Multivariate Gaussian Network Structure Learning

BY XINGQI DU

Department of Statistics, North Carolina State University, 5109 SAS Hall, Campus Box 8203, Raleigh, North Carolina 27695, USA
xdu8@ncsu.edu

AND SUBHASHIS GHOSAL

Department of Statistics, North Carolina State University, 5109 SAS Hall, Campus Box 8203, Raleigh, North Carolina 27695, USA
sgosal@ncsu.edu

SUMMARY

We consider a graphical model where a multivariate normal vector is associated with each node of the underlying graph and estimate the graphical structure. We minimize a loss function obtained by regressing the vector at each node on those at the remaining ones under a group penalty. We show that the proposed estimator can be computed by a fast convex optimization algorithm. We show that as the sample size increases, the estimated regression coefficients and the correct graphical structure are correctly estimated with probability tending to one. By extensive simulations, we show the superiority of the proposed method over comparable procedures. We apply the technique on two real datasets. The first one is to identify gene and protein networks showing up in cancer cell lines, and the second one is to reveal the connections among different industries in the US.

Some key words: Graphical model; group penalty; multivariate normal; rate of convergence.

1. INTRODUCTION

Finding structural relations in a network of random variables \( (X_i : i \in V) \) is a problem of significant interest in modern statistics. The intrinsic dependence between variables in a network is appropriately described by a graphical model, where two nodes \( i, j \in V \) are connected by
an edge if and only if the two corresponding variables \(X_i\) and \(X_j\) are conditionally dependent given all other variables. If the joint distribution of all variables is multivariate normal with precision matrix \(\Omega = \left(\omega_{ij}\right)\), the conditional independence between the variable located at node \(i\) and that located at node \(j\) is equivalent of having zero at the \((i,j)\)th entry of \(\Omega\). In a relatively large network of variables, generally conditional independence is abundant, meaning that in the corresponding graph edges are sparsely present. Thus in a Gaussian graphical model, the structural relation can be learned from a sparse estimate of \(\Omega\), which can be naturally obtained by regularization method with a lasso-type penalty. Friedman et al. (2008) and Banerjee et al. (2008) proposed the graphical lasso (glasso) estimator by minimizing the sum of the negative log-likelihood and the \(\ell_1\)-norm of \(\Omega\), and its convergence property was studied by Rothman et al. (2008). A closely related method was proposed by Yuan & Lin (2007). An alternative to the graphical lasso is an approach based on regression of each variable on others, since \(\omega_{ij}\) is zero if and only if the regression coefficient \(\beta_{ij}\) of \(X_j\) in regressing \(X_i\) on other variables is zero. Equivalently this can be described as using a pseudo-likelihood obtained by multiplying one-dimensional conditional densities of \(X_i\) given \((X_j, j \neq i)\) for all \(i \in V\) instead of using the actual likelihood obtained from joint normality of \((X_i, i \in V)\). The approach is better scalable with dimension since the optimization problem is split into several optimization problems in lower dimensions. The approach was pioneered by Meinshausen & Bühlmann (2006), who imposed a lasso-type penalty on each regression problem to obtain sparse estimates of the regression coefficients, and showed that the correct edges are selected with probability tending to one. However, a major drawback of their approach is that the estimator of \(\beta_{ij}\) and that of \(\beta_{ji}\) may not be simultaneously zero (or non-zero), and hence may lead to logical inconsistency while selecting edges based on the estimated values. Peng et al. (2009) proposed the Sparse PArtial Correlation Estimation (space) by taking symmetry of the precision matrix into account. The method is shown to lead to convergence and correct edge selection with high probability, but it may be computationally challenging. A weighted version of space was considered by Khare et al. (2015), who showed that a specific choice of weights guarantees convergence of the iterative algorithm due to the convexity of the objective function in its arguments. Khare et al. (2015) named their estimator the CONvex CORrelation selection methoD (concord), and proved that the estimator inherits the theoretical convergence properties of space. By extensive simulation and numerical illustrations, they showed that concord has good accuracy for reasonable sample sizes and can be computed very efficiently.
However, in many situations, such as if multiple characteristics are measured, the variables $X_i$ at different nodes $i \in V$ may be multivariate. The methods described above apply only in the context when all variables are univariate. Even if the above methods are applied by treating each component of these variables as separate one-dimensional variables, ignoring their group structure may be undesirable, since all component variables refer to the same subject. For example, we may be interested in the connections among different industries in the US, and may like to see if the GDP of one industry has some effect on that of other industries. The data is available for 8 regions, and we want to take regions into consideration, since significant difference in relations may exist because of regional characteristics, which are not possible to capture using only national data. It seems that the only paper which addresses multi-dimensional variables in a graphical model context is Kolar et al. (2014), who pursued a likelihood based approach. In this article, we propose a method based on a pseudo-likelihood obtained from multivariate regression on other variables. We formulate a multivariate analog of concord, to be called $m_{\text{concord}}$, because of the computational advantages of concord in univariate situations. Our regression based approach appears to be more scalable than the likelihood based approach of Kolar et al. (2014), which is improper in the more complex simulation cases involving more nodes than observations. Moreover, we provide theoretical justification by studying large sample convergence properties of our proposed method, while such properties have not been established to the procedure introduced by Kolar et al. (2014).

The paper is organized as follows. Section 2 introduces the $m_{\text{concord}}$ method and describes its computational algorithm. Asymptotic properties of $m_{\text{concord}}$ are presented in Section 3. Section 4 illustrates the performance of $m_{\text{concord}}$, compared with other methods mentioned above. In Section 5, the proposed method is applied to two real data sets on gene/protein profiles and GDP respectively. Proofs are presented in Section 6 and in the appendix.

2. Method description

2.1. Model and estimation procedure

Consider a graph with $p$ nodes, where at the $i$th node there is an associated $K_i$-dimensional random variable $Y_i = (Y_{i1}, \ldots, Y_{iK_i})^T$, $i = 1, \ldots, p$. Let $Y = (Y_1^T, \ldots, Y_p^T)^T$. Assume that $Y$ has multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ijkl})$, where $\sigma_{ijkl} = \text{cov}(Y_{ik}, Y_{jl})$, $k = 1, \ldots, K_i$, $l = 1, \ldots, K_j$, $i, j = 1, \ldots, p$. Let the precision matrix $\Sigma^{-1}$ be denoted by $\Omega = (\omega_{ijkl})$, which can also be written as a block-matrix $(\Omega_{ij})$. The pri-
mary interest is in the graph which describes the conditional dependence (or independence) between \( Y_i \) and \( Y_j \) given the remaining variables. We are typically interested in the situation where \( p \) is relatively large and the graph is sparse, that is, most pairs \( Y_i \) and \( Y_j, i \neq j, i, j = 1, \ldots, p \), are conditionally independent given all other variables. When \( Y_i \) and \( Y_j \) are conditionally independent given other variables, there will be no edge connecting \( i \) and \( j \) in the underlying graph; otherwise there will be an edge. Under the assumed multivariate normality of \( Y \), it follows that there is an edge between \( i \) and \( j \) if and only if \( \Omega_{ij} \) is a non-zero matrix. Therefore the problem of identifying the underlying graphical structure reduces to estimating the matrix \( \Omega \) under the sparsity constraint that most off-diagonal blocks \( \Omega_{ij} \) in the grand precision matrix \( \Omega \) are zero.

Suppose that we observe \( n \) independent and identically distributed (i.i.d.) samples from the graphical model, which are collectively denoted by \( Y \), while \( Y_i \) stands for the sample of \( n \) many \( K_i \)-variate observations at node \( i \) and \( Y_{ik} \) stands for the vector of observations of the \( k \)th component at node \( i, k = 1, \ldots, K_i, i = 1, \ldots, p \). Following the estimation strategies used in univariate Gaussian graphical models, we may propose a sparse estimator for \( \Omega \) by minimizing a loss function obtained from the conditional densities of \( Y_i \) given \( Y_j, j \neq i \), for each \( i \) and a penalty term. However, since sparsity refers to off-diagonal blocks rather than individual elements, the lasso-type penalty used in univariate methods like space or concord should be replaced by a group-lasso type penalty, involving the sum of the Frobenius-norms of each off-diagonal block \( \Omega_{ij} \). A multivariate analog of the loss used in a weighted version of space is given by

\[
L_n(\omega, \sigma, Y) = \frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{K_i} \left( -\log \sigma_{ik} + \frac{w_{ik}}{n} ||Y_{ik} + \sum_{j \neq i}^{K_j} \sqrt{\omega_{ijkl}/\sigma_{ik}} Y_{jl} ||^2_{2} \right),
\]

(1)

where \( \sigma_{ik} = \omega_{ikkk}, w = (w_{11}, \ldots, w_{pK_p}) \) are nonnegative weights and \( \omega_{ijkl} = \omega_{jilk} \) due to the symmetry of precision matrix. Writing the quadratic term in the above expression as

\[
w_{ik} ||Y_{ik} + \sum_{j \neq i}^{K_j} \sqrt{\omega_{ijkl}/\sigma_{ik}} Y_{jl} ||^2_{2} = \frac{w_{ik}}{(\sigma_{ik})^2} \sigma_{ik} ||Y_{ik} + \sum_{j \neq i}^{K_j} \omega_{ijkl} Y_{jl} ||^2_{2},
\]

and, as in concord choosing \( w_{ik} = (\sigma_{ik})^2 \) to make the optimization problem convex in the arguments, we can write the quadratic term in the loss function as \( ||\sigma_{ik}Y_{ik} + \sum_{j \neq i}^{K_j} \omega_{ijkl} Y_{jl} ||^2_{2} \). Applying the group penalty we finally arrive at the objective function

\[
\frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{K_i} \left( -\log \sigma_{ii} + \frac{1}{n} ||\sigma_{ik} Y_{ik} + \sum_{j \neq i}^{K_j} \omega_{ijkl} Y_{jl} ||^2_{2} \right) + \lambda \sum_{i<j} \left( \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \omega_{ijkl}^2 \right)^{1/2}.
\]

(2)
Multivariate Gaussian Network

2.2. Algorithm

To obtain a minimizer of (2), we periodically minimize it with respect to the arguments of $\Omega_{ij}$, $i \neq j$, $i, j = 1, \ldots, p$. For each fixed $(i, j)$, $i \neq j$, suppressing the terms not involving any element of $\Omega_{ij}$, we may write the objective function as

$$\frac{1}{2n} \left( \sum_{k=1}^{K_i} \left\| \sigma^{ik} Y_{ik} \right\|_2^2 + \sum_{j' \neq i} \sum_{l=1}^{K_j} \omega_{ij'k} Y_{ij'l} \right) + \frac{1}{2n} \left( \sum_{l=1}^{K_i} \left\| \sigma^{jl} Y_{jl} \right\|_2^2 + \sum_{i' \neq j} \sum_{k=1}^{K_i} \omega_{i'jik} Y_{i'jlk} \right) + \lambda \left\| \omega_{ij} \right\|_2,$$

where $\omega_{ij} = \text{vec}(\Omega_{ij})$. Without loss of generality, we assume $i < j$ and rewrite the expression as

$$\frac{1}{2n} \left( \sum_{k=1}^{K_i} \left\| \sigma^{ik} Y_{ik} + B_{1jk} \omega_{ij} \right\|_2^2 + \sum_{j' > i, j' \neq i} \sum_{l=1}^{K_j} B_{1j'k} \omega_{i'j} \right) + \frac{1}{2n} \left( \sum_{l=1}^{K_i} \left\| \sigma^{jl} Y_{jl} + B_{2il} \omega_{ij} \right\|_2^2 + \sum_{i' > j, i' \neq i} \sum_{l=1}^{K_j} B_{2l'i} \omega_{i'j} \right) + \lambda \left\| \omega_{ij} \right\|_2,$$

where $B_{1jk}$ and $B_{2il}$ are $n \times K_i K_j$ matrices specified as follows: ($(k - 1) K_j + 1, \ldots, k K_j$)th columns of $B_{1jk}$ are $Y_{ij}$, the $(k, K_j + k, \ldots, (K_i - 1) K_j + k)$th columns of $B_{2il}$ are $Y_{il}$, and other columns are zero. This leads to the following algorithm.

Algorithm:

Initialization: For $k = 1, \ldots, K_i$, and $i = 1, \ldots, p$, set the initial values $\hat{\sigma}^{ik} = 1/\text{var}(Y_{ik})$ and $\hat{\omega}_{ij} = 0$.

Iteration: For all $1 \leq i \leq p$ and $1 \leq k \leq K_i$, repeat the following steps until certain convergence criterion is satisfied:

Step 1: Calculate the vectors of errors for $\omega_{ij}$:

$$r_{ijk} = \hat{\sigma}^{ik} Y_{ik} + \sum_{j' > i, j' \neq i} B_{2j'k} \hat{\omega}_{ij} + \sum_{j' > i, j' \neq i} B_{1j'k} \hat{\omega}_{i'j},$$

$$r_{jil} = \hat{\sigma}^{jl} Y_{jl} + \sum_{i' > j} B_{1l'i} \hat{\omega}_{j'i} + \sum_{i' < j, i' \neq i} B_{2l'i} \hat{\omega}_{i'j}.$$

Step 2: Regress the errors on the specified variables to obtain

$$\hat{\omega}_{ij} = \text{arg min} \left\{ \frac{1}{2n} \sum_{k=1}^{K_i} B_{1kj}^T B_{1jk} \omega_{ij} + \sum_{l=1}^{K_j} B_{2il}^T B_{2li} \omega_{ij} \right\},$$

by the proximal gradient algorithm described as follows:
Given $\omega_{ij}^{(t)}$, $r_{ij}^{(t+1)}$ and $r_{ijkl}^{(t+1)}$, compute

$$
\begin{align*}
f(\omega_{ij}^{(t)}) &= \frac{1}{2n} \left[ \omega_{ij}^{(t)T} \left( \sum_{k=1}^{K_i} B_{1jk}^T B_{1jk} + \sum_{l=1}^{K_j} B_{2il}^T B_{2il} \right) \omega_{ij}^{(t)} \\
&\quad + 2 \left( \sum_{k=1}^{K_i} r_{ij}^{(t+1)T} B_{1jk} + \sum_{l=1}^{K_j} r_{ijkl}^{(t+1)T} B_{2il} \right) \omega_{ij}^{(t)} \right] \\
g &= \frac{1}{n} \left( \sum_{k=1}^{K_i} \left( B_{jk}^T B_{jk} \omega_{ij}^{(t)} + r_{ij}^{(t+1)T} B_{jk} \right) + \sum_{l=1}^{K_j} \left( B_{il}^T B_{ij} \omega_{ij}^{(t)} + r_{ijkl}^{(t+1)T} B_{il} \right) \right)
\end{align*}
$$

Set $s \leftarrow 1$ and repeat

- $z_{ij} \leftarrow \omega_{ij}^{(t)} - sg$,
- if $\|z_{ij}\|_2 \geq \lambda s \|z_{ij}\|_2$, set $\omega_{ij}^{(t+1)} \leftarrow \left(1 - \frac{\lambda s}{\|z_{ij}\|_2}\right) z_{ij}$; else set $\omega_{ij}^{(t+1)} \leftarrow 0$,
- replace $s$ by $s/2$,

until $f(\omega_{ij}^{(t)}) \leq f(\omega_{ij}^{(t+1)}) + g^T(\omega_{ij}^{(t+1)} - \omega_{ij}^{(t)}) + \frac{1}{2s} \|\omega_{ij}^{(t+1)} - \omega_{ij}^{(t)}\|_2^2$.

**Step 3**: For $1 \leq i \leq p$ and $1 \leq k \leq K_i$, update $\hat{\sigma}_{ik}$ to

$$
\begin{align*}
- \frac{Y_{ik}^T \left( \sum_{j<i} B_{2jk} \hat{\omega}_{ij} + \sum_{j>i} B_{1jk} \hat{\omega}_{ij} \right) + \sqrt{\left(Y_{ik}^T \left( \sum_{j<i} B_{2jk} \hat{\omega}_{ij} + \sum_{j>i} B_{1jk} \hat{\omega}_{ij} \right) \right)^2 + 2n Y_{ik}^T Y_{ik}}}{2Y_{ik}^T Y_{ik}}.
\end{align*}
$$

If the total number of variables at all nodes $\sum_{i=1}^{p} K_i$ is less than or equal to the available sample size $n$, then the objective function is strictly convex, there is a unique solution to the minimization problem (2) and the iterative scheme converges to the global minimum (Tseng, 2001). However, if $\sum_{i=1}^{p} K_i > n$, the objective function need not be strictly convex, and hence a unique minimum is not guaranteed. However, as in univariate concord, the algorithm converges to a global minimum. This follows by arguing as in the proof of Theorem 1 of Khare et al. (2015) after observing that the objective function of mconcord differs from that of concord only in two aspects — the loss function does not involve off-diagonal entries of diagonal blocks, and the penalty function has grouping, neither of which affect the structure of the concord described by Equation (33) of Khare et al. (2015).

3. LARGE SAMPLE PROPERTIES

In this section, we study large sample properties of the proposed mconcord method. As in the univariate concord method, we consider the estimator obtained from the minimization
problem
\[
\frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{K_i} \left( -\log \hat{\sigma}^{ik} + \frac{\omega_{ik}}{n} ||Y_{ik} + \lambda \sum_{j \neq i}^{K_j} \frac{\omega_{ijkl}}{\hat{\sigma}^{ik}} Y_{jl} ||^2 \right) + \lambda_n \sum_{i < j}^{K_i} \sum_{k=1}^{K_j} \omega_{ijkl}^2 \right)^{1/2}
\]
with a general weight \(w_{ik}\) and a suitably consistent estimator \(\hat{\sigma}^{ik}\) of \(\sigma^{ik}\) plugged in for all \(k = 1, \ldots, K_i, i = 1, \ldots, p, \) and for some suitable sequence \(\lambda_n\). Existence of such an estimator is also shown.

Introduce the notation
\[
L(\omega, \sigma, Y) = \frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{K_i} w_{ik} \left( Y_{ik} + \sum_{j \neq i}^{K_j} \frac{\omega_{ijkl}}{\sigma^{ik}} Y_{jl} \right)^2,
\]
where \(\sigma = (\sigma^{ik} : k = 1, \ldots, K_i, i = 1, \ldots, p)\) and \(\omega = (\omega_{ijkl} : k = 1, \ldots, K_i, l = 1, \ldots, K_j, i,j = 1, \ldots, p, i \neq j)\). Let \(\bar{\omega}\) and \(\bar{\sigma}\) respectively stand for true values of \(\Omega\) and \(\sigma\) respectively. All probability and expectation statements made below are understood under the distributions obtained from the true parameter values. Let \(\bar{L}'_{ijkl}(\omega, \sigma, Y) = E(\frac{\partial}{\partial \omega_{ijkl}} L(\omega, \sigma, Y) | \omega = \bar{\omega}, \sigma = \bar{\sigma})\) and \(\bar{L}''_{ijkl,i'j'k'l'}(\bar{\omega}, \bar{\sigma}) = E(\frac{\partial^2}{\partial \omega_{ijkl} \partial \omega_{i'j'k'l'}} L(\omega, \sigma, Y) | \omega = \bar{\omega}, \sigma = \bar{\sigma})\) be the expected first and second order partial derivatives of \(L\) at the true parameter respectively. Also let \(\bar{L}'_{ijkl,S}\) and \(\bar{L}''_{ijkl,S}\) stand for the row vector \((\bar{L}'_{ijkl,i'j'k'l'} : (i,j,k,l) \in S)\) and \(\bar{L}''_{ijkl,S}\) for the matrix \((\bar{L}''_{ijkl,i'j'k'l'} : i,j,k,l \in S)\), where \(S \subset T := \{(i,j,k,l) : 1 \leq i \neq j \leq p, 1 \leq k \leq K_i, 1 \leq l \leq K_j\}\). Note that \(\bar{L}''_{ijkl,i'j'k'l'}(\bar{\omega}, \bar{\sigma}) = E[Y_{ji}Y_{jl} + Y_{ik}Y_{l'k}] = \sigma_{ji,j'l'} + \sigma_{ik,l'k}\).

Let \(A_0 = \{ (i,j) : \exists k \in \{1, \ldots, K_i\}, \exists l \in \{1, \ldots, K_j\}, \omega_{ijkl} \neq 0 \}\), and \(q_n = |A_0|\). We further define that \(A = \{ (i,j,k,l) : (i,j) \in A_0, 1 \leq k \leq K_i, 1 \leq l \leq K_j \}\), and thus there are \(\sum_{(i,j) \in A_0} K_i K_j\) elements in \(A\). Let \(K_{\max} = \max\{K_i : i = 1, \ldots, p\}\). The following assumptions will be made throughout.

(C0) The weights satisfy \(0 < w_0 \leq \min(w_{ik}) \leq \max(w_{ik}) \leq w_\infty < \infty\) and \(K_{\max}\) and \(p\) grow at most like a power of \(n\).

(C1) There exist constants \(0 < \lambda_{\min} \leq \lambda_{\max}\) depending on the true parameter value such that the minimum and maximum eigenvalues of the true covariance \(\Sigma\) satisfies \(0 < \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \Lambda_{\max} < \infty\).

(C2) There exists a constant \(\delta < 1\) such that for all \((i,j,k,l) \notin A\),
\[
|\bar{L}''_{ijkl,A}(\bar{\omega}, \bar{\sigma})| \frac{1}{|A_A(\bar{\omega}, \bar{\sigma})|^{-1}} |M| \leq \delta,
\]
where \(M\) is a column-vector with elements \(\bar{\omega}_{ijkl} / \sqrt{\sum_{k',l'} \bar{\omega}_{ik'j'l'}}, (i,j,k,l) \in A\).
(C3) There is an estimator $\hat{\sigma}_{ik}$ of $\sigma_{ik}$, $k = 1, \ldots, K_i$, satisfying $\max\{|\hat{\sigma}_{ik} - \hat{\sigma}_{ik}| : 1 \leq i \leq p, 1 \leq k \leq K_i\} \leq C_n \sqrt{(\log n)/n}$ for every $C_n \to \infty$ with probability tending to 1.

The following result concludes that Condition C3 holds if the total dimension is less than a fraction of the sample size.

**Proposition 1.** Suppose that $\sum_{i=1}^p K_i \leq \beta n$ for some $0 < \beta < 1$. Let $e_{ik}$ stand for the vector of regression residuals of $Y_{ik}$ on $\{Y_{il} : l \neq k\}$. Then the estimator $\hat{\sigma}_{ik} = 1/\hat{\sigma}_{ik,-ik}$, where $\hat{\sigma}_{ik,-ik} = (n - \sum_{j \neq i} K_j)^{-1} e_{ik}^T e_{ik}$, satisfies Condition C3.

We adapt the approach in Peng et al. (2009) to the multivariate Gaussian setting. The approach consists of first showing that if the estimator is restricted to the correct model, then it converges to the true parameter at a certain rate as the sample size increases to infinity. The next step consists of showing that with high probability no edge is falsely selected. These two conclusions combined yield the result.

**Theorem 1.** Let $K_{\max} q_n = o(\sqrt{n/\log n})$, $\lambda_n \sqrt{n/\log n} \to \infty$ and $K_{\max} \sqrt{q_n} \lambda_n = o(1)$ as $n \to \infty$. Then the following events hold with probability tending to 1:

(i) there exists a solution $\hat{\omega}_{A_n}^\lambda = \hat{\omega}_{A_n}^\lambda(\hat{\sigma})$ of the restricted problem

$$\arg\min_{\omega, \omega_{A_n}^C = 0} L_n(\omega, \hat{\sigma}, Y) + \lambda_n \sum_{i < j} \|\omega_{ij}\|_2.$$  \hspace{1cm} (4)

(ii) (estimation consistency) for any sequence $C_n \to \infty$, any solution $\hat{\omega}_{A_n}^\lambda$ of the restricted problem (4) satisfies $\|\hat{\omega}_{A_n}^\lambda - \hat{\omega}_{A}^\lambda\|_2 \leq C_n K_{\max} \sqrt{q_n} \lambda_n$.

**Theorem 2.** Suppose that $K_{\max} \leq O(n^\kappa)$ for some $\kappa \geq 0$, $K_{\max} q_n = o(\sqrt{n/\log n})$, $K_{\max} \sqrt{q_n} \log n/n = o(\lambda_n)$, $\lambda_n \sqrt{n/\log n} \to \infty$ and $K_{\max} \sqrt{q_n} \lambda_n = o(1)$ as $n \to \infty$. Then with probability tending to 1, the solution of (4) satisfies $\max\{|L_{n,i,j,k,l}'(\hat{\Omega}_{A,\lambda_n}^\lambda, \hat{\sigma}, Y)| : (i,j,k,l) \in A_n\} < \lambda_n$, where $L_{n,i,j,k,l}' = \partial L_n/\partial \omega_{i,j,k,l}$.

**Theorem 3.** Assume that the sequences $K_{\max}, p, q_n$ and $\lambda_n$ satisfy the conditions in Theorem 2. Then with probability tending to 1, a minimizer $\hat{\omega}_{A_n}^\lambda$ of $L_n(\omega, \hat{\sigma}, Y) + \lambda_n \sum_{i < j} \|\omega_{ij}\|_2$ satisfies

(i) (estimation consistency) for any sequence $C_n \to \infty$, $\|\hat{\omega}_{A_n}^\lambda - \hat{\omega}_{A}^\lambda\|_2 \leq C_n K_{\max} \sqrt{q_n} \lambda_n$.

(ii) (selection consistency) $\hat{A} = A$, where $\hat{A} = \{(i,j) : \hat{\omega}_{ij}^\lambda \neq 0\}$. 
In this section, two simulation studies are conducted to examine the performance of mconcord and compare with space, concord, glasso and multi, the method of Kolar et al. (2014) in regards of estimation accuracy and model selection. For space, concord and glasso, all components of each node are treated as separate univariate nodes, and we put an edge between two nodes as long as there is at least one non-zero entry in the corresponding submatrix.

4.1. Estimation Accuracy Comparison

In the first study, we evaluate the performance of each method at a series of different values of the tuning parameter $\lambda$. Four random networks with $p = 30$ (44% density), $p = 50$ (21% density), $p = 100$ (6% density) and $p = 200$ (2% density) nodes are generated, and each node has a $K$-dimensional Gaussian variable associate with it, $K = 3, 5, 8$. Based on each network, we construct a $pK \times pK$ precision matrix, with non-zero blocks corresponding to edges in the network. Elements of diagonal blocks are set as random numbers from $[0.5, 1]$. If node $i$ and node $j$ ($i < j$) are not connected, then the entire $(i, j)$th and $(j, i)$th blocks would take values zero. If node $i$ and node $j$ ($i < j$) are connected, the $(i, j)$th block would have elements taking values in $(0, 0.05, -0.05, -0.2, 0.2)$ with equal probabilities so that both strong and weak signals are included. The $(j, i)$th block can be obtained by symmetry. Finally, we add $\rho I$ to the precision matrix to make it positive-definite, where $\rho$ is the absolute value of the smallest eigenvalue plus 0.5 and $I$ is the identity matrix. Using each precision matrix, we generate 50 independent datasets consisting of $n = 50$ (for the $p = 30$ and $p = 50$ networks) and $n = 100$ (for the $p = 100$ and $p = 200$ networks) i.i.d. samples. Results are given in Figure 1 to Figure 4. All figures show the number of correctly detected edges ($N_c$) versus the number of total detected edges ($N_t$), averaged across the 30 independent datasets.

We can observe that for all methods, $N_t$ decreases when we increase $\lambda$. It can be seen that mconcord consistently outperforms its counterparts, as it detects more correct edges than the other methods for the same number of total edges detected, especially when we have large $K$ or large $p$. In all scenarios, space, concord and glasso give very similar results. With large $K$ and $p$, multi performs better than univariate methods.

The better performance of mconcord over space, concord and glasso is largely due to the fact that mconcord is designed for multivariate network, and treating the precision matrix
by different blocks is more likely to catch an edge even when the signal is comparably weak. On the contrary, the univariate approaches tend to select more unwanted edges since there is high probability that there is at least one non-zero element in the block due to randomness.

In high dimensional settings, regression based methods have simpler quadratic loss function and are computationally faster and more efficient than that of penalized likelihood methods, which optimize with respect to the entire precision matrix at once. The running time for mconcord is about one-third of that for multi. We also believe that this helps explain why mconcord outperforms multi, especially when we have large $p$, which is reminiscent of the same phenomenon in the univariate case.

### 4.2. Model Selection Comparison

Next in the second study, we compare the model selection performance of the above approaches. We fix $K = 4$, and conduct simulation studies for several combinations of $n$ and $p$ with different densities which vary from 41% to 1%. The precision matrices are generated using the same technique as in the first study. The tuning parameter $\lambda$ is selected using a 5-fold cross-validation for all methods. We also studied the performance of the Bayesian Information Criterion (BIC) for model selection, but it seems that BIC does not work in the multi dimensional settings. In fact, BIC in most cases tends to choose the smallest model where no edge can be detected. Here we compare sensitivity (TPR), precision (PPV) and Matthew’s Correlation Coefficient (MCC) defined by

$$TPR = \frac{TP}{TP + FN}, \quad PPV = \frac{TP}{TP + FP}, \quad MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

where $TP$, $TN$, $FP$ and $FN$ denote true positives (number of edges correctly detected), true negatives (number of edges correctly excluded), false positives (number of edges detected but absent in the true model) and false negatives (number of edges falsely excluded). For each network, all final numbers are averaged across 30 independent datasets.

Table 1 shows that substantial gain is achieved by considering the multivariate aspect in mconcord compared with the univariate methods space and concord in regards of both sensitivity and precision, except for the case $p = 30$ and $n = 50$ where these two methods score slightly better TPR due to more selection of edges. Both glasso and multi select very dense models in nearly all cases, and as a consequence their TPR are higher. However, in terms of MCC which accounts for both correct and incorrect selections, mconcord performs consistently better than all the other methods.
5. **APPLICATION**

5.1. **Gene/Protein Network Analysis**

According to the NCI website [https://dtp.cancer.gov/discovery_development/nci-60](https://dtp.cancer.gov/discovery_development/nci-60), “the US National Cancer Institute (NCI) 60 human tumor cell lines screening has greatly served the global cancer research community for more than 20 years. The screening method was developed in the late 1980s as an in vitro drug-discovery tool intended to supplant the use of transplantable animal tumors in anticancer drug screening. It utilizes 60 different human tumor cell lines to identify and characterize novel compounds with growth inhibition or killing of tumor cell lines, representing leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney cancers”.

We apply our method to a dataset from the well-known NCI-60 database, which consists of protein profiles (normalized reverse-phase lysate arrays for 94 antibodies) and gene profiles (normalized RNA microarray intensities from Human Genome U95 Affymetrix chip-set for more than 17000 genes). Our analysis will be restricted to a subset of 94 genes/proteins for which both types of profiles are available. These profiles are available across the same set of 60 cancer cell lines. Each gene-protein combination is represented by its Entrez ID, which is a unique identifier common for a protein and a corresponding gene that encodes this protein.

Three networks are studied: a network based on protein measurements alone, a network based on gene measurements alone, and a gene-protein multivariate network. For protein alone and gene alone networks, we use concord, and for gene-protein network, we use mconcord. The tuning parameter \( \lambda \) is selected using 5-fold cross-validation for all three networks.

From the gene-protein network 531 edges are selected. For the protein network, 798 edges are selected and for the gene network, 784 edges are selected. Protein and gene-protein networks share 313 edges, while gene and gene-protein networks share 287 edges. However, protein and gene networks only share 167 edges. Table 2 provides summary statistics for these networks.

In Table 3, we also list the top 20 most connected components for all three networks. Among them, the gene-protein network and the protein network share 11, the gene-protein network and the gene network share 10, while the protein network and the gene network share only 6.

5.2. **GDP Network Analysis**

In this analysis, we apply our method to the regional GDP data obtained from U.S. Department of Commerce website [https://www.bea.gov/index.htm](https://www.bea.gov/index.htm), which contains GDP
data including the following 20 different industries with labels: (1) utilities (uti), (2) construction (cons), (3) Manufacturing (manu), (4) Durable goods manufacturing (durable), (5) nondurable goods manufacturing (nondu), (6) wholesale trade (wholesale), (7) retail trade (retail), (8) transportation and warehousing (trans), (9) information (info), (10) finance and insurance (finance), (11) real estate and rental and leasing (real), (12) professional, scientific and technical services (prof), (13) management of companies and enterprises (manage), (14) administrative and waste management services (admin), (15) educational services (edu), (16) health care and social assistance (health), (17) arts, entertainment and recreation (arts), (18) accommodation and food services (food), (19) other services except government (other) and (20) government (gov).

The data is available from the first quarter of 2005 to the second quarter of 2016, and in 8 regions in the US, including New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island and Vermont), Mideast (Delaware, D.C., Maryland, New Jersey, New York and Pennsylvania), Great Lakes (Illinois, Indiana, Michigan, Ohio and Wisconsin), Plains (Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota), Souteast (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia and West Virginia), Southwest (Arizona, New Mexico, Oklahoma and Texas), Rocky Mountain (Colorado, Idaho, Montana, Utah and Wyoming) and Far West (Alaska, California, Hawaii, Nevada, Oregon and Washington).

We reduce correlation in the time series data by taking differences of the consecutive observations. A multivariate network consisting of 20 nodes and 8 attributes for each node is studied. After using 5-fold cross-validation to select the tuning parameter $\lambda$, 47 edges are detected, with density of 24.7% and average node degree of 4.7. The 5 most connected industries are retail trade, transportation, wholesale trade, accommodation and food services, and professional and technical services. The network is shown in Figure 4(a). It is obvious to see hubs comprising of wholesale trade and retail trade. This is very natural for the consumer-driven economy of the US. Both of these two modes are connected to transportation, as both of these industries heavily rely on transporting goods. Another noticeable fact is that education is connected with government. As part of the services provided by government, it is natural that the quality as well as GDP of educational services can both be influenced by government.

The univariate network using the nationwide GDP data only is also studied for comparison using concord. For the tuning parameter $\lambda$, 5-fold cross-validation is applied, and 95 networks are selected, with density of 50% and average node degree of 9.5. The 5 most connected in-
dustries are administrative and waste management services, accommodation and food services, wholesale trade, professional and technical services and health care and social assistance. The network is shown in Figure 4(b). The more modest degree of connections in the multivariate network seems to be more interpretable.

6. PROOF OF THE THEOREMS

We rewrite (3) as
\[ L(\omega, \sigma, Y) = \frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{K_i} w_{ik} \left( Y_{ik} + \sum_{j \neq i} \sum_{l=1}^{K_j} \omega_{ijkl} \hat{Y}_{jl} \right)^2, \]
where \( \hat{Y}_{ik} = Y_{ik} / \sigma_{ik} \).

For any subset \( S \subset T \), the Karush-Kuhn-Tucker (KKT) condition characterizes a solution of the optimization problem
\[
\begin{align*}
\arg \min_{\omega} & \quad \omega_{Sc} = 0 \\
& \quad L_n(\omega, \hat{\sigma}, Y) + \lambda_n \sum_{1 \leq i < j \leq p} \sqrt{\sum_{k=1}^{K_i} \sum_{l=1}^{K_l} \omega_{ijkl}^2}.
\end{align*}
\]

A vector \( \hat{\omega} \) is a solution if and only if for any \( (i, j, k, l) \in S \)
\[
L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y) = -\lambda_n \frac{\hat{\omega}_{ijkl}}{\sqrt{\sum_{k',l'} \hat{\omega}_{ijkl'}^2}} \quad \text{if } \exists 1 \leq k \leq K_i, 1 \leq l \leq K_j, \hat{\omega}_{ijkl} \neq 0
\]
\[
|L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y)| \leq \lambda_n, \quad \text{if } \hat{\omega}_{ijkl} = 0, k = 1, \ldots, K_i, l = 1, \ldots, K_j.
\]

The following lemmas will be needed in the proof of Theorems 1–3. Their proofs are deferred to the Appendix.

**Lemma 1.** The following properties hold.

(i) For all \( \omega \) and \( \sigma \), \( L(\omega, \sigma, Y) \geq 0 \).

(ii) If \( \sigma^{ik} > 0 \) for all \( 1 \leq k \leq K_i \) and \( i = 1, \ldots, p \), then \( L(\cdot, \sigma, Y) \) is convex in \( \omega \) and is strictly convex with probability one.

(iii) For every index \( (i, j, k, l) \) with \( i \neq j \), \( L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}) = 0 \).

(iv) All entries of \( \Sigma \) are bounded and bounded below. Also, there exist constants \( 0 < \hat{\sigma}_0 \leq \hat{\sigma}_\infty < \infty \) such that
\[
\hat{\sigma}_0 \leq \min\{\hat{\sigma}^{ik} : 1 \leq i \leq p, 1 \leq k \leq K_i\} \leq \max\{\hat{\sigma}^{ik} : 1 \leq i \leq p, 1 \leq k \leq K_i\} \leq \hat{\sigma}_\infty.
\]

(v) There exists constants \( 0 < \Lambda_{\min}^{L}(\hat{\omega}, \hat{\sigma}) \leq \Lambda_{\max}^{L}(\hat{\omega}, \hat{\sigma}) < \infty \), such that
\[
0 < \Lambda_{\min}^{L}(\hat{\omega}, \hat{\sigma}) \leq \lambda_{\min}(\tilde{L}'(\hat{\omega}, \hat{\sigma})) \leq \lambda_{\max}(\tilde{L}'(\hat{\omega}, \hat{\sigma})) \leq \Lambda_{\max}^{L}(\hat{\omega}, \hat{\sigma}) < \infty.
\]
Lemma 2. (i) There exists a constant $N < \infty$, such that for all $1 \leq i \neq j \leq p$ and $1 \leq k \leq K_i$, $1 \leq l \leq K_j$, $\bar{L}_{ijkl,ijkl}'(\bar{\omega}, \bar{\sigma}) \leq N$.  

(ii) There exists constants $M_1, M_2 < \infty$, such that for any $1 \leq i < j \leq p$, 

$$\text{Var}(L_{ijkl}'(\bar{\omega}, \bar{\sigma}, Y)) \leq M_1, \quad \text{Var}(L_{ijkl}'(\bar{\omega}, \bar{\sigma}, Y)) \leq M_2.$$  

(iii) There exists a positive constant $g$, such that for all $(i, j, k, l) \in A$, 

$$L_{ijkl,ijkl}'(\bar{\omega}, \bar{\sigma}) - L_{ijkl,Aijkl}'(\bar{\omega}, \bar{\sigma})[L_{Aijkl,ijkl}'(\bar{\omega}, \bar{\sigma})]^{-1}L_{Aijkl,ijkl}'(\bar{\omega}, \bar{\sigma}) \geq g,$$

where $A_{ijkl} = A \setminus \{(i, j, k, l)\}$. 

(iv) For any $(i, j, k, l) \in A^c$, $\|L_{ijkl,A}'(\bar{\omega}, \bar{\sigma})[L_{Aijkl}'(\bar{\omega}, \bar{\sigma})]^{-1}\|_2 \leq M_3$, for some constant $M_3$.  

Lemma 3. There exists a constant $M_4 < \infty$, such that for any $1 \leq i \leq j \leq p$ and $1 \leq k \leq K_i$, $1 \leq l \leq K_j$, $|E[Y_{ik}Y_{jl}\hat{Y}^T]| \leq M_4$.  

Lemma 4. Let the conditions of Theorem 2 hold. Then for any sequence $C_n \to \infty$, 

$$\max_{1 \leq i < j \leq p, 1 \leq k, l \leq K} \left| L_{ijkl}'(\bar{\omega}, \bar{\sigma}, Y) - L_{ijkl}'(\bar{\omega}, \bar{\sigma}, Y) \right| \leq C_n \sqrt{\frac{\log n}{n}},$$

$$\max_{i < j, l < h} \left| L_{ijkl,ijkl}'(\bar{\omega}, \bar{\sigma}, Y) - L_{ijkl,ijkl}'(\bar{\omega}, \bar{\sigma}, Y) \right| \leq C_n \sqrt{\frac{\log n}{n}},$$

hold with probability tending to 1.  

Lemma 5. If $K_{\max}^2 q_n = o(\sqrt{n/\log n})$, then for any sequence $C_n \to \infty$ and any $u$, the following hold with probability tending to 1: 

$$\|L_{n,A}'(\bar{\omega}, \bar{\sigma}, Y)\|_2 \leq C_n K_{\max} \sqrt{\frac{q_n \log n}{n}},$$

$$|u^T L_{n,A}'(\bar{\omega}, \bar{\sigma}, Y)| \leq C_n \|u\|_2 K_{\max} \sqrt{\frac{q_n \log n}{n}},$$

$$|u^T L_{n,A,A}'(\bar{\omega}, \bar{\sigma}, Y) u - u^T L_{A,A}'(\bar{\omega}, \bar{\sigma}) u| \leq C_n \|u\|_2^2 K_{\max}^2 q_n \sqrt{\frac{\log n}{n}},$$

$$\|L_{n,A,A}'(\bar{\omega}, \bar{\sigma}, Y) u - L_{A,A}'(\bar{\omega}, \bar{\sigma}) u\|_2 \leq C_n \|u\|_2^2 K_{\max}^2 q_n \sqrt{\frac{\log n}{n}}.$$  

Lemma 6. Assume that the conditions of Theorem 1 hold. Then exists a constant $\bar{C}_1 > 0$, such that with probability tending to 1, there exists a local minimum of the restricted problem (4) within the disc $\{\omega : \|\omega - \bar{\omega}\|_2 \leq \bar{C}_1 K_{\max} \sqrt{q_n} \lambda_n\}$. 


LEMMA 7. Assume the conditions of Theorem 1. Then exists a constant $\bar{C}_2 > 0$ such that for any $\omega$ satisfying $\|\omega - \bar{\omega}\|_2 \geq \bar{C}_2 K_{\max} \sqrt{q_n} \lambda_n$ and $\omega_{A^c} = 0$, we have $\|L'_{n,A}(\omega, \hat{\sigma}, Y)\|_2 > K_{\max} \sqrt{q_n} \lambda_n$ with probability tending to 1.

LEMMA 8. Let $D_{A,A}(\bar{\omega}, \bar{\sigma}, Y) = L''_{A,A}(\bar{\omega}, \bar{\sigma}, Y) - \bar{L}''_{A,A}(\bar{\omega}, \bar{\sigma})$. Then there exists a constant $M_5 < \infty$, such that for any $(i, j, k, l) \in A$, $\lambda_{\max} \{(\nu(D_{A,ijkl}(\bar{\omega}, \bar{\sigma}, Y))) \leq M_5$.

Proof of Theorem 1. The existence of a solution of (4) follows from Lemma 6. By the KKT condition, any solution $\hat{\omega}$ of (4), satisfies $\|L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y)\|_\infty \leq \lambda_n$, implying $\|L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y)\|_2 \leq K_{\max} \sqrt{q_n} \|L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y)\|_\infty \leq K_{\max} \sqrt{q_n} \lambda_n$. Thus by Lemma 7, with probability tending to 1, all solutions of (4) are inside the disc $\{\omega : \|\omega - \bar{\omega}\|_2 \leq \bar{C}_2 K_{\max} \sqrt{q_n} \lambda_n\}$. Hence with probability tending to 1, $\|\hat{\omega}_A - \bar{\omega}\|_2 \leq \bar{C}_2(\hat{\omega}) K_{\max} \sqrt{q_n} \lambda_n$. □

Proof of Theorem 2. By the KKT condition and the expansion of $L'_{n,A}(\hat{\omega}_A^\lambda, \hat{\sigma}, Y)$ at $\bar{\omega}$,

$$-\lambda_n \bar{M}^A = L'_{n,A}(\hat{\omega}_A^\lambda, \hat{\sigma}, Y) = L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y) + L''_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y) \nu_n$$

$$= \bar{L}''_{A,A}(\hat{\omega}, \hat{\sigma}) \nu_n + L''_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y) + \left[ L''_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y) - \bar{L}''_{A,A}(\hat{\omega}, \hat{\sigma}) \right] \nu_n,$$

where $\nu_n := \hat{\omega}_A^\lambda - \bar{\omega}$ and $\bar{M}^A = (\hat{\omega}_{ijkl}/\sqrt{\sum_{k',l'} \hat{\omega}_{ijkl}^2} : (i, j, k, l) \in A)^T$. Therefore

$$\nu_n = -\lambda_n \left[L''_{A,A}(\hat{\omega}, \hat{\sigma})\right]^{-1} \bar{M}^A - \left[L''_{A,A}(\hat{\omega}, \hat{\sigma})\right]^{-1} \left[L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y) + D_{n,A}(\hat{\omega}, \hat{\sigma}, Y) \nu_n \right], \quad (5)$$

where $D_{n,A}(\hat{\omega}, \hat{\sigma}, Y) = L''_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y) - \bar{L}''_{A,A}(\hat{\omega}, \hat{\sigma})$. Next, fix $(i, j, k, l) \in A^c$, and consider the expansion of $L'_{n,ijkl}(\hat{\omega}_A^\lambda, \hat{\sigma}, Y)$ around $\bar{\omega}$ is given by

$$L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y) + L''_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y) \nu_n$$

$$= L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y) + L''_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}) \nu_n + \left[ L''_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y) - \bar{L}''_{ijkl,A}(\hat{\omega}, \hat{\sigma}) \right] \nu_n$$

$$= L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y) + L''_{ijkl,A}(\hat{\omega}, \hat{\sigma}) \nu_n + D_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y) \nu_n. \quad (6)$$

Then plugging (5) into (6) and rearranging, $L'_{n,ijkl}(\hat{\omega}_A^\lambda, \hat{\sigma}, Y)$ is given by

$$L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y) - \lambda_n L''_{ijkl,A}(\hat{\omega}, \hat{\sigma}) \left[L''_{A,A}(\hat{\omega}, \hat{\sigma})\right]^{-1} \bar{M}^A$$

$$- L''_{ijkl,A}(\hat{\omega}, \hat{\sigma}) \left[L''_{A,A}(\hat{\omega}, \hat{\sigma})\right]^{-1} L'_{ijkl}(\hat{\omega}, \hat{\sigma})$$

$$+ \left[ D_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y) - L''_{ijkl,A}(\hat{\omega}, \hat{\sigma}) \left[L''_{A,A}(\hat{\omega}, \hat{\sigma})\right]^{-1} D_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y) \right] \nu_n. \quad (7)$$

By Condition C2, for any $(i, j, k, l) \in A^c : |L''_{ijkl,A}(\hat{\omega}, \hat{\sigma})| |L''_{A,A}(\hat{\omega}, \hat{\sigma})|^{-1} M| \leq \delta < 1$. By Theorem 1, we have $\|\hat{\omega}_A^\lambda - \bar{\omega}\|_2 = O_p(K_{\max} \sqrt{q_n} \lambda_n) = o_p(1)$, then $|\bar{M}^A - M| = o_p(1)$. Hence for any $(i, j, k, l) \in A^c : |L''_{ijkl,A}(\hat{\omega}, \hat{\sigma})| |L''_{A,A}(\hat{\omega}, \hat{\sigma})|^{-1} \bar{M}^A| \leq \delta < 1$. Thus it suf-
fices to prove that the remaining term in (7) are $o(\lambda_n)$ with probability tending to 1 uniformly for all $(i,j,k,l) \in \mathcal{A}^4$. Then since $|A^c| \leq K_{\max}^2 p^2 = O(n^{2\kappa})$, the event

$$\max_{(i,j,k,l)\in A^c} |L'_{n,ijkl}(\hat{\omega}, Y)| < \lambda_n$$

happens with probability tending to 1.

By Lemma 2(iv), for any $(i,j,k,l) \in A^c$, $\|\hat{L}'_{ijkl,A}(\hat{\omega}, \hat{\sigma})|L''_{n,A,A}(\hat{\omega}, \hat{\sigma})|^{-1}\| \leq M_3(\hat{\omega}, \hat{\sigma})$. Therefore by Lemma 5,

$$\max_{(i,j,k,l)\in A^c} |\hat{L}'_{ijkl,A}(\hat{\omega}, \hat{\sigma})|L''_{n,A,A}(\hat{\omega}, \hat{\sigma})|^{-1}L'_{n,A}(\hat{\omega}, \hat{\sigma}, Y)\leq CnK_{\max}\sqrt{\frac{q_n\log n}{n}} = o(\lambda_n)$$

with probability tending to 1, choosing a sufficiently slow $C_n \to \infty$. By Lemma 2(ii), $\text{Var}(L'_{ijkl}(\hat{\omega}, \hat{\sigma}, Y)) \leq M_1(\hat{\omega}, \hat{\sigma})$. Then as in Lemma 5, with probability tending to 1, $\max_{i,j,k,l} |L'_{n,ijkl}(\hat{\omega}, \hat{\sigma}, Y)| \leq Cn\sqrt{(\log n)/n} = o(\lambda_n)$, by virtue of the assumption that $\lambda_n\sqrt{n/\log n} \to \infty$.

Note that by Theorem 1, $\|\nu_n\|_2 \leq C_nK_{\max}\sqrt{q_n}\lambda_n$ with probability tending to 1. Thus as in Lemma 5, for sufficiently slowly growing sequence $C_n \to \infty$, $|D_{n,ijkl,A}(\hat{\omega}, \hat{\sigma}, Y)\nu_n| \leq CnK_{\max}\sqrt{q_n(\log n)/n}K_{\max}\sqrt{q_n}\lambda_n = o(\lambda_n)$ with probability tending to 1. This claim follows from the assumption $K_{\max}^2q_n = o(\sqrt{n/\log n})$.

Finally, let $b^T = \hat{L}'_{ijkl,A}(\hat{\omega}, \hat{\sigma})|L''_{n,A,A}(\hat{\omega}, \hat{\sigma})|^{-1}$. By the Cauchy-Schwartz inequality

$$|b^TD_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y)\nu_n| \leq \|b^TD_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y)\|_2\|\nu_n\|_2$$

$$\leq K_{\max}^2q_n\lambda_n\max_{(i',j',k',l')\in A} |b^TD_{n,A,i'j'k'l'}(\hat{\omega}, \hat{\sigma}, Y)|.$$

In order to show that the right hand side is $o(\lambda_n)$ with probability tending to 1, it suffices to show

$$\max_{(i',j',k',l')\in A} |b^TD_{n,A,i'j'k'l'}(\hat{\omega}, \hat{\sigma}, Y)| = O\left(\frac{\sqrt{\log n}}{n}\right)$$

with probability tending to 1, because of the assumption $K_{\max}^2q_n = o(\sqrt{n/\log n})$. This is implied by $\mathbb{E}(|b^T D_{A,i'j'k'l'}(\hat{\omega}, \hat{\sigma}, Y)|^2) \leq \|b\|_2^2\lambda_{\max} \text{Var}(D_{A,i'j'k'l'}(\hat{\omega}, \hat{\sigma}, Y))$ being bounded, which follows immediately from Lemma 1(iv) and Lemma 8. Finally, as in Lemma 5,

$$\|b^TD_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y)\nu_n| \leq |b^TD_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y)\nu_n|$$

$$+|b^T(D_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y) - D_{n,A,A}(\hat{\omega}, \hat{\sigma}, Y))\nu_n|,$$

where by Lemma 4, the second term on the right hand side is bounded by $O_p(\sqrt{(\log n)/n})\|b\|_2\|\nu_n\|_2$. Note that $\|b\|_2 = O(K_{\max}\sqrt{q_n})$, thus the second term is also of order $o(\lambda_n)$ by the assumption $K_{\max}^2q_n = o(\sqrt{n/\log n})$. \qed
Proof of Theorem 3. By Theorems 1 and 2 and the KKT condition, with probability tending to 1, a solution of the restricted problem is also a solution of the original problem. This shows the existence of the desired solution. If the objective function is strictly convex, the solution is also unique, so this will be the only solution for the original problem.

Finally, convergence properties of the estimator of $\sigma$ claimed in Proposition 1 is shown.

Proof of Proposition 1. Observe that when $\sum_{i=1}^{p} K_i < \beta n$, $e_{ik}$ can be expressed as $e_{ik} = Y_{ik} - Y_{ik}^T (Y_{ik}^T Y_{ik})^{-1} Y_{ik}$. As argued in Peng et al. (2009), $E(e_{ik}^T e_{ik}) = 1/\bar{\sigma}_{ik}$. Therefore, by Lemma 9 of the Appendix and Lemma 1(iv), we have $\max \{ |\bar{\sigma}_{ik} - \bar{\sigma}_{ik}| : 1 \leq k \leq K_i, 1 \leq i \leq p \} = O_p(\sqrt{\log n}/n)$.

Acknowledgement

This research is partially supported by National Science Foundation grant DMS-1510238.

Appendix A. Proof of the Lemmas

Proof of Lemma 1. The assertions (i) and (ii) are self-evident from the definition of $L$. To prove (iii), denote the residual for the $i$th term by $e_{ik}(\omega, \sigma) = Y_{ik} + \sum_{j \neq i} K_j \omega_{ijkl} \bar{Y}_{jl}$. Then evaluated at the true parameter values $(\tilde{\omega}, \tilde{\sigma}),$ we have $e_{ik}(\tilde{\omega}, \tilde{\sigma})$ uncorrelated with $Y_{jl}$ and $E(e_{ik}(\tilde{\omega}, \tilde{\sigma})) = 0$. Since $\partial L(\omega, \sigma, Y)/\partial \omega_{ijkl} = w_{ik} e_{ik}(\omega, \sigma) Y_{jl} + w_{jl} e_{jl}(\omega, \sigma) Y_{ik}$, (iii) follows by taking expectation.

Since all eigenvalues of $\tilde{\Sigma}$ lie between two positive numbers, so do all diagonal entries because these are values of quadratic forms for unit vectors having 1 at one place. All off-diagonals entries lie in $[-\Lambda_{\max}, \Lambda_{\max}]$ because these are values of bilinear forms at such unit vectors. This shows (iv).

To prove (v), let $\tilde{\bar{X}} = (\tilde{\bar{X}}_{(11,21)}, \ldots, \tilde{\bar{X}}_{(11,2K_2)}, \ldots, \tilde{\bar{X}}_{(1K_1,2K_2)}, \ldots, \tilde{\bar{X}}_{((p-1)K_{p-1},pK_{p})})$, with $\tilde{\bar{X}}_{(ik,jl)} = (0, \ldots, 0, \tilde{\bar{Y}}_{jl}, 0, \ldots, 0, \bar{Y}_{ik}, 0, \ldots, 0)^T$, a matrix of order $p \sum_{i=1}^{K_i} K_i \times \sum_{i<j} K_i K_j$, where only the $(i, k)$th and $(j, l)$th elements are non zero. The loss function can be written as $L(\omega, \sigma, Y) = \frac{1}{2} \| w^{1/2} (Y - \tilde{\bar{X}} \omega) \|^2_2$, where $w^{1/2} = \text{diag}(\sqrt{w_{11}}, \ldots, \sqrt{w_{pK_p}})$. Thus $L''(\omega, \sigma) = E[\tilde{\bar{X}}^T (w^{1/2})^2 \tilde{\bar{X}}]$. Let $d = \sum_{i<j} K_i K_j$, the number of columns in $\tilde{\bar{X}}$, and denote its $(i, k)$th row by $X_{ik}^T$, $1 \leq k \leq K_i$, $1 \leq i \leq p$. Then for any unit vector $a \in \mathbb{R}^d$, we have

$$a^T L''(\omega, \sigma) a = E(a^T \tilde{\bar{X}}^T (w^{1/2})^2 \tilde{\bar{X}} a) = E\left( \sum_{i=1}^{p} \sum_{k=1}^{K_i} w_{ik} (X_{ik}^T a)^2 \right).$$
Index the elements of a as \( (a_{1(1,2)}, \ldots, a_{1(2,2)}, \ldots, a_{1K_1,2K_2}, \ldots, a_{((p-1)K_{p-1}, pK_p)})^T \), and for each \( 1 \leq i \leq p \) and \( 1 \leq k \leq K_i \), define \( a_{ik} \in \mathbb{R}^{K_i \times p} \) by
\[
a_{ik} = \begin{cases} 
(0, \ldots, 0, a_{1(1,2)}, \ldots, a_{1(2,2)}, \ldots, a_{1K_1,2K_2}, \ldots, a_{1K_i,pK_p})^T, & i = 1, \\
(a_{ik,1}, \ldots, a_{ik,1K_1}, \ldots, a_{ik,(p-1)K_{p-1}}, 0, \ldots, 0, 0)^T, & i = p, \\
(a_{1(1,ik)}, \ldots, a_{1((i-1)K_{i-1}, ik)}, \ldots, a_{1(ik,ik)}, \ldots, a_{1K_i,ik})^T, & 1 < i < p, 
\end{cases}
\]
with exactly \( K_i \) zeros and \( \sum_{j \neq i} K_j \) non-zeros. Then by definition \( X^T_k a = \hat{Y}^T a_{ik} \). Also note that
\[
\sum_{i=1}^p \sum_{k=1}^{K_i} \|a_{ik}\|_2^2 = 2\|a\|_2^2 = 2. 
\]
This is because, each element of \( a_{ik} \) appears exactly twice in \( a \). Therefore, since \( \hat{L}'(\omega, \sigma) = \hat{E} \hat{Y}^T \), we have
\[
a^T \hat{L}'(\omega, \sigma) a = \sum_{i=1}^p \sum_{k=1}^{K_i} w_{ik} a_{ik}^T \hat{Y}_{ik} \geq \sum_{i=1}^p \sum_{k=1}^{K_i} w_{ik} \lambda_{\min}(\hat{\Sigma}) \|a_{ik}\|_2^2 \geq 2w_0 \lambda_{\min}(\hat{\Sigma}),
\]
where \( \hat{\Sigma} = \var(\hat{Y}) \). Similarly, \( a^T \hat{L}''(\omega) a \leq 2w_\infty \lambda_{\max}(\hat{\Sigma}) \). By Condition C1, \( \hat{\Sigma} \) has bounded eigenvalues, and hence (v) follows. \( \square \)

**Proof of Lemma 2.** The proof of (i) follows because \( \hat{L}''_{ijkl,i'j'k'l'}(\omega, \sigma) = \sigma_{jl,j'l'} + \sigma_{ik,ik'} \), and the entries of \( \hat{\Sigma} \) are bounded by Lemma 1(iv).

For (ii), note that \( \text{Var}(e_{ik}(\omega, \sigma)) = 1/\sigma_{ik} \) and \( \text{Var}(Y_{ik}) = \sigma_{ik,ik} \),
\[
\text{Var}(L_{n,ijkl}(\omega, \sigma, Y)) = \text{Var}(w_{ik} e_{ik}(\omega, \sigma) Y_{jl}) + \text{Var}(w_{jl} e_{jl}(\omega, \sigma) Y_{ik})
\]
\[
\leq \text{E}(w_{ik}^2 e_{ik}^2(\omega, \sigma) Y_{jl}^2) + \text{E}(w_{jl}^2 e_{jl}^2(\omega, \sigma) Y_{ik}^2) = \frac{w_{ik}^2 \sigma_{jl,jl}}{\sigma_{ik}} + \frac{w_{jl}^2 \sigma_{ik,ik}}{\sigma_{jl}}.
\]
The right hand side is bounded because of Condition C0 and Lemma 1(iv), and the fact that \( e_{ik}(\omega, \sigma) \) and \( Y_{jl} \) are independent.

For \( i, j, k, l \in \mathcal{A} \), denote
\[
D := \hat{L}''_{ijkl}(\omega, \sigma) - \hat{L}''_{ijkl,A^{-}\lambda_{ijkl}}(\omega, \sigma) \left[ \hat{L}''_{A^{-}\lambda_{ijkl},A^{-}\lambda_{ijkl}}(\omega, \sigma) \right]^{-1} \hat{L}''_{A^{-}\lambda_{ijkl},ijkl}(\omega, \sigma).
\]
Then \( D^{-1} \) is the \((ijkl, i'j'k'l')\)th entry in \( \left[ \hat{L}''_{A,A}(\omega, \sigma) \right]^{-1} \). Thus by Lemma 1(v), \( D^{-1} \) is positive and bounded from above, so \( D \) is bounded away from zero. This proves (iii).

Note that \( \|\hat{L}''_{ijkl,A}(\omega, \sigma)\|_2 \geq \lambda_{\min}(\hat{\Sigma})^{-1} \) is bounded from above, thus it suffices to show that \( \|\hat{L}''_{ijkl,A}(\omega, \sigma)\|_2 \) is bounded. Define \( \mathcal{A}^+ := \{(i, j, k, l) \cup \mathcal{A} \} \). Then \( \hat{L}''_{ijkl,A}(\omega, \sigma) = \hat{L}''_{ijkl}(\omega, \sigma) \left[ \hat{L}''_{A,A}(\omega, \sigma) \right]^{-1} \hat{L}''_{ijkl}(\omega, \sigma) \) is the inverse of the \((kl, kl)\) entry of \( \hat{L}''_{A,A}(\omega, \sigma) \). Thus by Lemma 1(iv), \( \hat{L}''_{ijkl,A}(\omega, \sigma) \) is bounded away from zero. Therefore by Lemma 2(i),
\[
\hat{L}''_{ijkl,A}(\omega, \sigma) \left[ \hat{L}''_{A,A}(\omega, \sigma) \right]^{-1} \hat{L}''_{ijkl}(\omega, \sigma) \geq \|\hat{L}''_{ijkl,A}(\omega, \sigma)\|_2 \lambda_{\min}(\|\hat{L}''_{ijkl,A}(\omega, \sigma)\|_2^{-1}),
\]
Multivariate Gaussian Network

and by Lemma 1(iv), $\lambda_{\min}(\mathcal{L}_{A,\mathcal{A}}^{\prime}(\tilde{\omega}, \tilde{\sigma}))^{-1}$ is bounded away from zero, we have $\|\mathcal{L}_{ijkl,A}^{\prime}(\tilde{\omega}, \tilde{\sigma})\|_2$ bounded from above. Thus (iv) follows. \hfill \Box

Proof of Lemma 3. The $(i'k',j'l')$th entry of the matrix $Y_{ik}Y_{jl}Y_jY_i^T$ is $Y_{ik}Y_{jl}Y_{jl}Y_{ik}^T$, for $1 \leq i < j \leq p$, $1 \leq k' \leq K_j'$ and $1 \leq l' \leq K_i'$. Hence, the $(i'k',j'l')$th entry of the matrix $E[Y_{ik}Y_{jl}Y_jY_i^T]$ is $E[Y_{ik}Y_{jl}Y_{jl}Y_{ik}^T] = (\tilde{\sigma}_{ik,jl}(\tilde{\sigma}_{ik,jl} + \tilde{\sigma}_{ik,jl}(\tilde{\sigma}_{ik,jl} + \tilde{\sigma}_{ik,jl}/(\tilde{\sigma}_{ik,jl})))$, where $\tilde{\sigma}_{ik,jl}$ denotes the covariance between $Y_{ik}$ and $Y_{jl}$. Thus, we can write

$$E[Y_{ik}Y_{jl}Y_jY_i^T] = \frac{1}{\tilde{\sigma}_{ik,jl}^2}(\tilde{\sigma}_{ik,jl}^2 + \tilde{\sigma}_{ik,jl}^2 + \tilde{\sigma}_{ik,jl}^2), \quad (A1)$$

where $\tilde{\sigma}_{ik,jl}$ is the $\sum_{j=1}^{K_j} K_j$ vector $(\tilde{\sigma}_{ik,jl} : l = 1, \ldots, K_j, j = 1, \ldots, p, j \neq i)$. Then we have

$$\|E[Y_{ik}Y_{jl}Y_jY_i^T]\| \leq \frac{1}{\tilde{\sigma}_{ik,jl}^2}(\|\tilde{\sigma}_{ik,jl}\| + 2\|\tilde{\sigma}_{ik,jl}\|\|\tilde{\sigma}_{jk,l}\|), \quad (A2)$$

where $\|\cdot\|$ is the operator norm. By Condition C1, $|\tilde{\sigma}_{ik,jl}|$ and $\|\tilde{\sigma}_{ik,jl}\|$ are uniformly bounded. Further $\tilde{\sigma}_{ik,jl} - \tilde{\sigma}_{ik,jl}^2 \tilde{\Sigma}_{(-ik)}^{-1} \tilde{\sigma}_{ik,jl}, > 0$, where $\tilde{\Sigma}_{(-ik)}$ is the submatrix of $\tilde{\Sigma}$ removing $ik$th row and column. From this, it follows that

$$\|\tilde{\sigma}_{ik,jl}\| = \|\tilde{\Sigma}_{(-ik)}^{-1/2} \tilde{\Sigma}_{-ik}^{-1/2} \tilde{\sigma}_{ik,jl}\| \leq \|\tilde{\Sigma}_{(-ik)}^{-1/2} \tilde{\Sigma}_{-ik}^{-1/2}\| \|\tilde{\sigma}_{ik,jl}\| \leq \sqrt{\|\tilde{\Sigma}\|}\sqrt{\|\tilde{\sigma}_{ik,jl}\|}, \quad (A3)$$

which follows from the fact that $\tilde{\Sigma}_{-ik}$ is a principal submatrix of $\tilde{\Sigma}$. \hfill \Box

Proof of Lemma 4. Observe that $L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y)$ is given by

$$\frac{1}{n} \sum_{m=1}^{n} w_{ik}(Y_{ik}^m + \sum_{j' \neq i, l'=1}^{K_j} \omega_{ik} \omega_{ik} Y_{ik}^m Y_{jl}^m + w_{jl}(Y_{jl}^m + \sum_{i' \neq j, k'=1}^{K_i} \omega_{ik} \omega_{ik} Y_{ik}^m Y_{ij}^m) Y_{ik}^m Y_{jl}^m, \quad (A4)$$

Thus $L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y) = L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y)$ is given by

$$w_{ik}(Y_{ik}^m Y_{jl}(\frac{1}{\tilde{\sigma}_{ik}} - \frac{1}{\tilde{\sigma}_{ik}}) + \sum_{j' \neq i, l'=1}^{K_j} \omega_{ik} \omega_{ik} Y_{ik}^m Y_{jl} Y_{jl}^m - w_{jl}(Y_{jl}^m Y_{ik}(\frac{1}{\tilde{\sigma}_{jl}} - \frac{1}{\tilde{\sigma}_{jl}}) + \sum_{i' \neq j, k'=1}^{K_i} \omega_{ik} \omega_{ik} Y_{ik}^m Y_{ij}^m Y_{ij}^m Y_{ik}^m), \quad (A5)$$

where $Y_{ik}Y_{jl} = \frac{1}{n} \sum_{m=1}^{n} Y_{ik}^m Y_{jl}^m$. By Lemma 1(iv), $\{\tilde{\sigma}_{ik,jl} : 1 \leq i, j \leq p, 1 \leq k, l \leq K\}$ are bounded from below and above, and hence $\max_{i,j,k,l} \|Y_{ik}Y_{jl} - \tilde{\sigma}_{ik,jl}\| = O_p(\sqrt{\log n}/n)$. This implies that $\max_{i,j,k,l} \|Y_{ik}Y_{jl}\| = O_p(1)$, and hence by Lemma 1(iv) and Condition C3 it follows that

$$\max_{i,j,k,l} \|L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y) - L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y)\| = O_p\left(\frac{\sqrt{\log n}}{n}\right).$$

The bound for $|L''_{n,ijkl,tstkl'}(\tilde{\omega}, \tilde{\sigma}, Y) - L''_{n,ijkl,tstkl'}(\tilde{\omega}, \tilde{\sigma}, Y)|$ follows similarly.

Proof of Lemma 5. If we replace $\hat{\sigma}$ by $\tilde{\sigma}$ on the left hand side and take $(i, j, k, l) \in \mathcal{A}$, then from the definition we have $L'_{n,ijkl}(\tilde{\omega}, \tilde{\sigma}, Y) = e_{ik}(\tilde{\omega}, \tilde{\sigma})^T Y_{jl} + e_{jl}(\tilde{\omega}, \tilde{\sigma})^T Y_{ik}$, and $Y_{ijl}$, where $e_{ik}$ are $n$ replications
of $e_{ik}(\omega, \sigma)$. Thus by Lemma 10 of the Appendix we obtain $\max \{ |L'_{n,ijkl}(\omega, \sigma, Y) : (i, j, k, l) \in A \} \leq C_n \sqrt{\log n}$. and hence by the Cauchy-Schwartz inequality

$$\|L'_{n,A}(\omega, \sigma, Y)\| \leq K_{\max} \sqrt{\log n} \max_{(i,j,k,l) \in A} |L'_{n,ijkl}(\omega, \sigma, Y)| \leq C_n K_{\max} \sqrt{q_n \log n},$$

and $\|L'_{n,A}(\omega, \sigma, Y)\| \leq \|L'_{n,A}(\omega, \sigma, Y)\| + \|L'_{n,A}(\omega, \sigma, Y) - L'_{n,A}(\omega, \sigma, Y)\|$. The second term on the right hand side has order $K_{\max} \sqrt{q_n \log n}$. Since there are $r_{max} q_n$ terms and by Lemma 4, they are uniformly bounded by $\sqrt{\log n}$. The rest of the lemma can be proved by similar arguments. \square

**Proof of Lemma 6.** Let $\alpha_n = K_{\max} \sqrt{q_n} \lambda_n$, and $L_n(\omega, \sigma, Y) = L_n(\omega, \sigma, Y) + \lambda \sum \sum_{i<j} \|\omega_{ij}\|$. Then for any given constant $\bar{C}_1 > 0$ and any vector $u$ such that $u_{A'} = 0$ and $\|u\| = \bar{C}_1$, the triangle inequality and the Cauchy-Schwartz inequality together imply that

$$\sum_{i<j} \|\omega_{ij}\| - \sum_{i<j} \|\omega_{ij} + \alpha_n u_{ij}\| \leq \alpha_n \sqrt{r_{max} q_n} \|u\| = \bar{C}_1 \alpha_n K_{\max} \sqrt{q_n}.$$

Thus $L_n(\omega + \alpha_n u, \sigma, Y, \lambda_n) - L_n(\omega, \sigma, Y, \lambda_n)$ can be written as

$$L_n(\omega + \alpha_n u, \sigma, Y) = L_n(\omega, \sigma, Y) + \lambda \sum \sum_{i<j} \|\omega_{ij}\| - \sum_{i<j} \|\omega_{ij} + \alpha_n u_{ij}\|$$

Thus for any sequence $C_n \to \infty$, with probability tending to 1,

$$L_n(\omega + \alpha_n u, \sigma, Y) - L_n(\omega, \sigma, Y) = \alpha_n u_{A}^{T} L'_{n,A}(\omega, \sigma, Y) u_{A} + \frac{1}{2} \alpha_n^2 u_{A}^{T} L''_{n,A}(\omega, \sigma, Y) u_{A}$$

$$\geq \frac{1}{2} \alpha_n^2 u_{A}^{T} L''_{n,A}(\omega, \sigma, Y) u_{A} - C_n \alpha_n^2 K_{max} q_n n^{-1/2} \sqrt{\log n} - C_n \alpha_n K_{max} q_n^{1/2} n^{-1/2} \sqrt{\log n}.$$

In the above, the first equation holds because the loss function $L(\omega, \sigma, Y)$ is quadratic in $\omega$ and $u_{A'} = 0$. The inequality is due to Lemma 5.

By the assumptions that $K_{\max} q_n = o(\sqrt{n/\log n})$ and $\lambda_n \sqrt{n/\log n} \to \infty$, we have that $\alpha_n^2 K_{\max} q_n n^{-1/2} \sqrt{\log n} = o(\alpha_n^2)$ and $\alpha_n K_{max} q_n^{1/2} n^{-1/2} \sqrt{\log n} = o(\alpha_n^2)$. Thus,

$$L_n(\omega + \alpha_n u, \sigma, Y, \lambda_n) - L_n(\omega, \sigma, Y, \lambda_n) \geq \frac{1}{4} \alpha_n^2 u_{A}^{T} L''_{n,A}(\omega, \sigma) u_{A} - \bar{C}_1 \alpha_n^2$$

with probability tending to 1. By Lemma 1 (iv), $u_{A}^{T} L''_{n,A} u_{A} \geq \Lambda_{\min}(\omega, \sigma) \|u_{A}\|^2 = \Lambda_{\min}(\omega, \sigma) \bar{C}_1^2$, thus if we take $\bar{C}_1 = 5/\Lambda_{\min}(\omega, \sigma)$, then

$$P\left[ \inf \{ L_n(\omega + \alpha_n u, \sigma, Y, \lambda_n) : u : u_{A'} = 0, \|u\|_2 = \bar{C}_1 \} \right] \to 1.$$ 

Hence a local minimum exists in $\{ \omega : \|\omega - \bar{\omega}\|_2 \leq \bar{C}_1 K_{\max} \sqrt{q_n} \lambda_n \}$ with probability tending to 1. \square
Proof of Lemma 7. Let $\alpha_n = K_{\max} \sqrt{\log n}$. Any $\omega$ in the statement of the lemma can be written as $\omega = \omega + \alpha_n u$, with $u_{A^c} = 0$ and $\|u\|_2 \geq \tilde{C}_2$, where $\tilde{C}_2 > 0$. Note that

$$L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y) = L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y) + \alpha_n L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y)u$$

$$= L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y) + \alpha_n \left(L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y) - L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y)\right)u + \alpha_n L_n^{\prime\prime}(\omega, \tilde{\sigma})u.$$ 

By the triangle inequality and Lemma 5, for any $C_n \to \infty$, $\|L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y)\|_2$ is bounded below by

$$\alpha_n \|L_n^{\prime\prime}(\omega, \tilde{\sigma})u\|_2 - C_n(K_{\max} q_{\max} n^{-1/2} \sqrt{\log n}) - C_n\|u\|_2 \left(\alpha_n K_{\max} q_{\max} n^{-1/2} \sqrt{\log n}\right)$$

with probability tending to 1. Thus, as argued in the proof of Lemma 6, $\alpha_n K_{\max} q_{\max} n^{-1/2} \sqrt{\log n} = o(\alpha_n)$ and $\alpha_n K_{\max} q_{\max} n^{-1/2} \sqrt{\log n} = o(\alpha_n)$, then $\|L_n^{\prime\prime}(\omega, \tilde{\sigma}, Y)\|_2 \geq \frac{1}{2} \alpha_n \|L_n^{\prime\prime}(\omega, \tilde{\sigma})u\|_2$ with probability tending to 1. By Lemma 1(iv), $\|L_n^{\prime\prime}(\omega, \tilde{\sigma})u\|_2 \geq \Lambda_{\min} \|\omega\|_2$. Therefore $\tilde{C}_2$ can be taken as $3/\Lambda_{\min} \|\omega\|_2$.

Proof of Lemma 8. Observe that $\text{Var}(D_{A,ijkl}(\omega, \tilde{\sigma}, Y)) = E(L_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma}, Y) L_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma}, Y)^T) - \bar{L}_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma}) L_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma})^T$. Thus it suffices to show that there exists a constant $M_5 > 0$, such that for all $(i, j, k, l)$, $\lambda_{\max}(L_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma}, Y) L_{ijkl}^{\prime\prime}(\omega, \tilde{\sigma}, Y)^T) \leq M_5$. We use the same notations as in the proof of Lemma 1(v). Note that $L_{ijkl}(\omega, \tilde{\sigma}, Y) = \bar{X}^T X_{ijkl}$ and $L_{ijkl}(\omega, \tilde{\sigma}, Y)^T$ is given by $E[Y_{ik} X_{ijkl} X_{ijkl}^T] + E[Y_{ik} X_{ijkl} X_{ijkl}^T] + E[Y_{ik} X_{ijkl} X_{ijkl}^T] + E[Y_{ik} X_{ijkl} X_{ijkl}^T]$, and for $a \in \mathbb{R}^d$, $a^T L_{ijkl}(\omega, \tilde{\sigma}, Y) L_{ijkl}(\omega, \tilde{\sigma}, Y)^T a = a^T E[Y_{ik} X_{ijkl} X_{ijkl}^T] a_{ik} + a_{ik}^T E[Y_{ik} X_{ijkl} X_{ijkl}^T] a_{ik} + 2a_{ik}^T E[Y_{ik} X_{ijkl} X_{ijkl}^T] a_{ik}$. Since $\sum_{i=1}^{K_i} \sum_{k=1}^{K_i} \|a_{ik}\|^2 = 2 \|a\|^2$, and by Lemma 3 $\lambda_{\max}(E[Y_{ik} X_{ijkl} X_{ijkl}^T]) \leq M_4$ for any $1 \leq i < j \leq p$ and $1 \leq k \leq K_i, 1 \leq l \leq K_j$, the conclusion follows.

APPENDIX B. AUXILIARY RESULTS

**Lemma 9.** Let $X_{ij} \sim N(0, \sigma_i^2)$, $i = 1, \ldots, m$ and $j = 1, \ldots, n$. For each $i$, $X_{i1}, \ldots, X_{in}$ are assumed to be i.i.d., but are arbitrarily dependent across $i$. Then for any sequence $C_n \to \infty$, with probability tending to 1, we have

$$\max_{1 \leq i \leq m} n^{-1} \sum_{j=1}^{n} X_{ij}^2 - \sigma_i^2 \leq C_n \sqrt{(\log m)/n}.$$ 

Proof. Let $Z_{ij} = X_{ij}/\sigma_i$, then for fixed $i$ and $r = 2, 3, \ldots$, we have

$$E[|n^{-1}(Z_{ij}^2 - 1)|^r] \leq \frac{r!}{n^r} E(Z_{ij}^{2r} + 1) \leq (2/n)^r r! = (2/n)^{-2} \frac{4}{n^2 r!}.$$ 

By Lemma 2.2.11 of Van Der Vaart & Wellner (1996), taking $M = 2/n$ and $v = 8/n$, it follows that $P\left(|n^{-1} \sum_{j=1}^{n} Z_{ij}^2 - 1| > x\right) \leq 2e^{-x^2/[2(8/n + 2x/n)]}$. Since $\sigma_i$ are bounded, Lemma 2.2.10 of Van Der Vaart & Wellner (1996) implies that for some $C > 0$, $E\left(\max_{1 \leq i \leq m} n^{-1} \sum_{j=1}^{n} X_{ij}^2 - \sigma_i^2\right) \leq C \sqrt{(\log m)/n}$, which implies the conclusion.
Lemma 10. Let $X_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_{x_i}^2)$ and $Y_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_{y_j}^2)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, and $X_{ij}$ and $Y_{ij}$ are independent for all $i$. Further assume that $0 < \sigma_{x_i}, \sigma_{y_j} \leq \sigma < \infty$. Then for any sequence $C_n \to \infty$, we have $\max_{1 \leq i \leq m} |n^{-1} \sum_{j=1}^{n} X_{ij} Y_{ij}| \leq \sqrt{(\log m)/n}$.

Proof. For fixed $i$ we can observe that

$$E|n^{-1}X_{i1}Y_{i1}|^r = \frac{1}{n^r} E|X_{i1}|^r E|Y_{i1}|^r \leq \frac{2^r \sigma^r (\Gamma(r+1/2))^2}{\pi} \leq (2\sigma/n)^r 4 \sigma^2/\pi n^2 r!.$$  

By Lemma 2.2.11 of Van Der Vaart & Wellner (1996), taking $M = 2\sigma/n$ and $v = 8\sigma^2/\pi n$, we have $P\left(\max_{1 \leq i \leq m} |n^{-1} \sum_{j=1}^{n} X_{ij} Y_{ij}| > x\right) \leq 2e^{-x^2/[2(8\sigma^2/\pi n+2\sigma x/n)]}$. Then by Lemma 2.2.10 of Van Der Vaart & Wellner (1996), for some $C > 0$, $E\left(\max_{1 \leq i \leq m} |n^{-1} \sum_{j=1}^{n} X_{ij} Y_{ij}|\right) \leq C \sqrt{(\log m)/n}$, which implies the conclusion. □

References

Banerjee, O., Ghaoui, L. E. & DaAspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research* 9, 485–516.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.

Khare, K., Oh, S.-Y. & Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 77, 803–825.

Kolar, M., Liu, H. & Xing, E. P. (2014). Graph estimation from multi-attribute data. *Journal of Machine Learning Research* 15, 1713–1750.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34, 1436–1462.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* 846, 735–746.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* 109, 475–494.

Van Der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35.
Fig. 1: Estimation accuracy comparison: total detected edges vs. correctly detected edges with 190 true edges (44%): (a) $K = 3$; (b) $K = 5$; (c) $K = 8$

Fig. 2: Estimation accuracy comparison: total detected edges vs. correctly detected edges with 262 true edges (21%): (a) $K = 3$; (b) $K = 5$; (c) $K = 8$

Fig. 3: Estimation accuracy comparison: total detected edges vs. correctly detected edges with 279 true edges (6%): (a) $K = 3$; (b) $K = 5$; (c) $K = 8$
Fig. 4: Estimation accuracy comparison: total detected edges vs. correctly detected edges with 412 true edges (2%): (a) $K = 3$; (b) $K = 5$; (c) $K = 8$

Fig. 5: Comparison of multivariate and univariate GDP networks
Table 1: Model selection comparison with \( p \) the number of nodes, \( q \) the number of true edges and \( n \) the sample size with the tuning parameter \( \lambda \) optimized by cross-validation. Cases considered below are (i) \( p = 30, \ q = 177 \) (41% density) (ii) \( p = 50, \ q = 137 \) (11% density), (iii) \( p = 100, \ q = 419 \) (8% density), (iv) \( p = 200, \ q = 617 \) (3% density), (v) \( p = 400, \ q = 782 \) (1% density) where the density is \( 100q/(p^2) \) in percentage.

| \( n \) | \( N_t(N_c) \) | mconcord | space | concord | glasso | multi |
|------|----------------|----------|--------|---------|--------|-------|
| (i)  | \( 50 \)       | 58(34)   | 70(35) | 85(42)  | 378(157)| 217(89)|
|      | TPR(PPV)       | 0.19(0.57)| 0.20(0.50) | 0.24(0.49) | 0.89(0.42) | 0.50(0.41) |
|      | MCC            | 0.14     | 0.08   | 0.09    | 0.04   | 0.01  |
|      | \( N_t(N_c) \) | 105(57)  | 47(10) | 46(9)   | 805(105)| 612(69)|
|      | TPR(PPV)       | 0.42(0.54)| 0.07(0.21) | 0.07(0.20) | 0.77(0.13) | 0.50(0.11) |
|      | MCC            | 0.42     | 0.06   | 0.05    | 0.08   | 0.01  |
| (ii) | \( 100 \)      | 191(64)  | 286(58) | 280(59) | 923(122)| 525(69)|
|      | TPR(PPV)       | 0.47(0.34)| 0.42(0.20) | 0.43(0.21) | 0.89(0.13) | 0.50(0.13) |
|      | MCC            | 0.30     | 0.16   | 0.17    | 0.11   | 0.05  |
|      | \( N_t(N_c) \) | 248(87)  | 202(40) | 267(51) | 2389(274)| 2501(211)|
|      | TPR(PPV)       | 0.21(0.35)| 0.10(0.20) | 0.12(0.19) | 0.65(0.11) | 0.50(0.08) |
|      | MCC            | 0.22     | 0.08   | 0.09    | 0.10   | 0.00  |
| (iii)| \( 200 \)      | 613(200) | 814(170)| 1005(196) | 1066(204)| 2380(201)|
|      | TPR(PPV)       | 0.48(0.33)| 0.41(0.21) | 0.47(0.20) | 0.49(0.19) | 0.48(0.08) |
|      | MCC            | 0.33     | 0.20   | 0.20    | 0.20   | 0.00  |
|      | \( N_t(N_c) \) | 481(112) | 84(12) | 133(18) | 5657(306)| 4797(240)|
|      | TPR(PPV)       | 0.18(0.23)| 0.02(0.14) | 0.03(0.14) | 0.50(0.05) | 0.39(0.05) |
|      | MCC            | 0.18     | 0.04   | 0.05    | 0.08   | 0.06  |
| (iv) | \( 100 \)      | 1250(300)| 892(143) | 976(151) | 6357(426)| 4392(226)|
|      | TPR(PPV)       | 0.49(0.24)| 0.23(0.16) | 0.24(0.15) | 0.69(0.07) | 0.37(0.05) |
|      | MCC            | 0.31     | 0.16   | 0.16    | 0.14   | 0.06  |
|      | \( N_t(N_c) \) | 764(129) | 31(3)  | 54(6)   | 14283(326)| 10229(259)|
|      | TPR(PPV)       | 0.16(0.17)| 0.00(0.10) | 0.00(0.11) | 0.42(0.02) | 0.33(0.03) |
|      | MCC            | 0.16     | 0.02   | 0.03    | 0.06   | 0.06  |
| (v)  | \( 200 \)      | 2063(378)| 396(62) | 404(53) | 16092(480)| 9648(240)|
|      | TPR(PPV)       | 0.48(0.18)| 0.08(0.16)| 0.07(0.13)| 0.61(0.03) | 0.31(0.02) |
|      | MCC            | 0.29     | 0.11   | 0.09    | 0.10   | 0.06  |
Table 2: Summary statistics for protein, gene and gene-protein networks

|                       | Protein network | Gene network | Gene-protein network |
|-----------------------|-----------------|--------------|----------------------|
| Number of edges       | 798             | 784          | 531                  |
| Density (%)           | 18              | 18           | 12                   |
| Maximum degree        | 24              | 24           | 20                   |
| Average node degree   | 16.98           | 16.68        | 11.30                |

Table 3: Top 20 most connected nodes for three networks (sorted by decreasing degrees)

| Gene-protein network | Entrez ID | Gene name | Protein network | Entrez ID | Gene name | Gene network | Entrez ID | Gene name |
|----------------------|-----------|-----------|-----------------|-----------|-----------|--------------|-----------|-----------|
| 302 ANXA2            |           |           | 4179 CD46       |           |           | 2064 ERBB2   |           |           |
| 7280 TUBB2A          |           |           | 983 CDK1        |           |           | 5605 MAP2K2  |           |           |
| 1398 CRK             |           |           | 3265 HRAS       |           |           | 307 ANXA4    |           |           |
| 4255 MGMT            |           |           | 3716 JAK1       |           |           | 5578 PRKCA   |           |           |
| 5578 PRKCA           |           |           | 10270 AKAP8     |           |           | 1173 AP2M1   |           |           |
| 5925 RB1             |           |           | 354 KLK3        |           |           | 1828 DSG1    |           |           |
| 9564 BCAR1           |           |           | 1019 CDK4       |           |           | 4179 CD46    |           |           |
| 307 ANXA4            |           |           | 6776 STAT5A     |           |           | 9961 MVP     |           |           |
| 354 KLK3             |           |           | 9564 BCAR1      |           |           | 1000 CDH2    |           |           |
| 2064 ERBB2           |           |           | 1398 CRK        |           |           | 2932 GSK3B   |           |           |
| 4163 MCC             |           |           | 3667 IRS1       |           |           | 4176 MCM7    |           |           |
| 6778 STAT6           |           |           | 4830 NME1       |           |           | 4436 MSH2    |           |           |
| 7299 TYR             |           |           | 307 ANXA4       |           |           | 5970 RELA    |           |           |
| 1173 AP2M1           |           |           | 1173 AP2M1      |           |           | 999 CDH1     |           |           |
| 983 CDK1             |           |           | 2017 CTTN       |           |           | 1001 CDH3    |           |           |
| 1001 CDH3            |           |           | 4255 MGMT       |           |           | 1398 CRK     |           |           |
| 1499 CTNNB1          |           |           | 1001 CDH3       |           |           | 2335 FN1     |           |           |
| 3716 JAK1            |           |           | 1020 CDK5       |           |           | 5925 RB1     |           |           |
| 4179 CD46            |           |           | 3308 HSPA4      |           |           | 7280 TUBB2A  |           |           |
| 4830 NME1            | 4176 MCM7  |           | 7299 TYR        |           |           |              |           |           |