_n. \end{aligned}$$

If we choose

$$\lambda_n = M\sqrt{\frac{\log p_n + \log q_n}{n}},$$

for some constant M, then $I_5 \leq \tau C_3\gamma_n^2$ for some constant τ, which makes I_5 to be dominated by K_3 with sufficiently large C_3.

Finally, we bound K_4,

$$\begin{aligned} K_4 &= \mathrm{tr}\{\Delta_V\mathbf{C}_X\Delta_V'\Delta_U\} = (\overline{\Delta_V})'\{\mathbf{C}_X \otimes \Delta_U\}\overline{\Delta_V} \\ &\leq \lambda_{\max}(\mathbf{C}_X)\lambda_{\max}(\Delta_U)\|\overline{\Delta_V}\|^2 \\ &\leq M_1 o(1)(C_3^2\gamma_n^2 + C_4^2\delta_n^2) = o(C_3^2\gamma_n^2), \end{aligned}$$

which is dominated by K_3. Using exactly the same argument as in Lam and Fan (2009), one can show that $|K_1| \leq L_1 + L_2$ where L_1 is dominated by K_3 and $I_2 - L_2 \geq 0$, and $|I_3|$ is dominated by K_3. This completes the proof of the theorem. □

6.3. *Proof of Theorem 4.* Let (Θ_1, Γ_1) be a maximizer of

$$\max \log\det\Theta - tr(\mathbf{S}_\Gamma\Theta) - \lambda\sum_{s,t} pen_1(\gamma_{st}) - \rho\sum_{t,t'} pen_2(\theta_{tt'}). \tag{8}$$

Define

$$q_1(\Theta_1, \Gamma_1) = -\log\det\Theta_1 + \mathrm{tr}S_{\Gamma_1}\Theta_1 + \lambda_n\sum_{ij}|\gamma^1_{ij}|/|\tilde{\gamma}_{ij}|^{\eta_1} + \rho_n\sum_{k\neq l}|\theta^1_{kl}|/|\tilde{\theta}_{kl}|^{\eta_2},$$

then the derivative for $q_1(\Theta_1,\Gamma_1)$ with respect to γ^1_{kl} for $(k,l)\in T^c$ is

$$\begin{aligned}\frac{\partial q_1(\Theta_1,\Gamma_1)}{\partial \gamma^1_{kl}} &= 2\big(\Theta_1(\Gamma_1-\Gamma)\mathbf{C}_X\big)_{kl} - 2(\frac{1}{n}\Theta_1\mathbf{W}'\mathbf{X})_{kl}\\ &\quad + \frac{\lambda_n}{|\tilde\gamma_{kl}|^{\eta_1}}\mathrm{sgn}(\gamma^1_{kl}),\end{aligned} \tag{9}$$

and the derivative for $q_1(\Theta_1,\Gamma_1)$ with respect to θ^1_{ij} for $(i,j)\in S^c$ is

$$\begin{aligned}\frac{\partial q_1(\Theta_1,\Gamma_1)}{\partial \theta^1_{ij}} &= (S_{\mathbf{W}})_{ij} - [(\Gamma_1-\Gamma)(\frac{1}{n}\mathbf{X}'\mathbf{W})]_{ij} - [(\frac{1}{n}\mathbf{W}'\mathbf{X})(\Gamma_1-\Gamma)']_{ij}\\ &\quad + [(\Gamma_1-\Gamma)\mathbf{C}_X(\Gamma_1-\Gamma)]_{ij} - \sigma^1_{ij} + \frac{\rho_n}{|\tilde\theta_{ij}|^{\eta_2}}\mathrm{sgn}(\theta^1_{ij}).\end{aligned} \tag{10}$$

We show that under the conditions in the theorem, the sign of (9) depends on $\mathrm{sgn}(\gamma^1_{kl})$ only and the sign of (10) depends on $\mathrm{sgn}(\theta^1_{ij})$ only both with probability tending to 1. Then the optimum will be at 0 so that $\hat\theta_{ij}=0$ for all $(i,j)\in S^c$ and $\hat\gamma_{kl}=0$ for all $(k,l)\in T^c$ with probability tending to 1. It can be shown that

$$\begin{aligned}&|\big(\Theta_1(\Gamma_1-\Gamma)\mathbf{C}_X\big)_{kl}| \le \|(\mathbf{C}_X\otimes\Theta)(\overline{\Gamma_1-\Gamma})\|_\infty + \|\big(\mathbf{C}_X\otimes(\Theta_1-\Theta)\big)(\overline{\Gamma_1-\Gamma})\|_\infty\\ &\quad \le M_1(\tau_1^{-1})\|\Gamma_1-\Gamma\|_F + o(\|\Gamma_1-\Gamma\|_F) = O\big(\sqrt{k_n\frac{\log p_n+\log q_n}{n}}\big),\end{aligned}$$

and from Lemma 1,

$$\begin{aligned}|(\frac{1}{n}\Theta_1\mathbf{W}'\mathbf{X})_{kl}| &\le \|\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\|_\infty + \|\frac{1}{n}(\Theta_1-\Theta)\mathbf{W}'\mathbf{X}\|_\infty\\ &= \|\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\|_\infty + o(\|\frac{1}{n}\Theta\mathbf{W}'\mathbf{X}\|_\infty)\\ &\le O(\sqrt{\frac{\log p_n+\log q_n}{n}}),\end{aligned}$$

so

$$\begin{aligned}|2\big(\Theta_1(\Gamma_1-\Gamma)\mathbf{C}_X\big)_{kl} - 2(\frac{1}{n}\Theta_1\mathbf{W}'\mathbf{X})_{kl}| &\le 2\sqrt{k_n\frac{\log p_n+\log q_n}{n}}\\ &\le O(e_n^{\eta_1}\lambda_n) \le O(\frac{\lambda_n}{|\tilde\gamma_{kl}|^{\eta_1}}),\end{aligned}$$

for $(k,l)\in T^c$. In other words, the sign of (9) is dominated by the $\mathrm{sgn}(\gamma^1_{kl})$. From Lam and Fan (2009), $\max_{ij}|(S_{\mathbf{W}})_{ij}-\sigma^1_{ij}| = O_P(\{\log p_n/n\}^{1/2}+a_n^{1/2})$ and using Lemma 2 and Lemma 3, we have

$$\max_{ij}|[(\Gamma_1-\Gamma)(\frac{1}{n}\mathbf{X}'\mathbf{W})]_{ij}| = \max_{ij}|[(\frac{1}{n}\mathbf{W}'\mathbf{X})(\Gamma_1-\Gamma)']_{ij}| \le |||\Gamma_1-\Gamma|||_\infty\|\frac{1}{n}\mathbf{X}'\mathbf{W}\|_\infty$$

$$\leq \sqrt{c_n}\sqrt{\frac{\log p_n + \log q_n}{n}},$$

and

$$\max_{ij} |[(\Gamma_1 - \Gamma)\mathbf{C}_X(\Gamma_1 - \Gamma)']_{ij}| \leq |||\Gamma_1 - \Gamma|||_\infty \|\mathbf{C}_X(\Gamma_1 - \Gamma)'\|_\infty.$$

On the other hand,

$$\begin{aligned}\|\mathbf{C}_X(\Gamma_1 - \Gamma)'\|_\infty &= \|(I_{p_n} \otimes \mathbf{C}_X)\overline{(\Gamma_1 - \Gamma)}\|_\infty \\ &\leq \|(I_{p_n} \otimes \mathbf{C}_X)\overline{(\Gamma_1 - \Gamma)}\| \\ &\leq M_1\|\Gamma_1 - \Gamma\|_F = O(\sqrt{k_n\frac{\log p_n + \log q_n}{n}}).\end{aligned}$$

So

$$\begin{aligned}&|(S_{\mathbf{W}})_{ij} - [(\Gamma_1 - \Gamma)(\frac{1}{n}\mathbf{X}'\mathbf{W})]_{ij} - [(\frac{1}{n}\mathbf{W}'\mathbf{X})(\Gamma_1 - \Gamma)']_{ij} \\ &\quad + [(\Gamma_1 - \Gamma)\mathbf{C}_X(\Gamma_1 - \Gamma)]_{ij} - \sigma^1_{ij}| \\ &\leq |(S_{\mathbf{W}})_{ij} - \sigma^1_{ij}| + 2|[(\Gamma_1 - \Gamma)(\frac{1}{n}\mathbf{X}'\mathbf{W})]_{ij}| + |(\Gamma_1 - \Gamma)\mathbf{C}_X(\Gamma_1 - \Gamma)]_{ij}| \\ &\leq O_P(\sqrt{a_n} + \sqrt{\log p_n/n} + 2\sqrt{c_n}\sqrt{\frac{\log p_n + \log q_n}{n}} + \sqrt{c_n}\sqrt{k_n\frac{\log p_n + \log q_n}{n}}) \\ &= O(f_n^{\eta_2}\rho_n) \leq O(\frac{\rho_n}{|\tilde{\theta}_{ij}|^{\eta_2}}),\end{aligned}$$

for $(i, j) \in S^c$. So the sign of (10) is dominated by $\mathrm{sgn}(\theta^1_{ij})$. This completes the proof. □

7. References.

1. Bai, Z.,Jack and Silverstein, W. (2006). Spectral Analysis of Large Dimensional Random Matrices. *Science Press, Beijing.*
2. Lam, C. and Fan, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. *The Annals of Statistics*, **37**, 4254-4278.
3. Mardia, K. V. , Kent, J. T. and Bibby, J. M. (1979). Multivariate Analysis. *Academic Press.*

Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
Philadelphia, PA 19104