Journal of Machine Learning Research 17 (2016) 1-51 Submitted 1/16; Revised 8/16; Published 8/16

Penalized Maximum Likelihood Estimation of Multi-layered Gaussian Graphical Models

Jiahe Lin * JIAHELIN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Sumanta Basu * SUMBOSE@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Moulinath Banerjee MOULIB@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

George Michailidis † GMICHAIL@UFL.EDU
Department of Statistics and Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611, USA

Editor: Jie Peng

Abstract

Analyzing multi-layered graphical models provides insight into understanding the conditional relationships among nodes within layers after adjusting for and quantifying the effects of nodes from other layers. We obtain the penalized maximum likelihood estimator for Gaussian multi-layered graphical models, based on a computational approach involving screening of variables, iterative estimation of the directed edges between layers and undirected edges within layers and a final refitting and stability selection step that provides improved performance in finite sample settings. We establish the consistency of the estimator in a high-dimensional setting. To obtain this result, we develop a strategy that leverages the biconvexity of the likelihood function to ensure convergence of the developed iterative algorithm to a stationary point, as well as careful uniform error control of the estimates over iterations. The performance of the maximum likelihood estimator is illustrated on synthetic data.

Keywords: graphical models, penalized likelihood, block coordinate descent, convergence, consistency

∗. Equal Contribution
†. Corresponding Author. Post Address: 205 Griffin Floyd Hall, 1 University Ave, Gainesville, FL, 32611.

1. Introduction.

The estimation of directed and undirected graphs from high-dimensional data has received a lot of attention in the machine learning and statistics literature (e.g., see Bühlmann and Van De Geer, 2011, and references therein), due to their importance in diverse applications including understanding of biological processes and disease mechanisms, financial systems stability and social interactions, just to name a few (Sachs et al., 2005; Wang et al., 2007; Sobel, 2000). In the case of undirected graphs, the edges capture conditional dependence relationships between the nodes, while for directed graphs they are used to model causal relationships (Bühlmann and Van De Geer, 2011).

However, in a number of applications the nodes can be *naturally partitioned* into sets that exhibit interactions both between them and amongst them. As an example, consider an experiment where one has collected data for both genes and metabolites for the same set of patient specimens. In this case, we have three types of interactions between genes and metabolites: regulatory interactions between the two of them and co-regulation within the gene and within the metabolic compartments. The latter two types of relationships can be expressed through undirected graphs within the sets of genes and metabolites, respectively, while the regulation of metabolites by genes corresponds to directed edges. Note that in principle there are feedback mechanisms from the metabolic compartment to the gene one, but these are difficult to detect and adequately estimate in the absence of carefully collected time course data. Another example comes from the area of financial economics, where one collects data on returns of financial assets (e.g. stocks, bonds) and also on key macroeconomic indicators (e.g. interest rate, prices indices, various measures of money supply and various unemployment indices). Once again, over short time periods there is influence from the economic variables to the returns (directed edges), while there are co-dependence relationships between the asset returns and the macroeconomic variables, respectively, that can be modeled as undirected edges.

Technically, such *layered* network structures correspond to multi-partite graphs that possess undirected edges and exhibit a directed acyclic graph structure between the layers, as depicted in Figure 1, where we use directed solid edges to denote the dependencies across layers and dashed undirected edges to denote within-layer conditional dependencies.

Selected properties of such so-called *chain graphs* have been studied in the work of Drton and Perlman (2008), with an emphasis on two alternative Markov properties including the LWF Markov property (Lauritzen and Wermuth, 1989; Frydenberg, 1990) and the AMP Markov property (Andersson et al., 2001).

While layered networks being interesting from a theoretical perspective and having significant scope for applications, their estimation has received little attention in the literature. Note that for a 2-layered structure, the directed edges can be obtained through a multivariate regression procedure, while the undirected edges in both layers through existing procedures for graphical models (for more technical details see Section 2.2). This is the strategy leveraged in the work of Rothman et al. (2010), where for a 2-layered network structure they proposed a multivariate regression with covariance estimation (MRCE) method for estimating the undirected edges in the second layer and the directed edges between them. A block coordinate descent algorithm was introduced to estimate the directed edges, while the popular glasso estimator (Friedman et al., 2008) was used for the undirected edges.

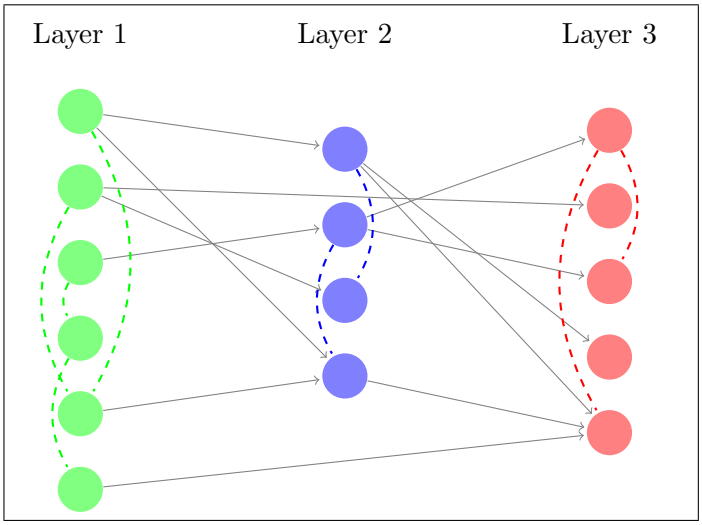


Figure 1: Diagram for a three-layered network

However, this method does not scale well according to the simulation results presented and no theoretical properties of the estimates were provided.

In follow-up work, Yin and Li (2011) used a cyclic block coordinate descent algorithm and claimed convergence to a stationary point leveraging a result in Tseng (2001) (see Proposition 2 in the Supplemental material). Unfortunately, a key assumption in Tseng (2001) -namely, that a corresponding coordinate wise optimization problem that is given by a high-dimensional lasso regression has unique minimum- fails and hence the convergence result does not go through.

In related work, Lee and Liu (2012) proposed the Plug-in Joint Weighted Lasso (PWL) and the Plug-in Joint Graphical Weighted Lasso (PWGL) estimator for estimating the same 2-layered structure, where they use a weighted version of the algorithm in Rothman et al. (2010) and also provide theoretical results for the low dimensional setting, where the number of samples exceeds the number of potential directed and undirected edges to be estimated. Finally, Cai et al. (2012) proposed a method for estimating the same 2-layered structure and provided corresponding theoretical results in the high dimensional setting. The Dantzig-type estimator (Candes and Tao, 2007) was used for the regression coefficients and the corresponding residuals were used as surrogates, for obtaining the precision matrix through the CLIME estimator (Cai et al., 2011). In another line of work (Sohn and Kim, 2012; Yuan and Zhang, 2014; McCarter and Kim, 2014), structured sparsity of directed edges was considered and the edges were estimated with a different parametrization of the objective function. We further elaborate on the connections of our work with these three papers in Section 5.

The above work assumed a Gaussian distribution for the data, in more recent work by Yang et al. (2014), the authors constructed the model under a general *mixed graphical model* framework, which allows each node-conditional distribution to belong to a potentially different univariate exponential family. In particular, with an underlying *mixed MRF* graph structure, instead of maximizing the joint likelihood, the authors proposed to esti-

mate the homogeneous and heterogeneous neighborhood for each node, by obtaining the ℓ_1 regularized M-estimator of the node-conditional distribution parameters, using traditional approaches (e.g. Meinshausen and Bühlmann, 2006) for neighborhood estimation. However, rather than estimating directed edges directly, the directed edges are obtained from a nonlinear transformation of the estimated homogeneous and heterogeneous neighborhood, whose sparsity pattern gets compromised during the process.

In this work, we obtain the regularized maximum likelihood estimator under a sparsity assumption on both directed and undirected parameters for multi-layered Gaussian graphical models and establish its consistency properties in a high-dimensional setting. As discussed in Section 3, the problem is *not jointly convex* on the parameters, but convex on selected subsets of them. Further, it turns out that the problem is *biconvex* if we consider a recursive multi-stage estimation approach that at each stage involves only regression parameters (directed edges) from preceding layers and precision matrix parameters (undirected edges) for the *last layer considered* in that stage. Hence, we decompose the multi-layer network structure estimation into a sequence of 2-layer problems that allows us to establish the desired results. Leveraging the biconvexity of the 2-layer problem, we establish the convergence of the iterates to the maximum-likelihood estimator, which under certain regularity conditions is arbitrarily close to the true parameters. The theoretical guarantees provided require a *uniform control* of the precision of the regression and precision matrix parameters, which poses a number of theoretical challenges resolved in Section 3.

In summary, despite the lack of overall convexity, we are able to provide theoretical guarantees for the MLE in a high dimensional setting. We believe that the proposed strategy is generally applicable to other non-convex statistical estimation problems that can be decomposed to two biconvex problems. Further, to enhance the numerical performance of the MLE in finite (and small) sample settings, we introduce a screening step that selects active nodes for the iterative algorithm used and that leverages recent developments in the high-dimensional regression literature (e.g., Van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhang and Zhang, 2014). We also post-process the final MLE estimate through a stability selection procedure. As mentioned above, the screening and stability selection steps are beneficial to the performance of the MLE in finite samples and hence recommended for similarly structured problems.

The remainder of the paper is organized as follows. In Section 2, we introduce the proposed methodology, with an emphasis on how the multi-layered network estimation problem is decomposed into a sequence of two-layered network estimation problem(s). In Section 3, we provide theoretical guarantees for the estimation procedure posited. In particular, we show consistency of the estimates and convergence of the algorithm, under a number of common assumptions in high-dimensional settings. In Section 4, we show the performance of the proposed algorithm with simulation results under different simulation settings, and introduce several acceleration techniques which speed up the convergence of the algorithm and reduce the computing time in practical settings.

2. Problem Formulation.

Consider an M-layered Gaussian graphical model. Suppose there are p_m nodes in Layer m, denoted by

$$\boldsymbol{X}^m = (X_1^m, \cdots, X_{p_m}^m)', \quad \text{for } m = 1, \cdots, M.$$

The structure of the model is given as follows:

- Layer 1. $\boldsymbol{X}^1 = (X_1^1, \cdots, X_{p_1}^1)' \sim \mathcal{N}(0, \Sigma^1)$.
- Layer 2. For $j = 1, \cdots, p_2$: $X_j^2 = (B_j^{12})'\boldsymbol{X}^1 + \epsilon_j^2$, with $B_j^{12} \in \mathbb{R}^{p_1}$, and $\boldsymbol{\epsilon}^2 = (\epsilon_1^2, \cdots, \epsilon_{p_2}^2)' \sim \mathcal{N}(0, \Sigma^2)$.

  $\vdots$

- Layer M. For $j = 1, 2, \cdots, p_M$:

$$X_j^M = \sum_{m=1}^{M-1} \{(B_j^{mM})'\boldsymbol{X}^m\} + \epsilon_j^M, \quad \text{where } B_j^{mM} \in \mathbb{R}^{p_m} \quad \text{for } m = 1, \cdots, M-1,$$

  and $\boldsymbol{\epsilon}^M = (\epsilon_1^M, \cdots, \epsilon_{p_M}^M)' \sim \mathcal{N}(0, \Sigma^M)$.

The parameters of interest are *all directed edges* that encode the dependencies across layers, that is,

$$B^{st} := \begin{bmatrix} B_1^{st} & \cdots & B_{p_t}^{st} \end{bmatrix}, \quad \text{for } 1 \le s < t \le M,$$

and *all undirected edges* that encode the conditional dependencies within layers after adjusting for the effects from directed edges, that is:

$$\Theta^m := (\Sigma^m)^{-1}, \quad \text{for } m = 1, \cdots, M.$$

It is assumed that B^{st} and Θ^m are *sparse* for all $1, \ldots, M$ and $1 \le s < t \le M$.

Given centered data for all M layers, denoted by $X^m = [X_1^m, \cdots, X_{p_m}^m] \in \mathbb{R}^{n \times p_m}$ for all $m = 1, \cdots, M$, we aim to obtain the MLE for all $B^{st}, 1 \le s < t \le M$ and all $\Theta^m, m = 1, \cdots, M$ parameters. Henceforth, we use $\boldsymbol{X}^m$ to denote random vectors, and X_j^m to denote the jth column in the data matrix $X_{n \times p_m}^m$ whenever there is no ambiguity.

Through Markov factorization (Lauritzen, 1996), the full log-likelihood function can be decomposed as

$$\begin{aligned} \ell(X^m; B^{st}, \Theta^m, 1 \le s < t \le M, 1 \le m \le M) &= \ell(X^M | X^{M-1}, \cdots, X^1; B^{1M}, \cdots, B^{M-1,M}, \Theta^M) \\ &\quad + \ell(X^{M-1} | X^{M-2}, \cdots, X^1; B^{1M-1}, \cdots, B^{M-2,M-1}, \Theta^{M-1}) \\ &\quad + \cdots + \ell(X^2 | X^1; B^{12}, \Theta^2) + \ell(X^1; \Theta^1) \\ &= \ell(X^1; \Theta^1) + \textstyle\sum_{m=2}^{M} \ell(X^m | X^1, \cdots, X^{m-1}; B^{1m}, \cdots, B^{m-1,m}, \Theta^m). \end{aligned}$$

Note that the summands share no common parameters, which enables us to maximize the likelihood with respect to individual parameters in the M terms separately. More importantly, by conditioning Layer m nodes on nodes in its previous $(m-1)$ layers, we can treat Layer m nodes as the"response" layer, and all nodes in the previous $(m-1)$ layer

combined as a super "parent" layer. If we ignore the structure within the bottom layer (X^1) for the moment, the M-layered network can be viewed as $(M-1)$ two-layered networks, each comprising a response layer and a parent layer. Thus, the network structure in Figure 1 can be viewed as a 2 two-layered network: for the first network, Layer 3 is the response layer, while Layers 1 and 2 combined form the "parent" layer; for the second network, Layer 2 is the response layer, and Layer 1 is the "parent" layer. Therefore, the problem for estimating all $\binom{M}{2}$ coefficient matrices and M precision matrices can be translated into estimating $(M-1)$ two-layered network structures with directed edges from the parent layer to the response layer, and undirected edges within the response layer, and finally estimating the undirected edges within the bottom layer separately.

Since all estimation problems boil down to estimating the structure of a 2-layered network, we focus the technical discussion on introducing our proposed methodology for a 2-layered network setting,[1]. The theoretical results obtained extend in a straightforward manner to an M-layered Gaussian graphical model.

Remark 1. For the M-layer network structure, we impose certain identifiability-type condition on the largest "parent" layer (encompassing $M-1$ layers), so that the directed edges of the entire network are estimable. The imposed condition translates into a minimum eigenvalue-type condition on the population precision matrix within layers, and conditions on the magnitude of dependencies across layers. Intuitively, consider a three-layered network: if $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ are highly correlated, then the proposed (as well as any other) method will exhibit difficulties in distinguishing the effect of $\boldsymbol{X}^1$ on $\boldsymbol{X}^3$ from that of $\boldsymbol{X}^2$ on $\boldsymbol{X}^3$. The (group) identifiability-type condition is thus imposed to obviate such circumstances. An in-depth discussion on this issue is provided in Section 3.4.

2.1 A Two-layered Network Setup.

Consider a two-layered Gaussian graphical model with p_1 nodes in the first layer, denoted by $\boldsymbol{X} = (X_1, \cdots, X_{p_1})'$, and p_2 nodes in the second layers, denoted by $\boldsymbol{Y} = (Y_1, \cdots, Y_{p_2})'$. The model is defined as

- $\boldsymbol{X} = (X_1, \cdots, X_{p_1})' \sim \mathcal{N}(0, \Sigma_X)$.
- For $j = 1, 2, \cdots, p_2$: $Y_j = B_j' \boldsymbol{X} + \epsilon_j$, $B_j \in \mathbb{R}^{p_1}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_{p_2})^\top \sim \mathcal{N}(0, \Sigma_\epsilon)$.

The parameters of interest are: $\Theta_X := \Sigma_X^{-1}, \Theta_\epsilon := \Sigma_\epsilon^{-1}$ and $B = [B_1, \cdots, B_{p_2}]$. As with most estimation problems in the high dimensional setting, we assume these parameters to be sparse.

Now given data $X = [X_1, \cdots, X_{p_1}] \in \mathbb{R}^{n \times p_1}$ and $Y = [Y_1, \cdots, Y_{p_2}] \in \mathbb{R}^{n \times p_2}$, both centered, we would like to use the penalized maximum likelihood approach to obtain estimates for Θ_X, Θ_ϵ and B. Throughout this paper, we use X, Y and E to denote the size-n realizations of the random vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$, respectively. Also, with a slight abuse of notation, we use $X_i, i = 1, 2, \cdots, p_1$ and $Y_j, j = 1, 2, \cdots, p_2$ to denote the columns of the data matrix X and Y, respectively, whenever there is no ambiguity.

1. In Appendix D we give a detail example on how our proposed method works under a 3-layered network setting.

The full log-likelihood can be written as

$$\ell(X, Y; B, \Theta_\epsilon, \Theta_X) = \ell(Y|X; \Theta_\epsilon, B) + \ell(X; \Theta_X) \tag{1}$$

Note that the first term only involves Θ_ϵ and B, and the second term only involves Θ_X. Hence, (1) can be maximized by maximizing $\ell(Y|X)$ w.r.t. (Θ_ϵ, B), and maximizing $\ell(X)$ w.r.t. Θ_X, respectively. $\widehat{\Theta}_X$ can be obtained using traditional methods for estimating undirected graphs, e.g., the Graphical Lasso (Friedman et al., 2008) or the Nodewise Regression prcoedure (Meinshausen and Bühlmann, 2006). Therefore, the rest of this paper will mainly focus on obtaining estimates for Θ_ϵ and B. In the next subsection, we introduce our estimation procedure for obtaining the MLE for Θ_ϵ and B.

Remark 2. Our proposed method is targeted towards maximizing $\ell(Y|X; \Theta_\epsilon, B)$ (with proper penalization) in (1) only, which gives the estimates for across-layers dependencies between the response layer and the parent layer, as well as estimates for the conditional dependencies within the response layer each time we solve a 2-layered network estimation problem. For an M-layered estimation problem, the maximization regarding $\ell(X; \Theta_X)$ occurs only when we are estimating the within-layer conditional dependencies for the bottom layer.

2.2 Estimation Algorithm.

The conditional likelihood for response Y given X can be written as

$$L(Y|X) \quad = (\tfrac{1}{\sqrt{2\pi}})^{np_2} |\Sigma_\epsilon \otimes I_n|^{-1/2} \exp\left\{-\tfrac{1}{2}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})^\top (\Sigma_\epsilon \otimes I_n)^{-1} (\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})\right\},$$

where $\mathcal{Y} = vec(Y_1, \cdots, Y_{p_2})$, $\mathcal{X} = I_{p_2} \otimes X$ and $\beta = vec(B_1, \cdots, B_{p_2})$. After writing out the Kronecker product, the log-likelihood can be written as

$$\ell(Y|X) = \text{constant} + \frac{n}{2} \log\det \Theta_\epsilon - \frac{1}{2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_2} \sigma_\epsilon^{ij} (Y_i - XB_i)^\top (Y_j - XB_j).$$

Here, σ_ϵ^{ij} denotes the ij-th entry of Θ_ϵ. With ℓ_1 penalization which induces sparsity, we formulate the following optimization problem using penalized log-likelihood, which was initially proposed in Rothman et al. (2010), and has also been examined in Lee and Liu (2012):

$$\min_{\substack{B \in \mathbb{R}^{p_1 \times p_2} \\ \Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}} \left\{ \frac{1}{n} \sum_{j=1}^{p_2} \sum_{i=1}^{p_2} \sigma_\epsilon^{ij} (Y_i - XB_i)^\top (Y_j - XB_j) - \log\det \Theta_\epsilon + \lambda_n \sum_{j=1}^{p_2} \|B_j\|_1 + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\}, \tag{2}$$

and the first term in (2) can be equivalently written as

$$\text{tr}\left\{ \tfrac{1}{n} \begin{bmatrix} (Y_1 - XB_1)^\top \\ \vdots \\ (Y_{p_2} - XB_{p_2})^\top \end{bmatrix} \begin{bmatrix} (Y_1 - XB_1) & \cdots & (Y_{p_2} - XB_{p_2}) \end{bmatrix} \Theta_\epsilon \right\} := \text{tr}(S\Theta_\epsilon).$$

where S is defined as the sample covariance matrix of $E \equiv Y - XB$. This gives rise to the following optimization problem:

$$\min_{\substack{B\in\mathbb{R}^{p_1\times p_2}\\ \Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}} \left\{ \operatorname{tr}(S\Theta_\epsilon) - \log\det\Theta_\epsilon + \lambda_n \sum_{j=1}^{p_2} \|B_j\|_1 + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\} \equiv f(B,\Theta_\epsilon), \tag{3}$$

where $\|\Theta\|_{1,\text{off}}$ is the absulote sum of the off-diagonal entries in Θ, λ_n and ρ_n are both positive tuning parameters.

Note that the objective function (3) is *not jointly convex* in (B,Θ_ϵ), but only convex in B for fixed Θ_ϵ and in Θ_ϵ for fixed B; hence, it is bi-convex, which in turn implies that the proposed algorithm may fail to converge to the global optimum, especially in settings where $p_1 > n$, as pointed out by Lee and Liu (2012). As is the case with most non-convex problems, good initial parameters are beneficial for fast convergence of the algorithm, a fact supported by our numerical work on the present problem. Further, a good initialization is critical in establishing convergence of the algorithm for this problem (see Section 3.1). To that end, we introduce a *screening step* for obtaining a good initial estimate for B. The theoretical justification for employing the screening step is provided in Section 3.3.

An outline of the computational procedure is presented in Algorithm 1, while the details of each step involved are discussed next.

Screening. For each variable $Y_j, j = 1,\cdots,p_2$ in the response layer, regress Y_j on X via the de-biased Lasso procedure proposed by Javanmard and Montanari (2014). The output consists of the p-value(s) for each predictor in each regression, denoted by P_j, with $P_j \in [0,1]^{p_1}$. To control the family-wise error rate of the estimates, we do a Bonferroni correction at level α: define $\alpha^\star = \alpha/p_1p_2$ and set $B_{j,k} = 0$ if the p-value obtained for the k'th predictor in the j'th regression $P_{j,k}$ exceeds $\alpha^\star$. Further, let

$$\mathcal{B}_j = \{B_j \in \mathbb{R}^{p_1} : B_{j,k} = 0 \text{ if } k \in \widehat{S}_j^c\} \subseteq \mathbb{R}^{p_1}, \tag{4}$$

where $\widehat{S}_j$ is the collection of indices for those predictors deemed "active" for response Y_j:

$$\widehat{S}_j = \{k : P_{j,k} < \alpha^\star\}, \quad \text{for } j = 1,\cdots,p_2.$$

Therefore, subsequent estimation of the elements of B will be restricted to $\mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}$.

Alternating Search. In this step, we utilize the bi-convexity of the problem and estimate B and Θ_ϵ by minimizing in an iterative fashion the objective function with respect to (w.r.t.) one set of parameters, while holding the other set fixed within each iteration.

As with most iterative algorithms, we need an initializer; for $\widehat{B}^{(0)}$ it corresponds to a Lasso/Ridge regression estimate with a small penalty, while for $\widehat{\Theta}_\epsilon$ we use the Graphical Lasso procedure applied to the residuals obtained from the first stage regression. That is, for each $j = 1,\cdots,p_2$,

$$\widehat{B}_j^{(0)} = \underset{B_j\in\mathcal{B}_j}{\operatorname{argmin}} \left\{ \|Y_j - XB_j\|_2^2 + \lambda_n^0 \|B_j\|_1 \right\}, \tag{5}$$

Algorithm 1: Computational procedure for estimating B and Θ_ϵ

Input : Data from the parent layer X and the response layer Y.

Screening:

  for $j = 1, \cdots, p_2$ **do**

    regress Y_j on X using the de-biased Lasso procedure in Javanmard and Montanari (2014) and obtain the corresponding vector of p-values P_j;

  end

  obtain adjusted p-values $\widetilde{P}_j$ by applying Bonferroni correction to $\text{vec}(P_1, \cdots, P_j)$;
  determine the support set $\mathcal{B}_j$ for each regression using (4).

Initialization:

  Initialize column $j = 1, \cdots, p_2$ of $\widehat{B}^{(0)}$ by solving (5).
  Initialize $\widehat{\Theta}_\epsilon^{(0)}$ by solving (9) using the graphical lasso (Friedman et al., 2008).

while $|f(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}) - f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k+1)})| \geq \epsilon$ **do**

  update $\widehat{B}$ with (6);

  update $\widehat{\Theta}_\epsilon$ with (8);

end

Refitting B and Θ_ϵ:

  for $j = 1, \cdots, p_2$ **do**

    Obtain the refitted $\widetilde{B}_j$ using (9);

  end

  re-estimate $\widetilde{\Theta}_\epsilon$ using (10) with W coming from stability selection.

Output: Final Estimates $\widetilde{B}$ and $\widetilde{\Theta}_\epsilon$.

where λ_n^0 is some small tuning parameter for initialization, and set $\widehat{E}_j^{(0)} := Y_j - X\widehat{B}_j^{(0)}$. An initial estimate for $\widehat{\Theta}_\epsilon$ is then given by solving for the following optimization problem with the graphical lasso (Friedman et al., 2008) procedure:

$$\widehat{\Theta}_\epsilon^{(0)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\text{argmin}} \left\{ \log\det\Theta_\epsilon - \text{tr}(\widehat{S}^{(0)}\Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\},$$

where $\widehat{S}^{(0)}$ is the sample covariance matrix based on $(\widehat{E}_1^{(0)}, \cdots, \widehat{E}_{p_2}^{(0)})$.

Next, we use an alternating block coordinate descent algorithm with ℓ_1 penalization to reach a stationary point of the objective function (3).

– Update B as

$$\widehat{B}^{(k+1)} = \underset{B \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}}{\text{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{p_2} \sum_{j=1}^{p_2} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (Y_i - XB_i)^\top (Y_j - XB_j) + \lambda_n \sum_{j=1}^{p_2} \|B_j\|_1 \right\}, \tag{6}$$

which can be obtained by cyclic coordinate descent w.r.t each column B_j of B, that is, update each column B_j by:

$$\widehat{B}_j^{(t+1)} = \underset{B_j \in \mathcal{B}_j}{\operatorname{argmin}} \left\{ \frac{(\widehat{\sigma}_\epsilon^{jj})^{(k)}}{n} \|Y_j + r_j^{(t+1)} - XB_j\|_2^2 + \lambda_n \|B_j\|_1 \right\}, \tag{7}$$

where

$$r_j^{(t+1)} = \frac{1}{(\widehat{\sigma}_\epsilon^{jj})^{(k)}} \left[\sum_{i=1}^{j-1} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (Y_i - X\widehat{B}_i^{(t+1)}) + \sum_{i=j+1}^{p_2} (\widehat{\sigma}_\epsilon^{ij})^{(k)} (Y_i - X\widehat{B}_i^{(t)}) \right],$$

and iterate over all columns until convergence. Here, we use k to index the outer iteration while minimizing w.r.t. B or Θ_ϵ, and use t to index the inner iteration while cyclically minimizing w.r.t. each column of B.

– Update Θ_ϵ as

$$\widehat{\Theta}_\epsilon^{(k+1)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\operatorname{argmin}} \left\{ \log\det \Theta_\epsilon - \operatorname{tr}(\widehat{S}^{(k+1)} \Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\}, \tag{8}$$

where $\widehat{S}^{(k+1)}$ is the sample covariance matrix based on $\widehat{E}_j^{(k+1)} = Y_j - X\widehat{B}_j^{(k+1)}, j = 1, \cdots, p_2$.

Refitting and Stabilizing. As noted in the introduction, this step is beneficial in applications, especially when one deals with large scale multi-layer networks and relatively smaller sample sizes. Denote the solution obtained by the above iterative procedure by B^∞ and Θ_ϵ^∞. For each $j = 1, \cdots, p_2$, set $\widetilde{\mathcal{B}}_j = \{B_j : B_{j,i} = 0 \text{ if } B_{j,i}^\infty = 0, B_j \in \mathbb{R}^{p_1}\}$ and the final estimate for B_j is given by ordinary least squares:

$$\widetilde{B}_j = \underset{B_j \in \widetilde{\mathcal{B}}_j}{\operatorname{argmin}} \|Y_j - XB_j\|^2. \tag{9}$$

For Θ_ϵ, we obtain the final estimate by a combination of stability selection (Meinshausen and Bühlmann, 2010) and graphical lasso (Friedman et al., 2008). That is, after obtaining the refitted residuals $\widetilde{E}_j := Y_j - X\widetilde{B}_j, j = 1, \cdots, p_2$, based on the stability selection procedure with the graphical lasso, we obtain the stability path, or probability matrix W for each edge, which records the proportion of each edge being selected based on bootstrapped samples of $\widetilde{E}_j$'s. Then, using this probability matrix W as a weight matrix, we obtain the final estimate of $\widetilde{\Theta}_\epsilon$ as follow:

$$\widetilde{\Theta}_\epsilon = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\operatorname{argmin}} \left\{ \log\det \Theta_\epsilon - \operatorname{tr}(\widetilde{S}\Theta_\epsilon) + \widetilde{\rho}_n \|(1 - W) * \Theta_\epsilon\|_{1,\text{off}} \right\}, \tag{10}$$

where we use $*$ to denote the element-wise product of two matrices, and $\widetilde{S}$ is the sample covariance matrix based on the refitted residuals $\widetilde{E}$. Again, (10) can be solved by the graphical lasso procedure (Friedman et al., 2008), with $\widetilde{\rho}_n$ properly chosen.

2.3 Tuning Parameter Selection.

To select the tuning parameters (λ_n, ρ_n), we use the Bayesian Information Criterion(BIC), which is the summation of a goodness-of-fit term (log-likelihood) and a penalty term. The explicit form of BIC (as a function of B and Θ_ϵ) in our setting is given by

$$\mathrm{BIC}(B, \Theta_\epsilon) = -\log\det\Theta_\epsilon + \mathrm{tr}(S\Theta_\epsilon) + \frac{\log n}{n}\left(\frac{\|\Theta_\epsilon\|_0 - p_2}{2} + \|B\|_0\right)$$

where

$$S := \frac{1}{n}\begin{bmatrix} (Y_1 - XB_1)^\top \\ \vdots \\ (Y_{p2} - XB_{p_2})^\top \end{bmatrix}\begin{bmatrix}(Y_1 - XB_1) & \cdots & (Y_{p_2} - XB_{p_2})\end{bmatrix},$$

and $\|\Theta_\epsilon\|_0$ is the total number of nonzero entries in Θ_ϵ. Here we penalize the non-zero elements in the upper-triangular part of Θ_ϵ and the non-zero ones in B. We choose the combination (λ_n^*, ρ_n^*) over a grid of (λ, ρ) values, and (λ_n^*, ρ_n^*) should minimize the BIC evaluated at $(B^\infty, \Theta_\epsilon^\infty)$.

3. Theoretical Results.

In this section, we establish a number of theoretical results for the proposed iterative algorithm. We focus the presentation on the two-layer structure, since as explained in the previous section the multi-layer estimation problem decomposes to a series of two-layer ones. As mentioned in the introduction, one key challenge for establishing the theoretical results comes from the fact that the objective function (3) is not jointly convex in B and Θ_ϵ. Consequently, if we simply used properties of block-coordinate descent algorithms, we would not be able to provide the necessary theoretical guarantees for the estimates we obtain. On the other hand, the biconvex nature of the objective function allows us to establish convergence of the alternating algorithm to a stationary point, provided it is initialized from a point close enough to the true parameters. This can be accomplished using a Lasso-based initializer for B and Θ_ϵ as previously discussed. The details of algorithmic convergence are presented in Section 3.1.

Another technical challenge is that each update in the alternating search step relies on estimated quantities—namely the regression and precision matrix parameters—rather than the raw data, whose estimation precision needs to be controlled *uniformly* across all iterations. The details of establishing consistency of the estimates for both fixed and random realizations are given in Section 3.2.

Next, we outline the structure of this section. In Section 3.1 Theorem 1, we show that for any fixed set of realization of X and E, [2] the iterative algorithm is guaranteed to converge to a stationary point if estimates for all iterations lie in a compact ball around the true value of the parameters. In Section 3.2, we show in Theorem 4 that for any random X and E, with high probability, the estimates for all iterations lie in a compact ball around the true value of the parameters. Then in Section 3.3, we show that asymptotically with

2. We actually observe X and Y, which is given by a corresponding set of realization in X and E based on the model.

$\log(p_1 p_2)/n \to 0$, while keeping the family-wise type I error under some pre-specified level, the screening step correctly identifies the true support set for each of the regressions, based upon which the iterative algorithm is provided with an initializer that is close to the true value of the parameters. Finally in Section 3.4, we provide sufficient conditions for both directed and undirected edges to be identifiable (estimable) for multi-layered network.

To aid the readability of the main results, we only present statements of theorems and propositions, while all proofs are relegated to the Appendix (Section A and B).

Throughout this section, to distinguish the estimates from the true values, we use B^* and Θ_ϵ^* to denote the true values.

3.1 Convergence of the Iterative Algorithm.

In this subsection, we prove that the proposed block relaxation algorithm converges to a stationary point for any fixed set of data, provided that the estimates for all iterations lie in a compact ball around the true value of the parameters. This requirement is shown to be satisfied with high probability in the next subsection 3.2.

Decompose the optimization problem in (3) as follows:

$$\min_{\substack{B\in\mathbb{R}^{p_1\times p_2}\\ \Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}} f(B,\Theta_\epsilon)\equiv f_0(B,\Theta_\epsilon)+f_1(B)+f_2(\Theta_\epsilon),$$

where

$$f_0(B,\Theta_\epsilon)=\frac{1}{n}\sum_{j=1}^{p_2}\sum_{i=1}^{p_2}\sigma_\epsilon^{ij}(Y_i-XB_i)'(Y_j-XB_j)-\log\det\Theta_\epsilon=\mathrm{tr}(S\Theta_\epsilon)-\log\det\Theta_\epsilon,$$

$$f_1(B)=\lambda_n\|B\|_1,\quad f_2(\Theta_\epsilon)=\rho_n\|\Theta_\epsilon\|_{1,\mathrm{off}}.$$

and $\mathbb{S}_{++}^{p_2\times p_2}$ is the collection of $p_2\times p_2$ symmetric positive definite matrices. Further, denote the limit point (if there is any) of $\{\widehat{B}^{(k)}\}$ and $\{\widehat{\Theta}_\epsilon^{(k)}\}$ by $B^\infty=\lim_{k\to\infty}\widehat{B}^{(k)}$ and $\Theta_\epsilon^\infty=\lim_{k\to\infty}\widehat{\Theta}_\epsilon^{(k)}$, respectively.

Definition 1 (stationary point (Tseng, 2001) pp.479)**.** Define z to be a stationary point of f if $z\in\mathrm{dom}(f)$ and $f'(z;d)\geq 0,\forall$ direction $d=(d_1,\cdots,d_N)$ where d_t is the t^{th} coordinate block.

Definition 2 (Regularity (Tseng, 2001) pp.479)**.** f is regular at $z\in\mathrm{dom}(f)$ if $f'(z;d)\geq 0$ for all $d=(d_1,\cdots,d_N)$ such that

$$f'(z;(0,\cdots,d_t,\cdots,0))\geq 0,\qquad t=1,2,\cdots,N.$$

Definition 3 (Coordinate-wise minimum)**.** Define $(B^\infty,\Theta_\epsilon^\infty)$ to be a coordinate-wise minimum if

$$\begin{aligned} f(B^\infty,\Theta_\epsilon) &\geq f(B^\infty,\Theta_\epsilon^\infty),\quad \forall\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2},\\ f(B,\Theta_\epsilon^\infty) &\geq f(B^\infty,\Theta_\epsilon^\infty),\quad \forall B\in\mathbb{R}^{p_1\times p_2}.\end{aligned}$$

Note for our iterative algorithm, we only have two blocks, hence with the above notation, $N=2$.

Remark 3. Tseng (2001) proved that if the level set $\{x : f(x) \leq f(x^0)\}$ is compact and f satisfies certain conditions (Tseng, 2001, see Theorem 4.1 (a), (b) and (c) for details), the limit point given by the general block-coordinate descent algorithm (with $N \geq 2$ blocks) is a stationary point of f. However, the conditions given in Theorem 4.1 (a), (b) and (c) are not satisfied for the objective function at hand. Hence, for the problem under consideration, a different strategy is needed to prove convergence of the $2-$block alternating algorithm to a stationary point, and the resulting statements hold true for all problems that use a 2-block coordinate descent algorithm.

Since $\text{dom}(f_0)$ is open and f_0 is Gâteaux-differentiable on the $\text{dom}(f_0)$, by Tseng (2001) Lemma 3.1, f is regular in the $\text{dom}(f)$. From the discussion on Page 479 of (Tseng, 2001), we then have:

Fact 1: Every coordinate-wise minimum is a stationary point of f.

The following theorem shows that any limit point $(B^\infty, \Theta_\epsilon^\infty)$ of the iterative algorithm described in Section 2.2 is a stationary point of f, as long as all the iterates are within a closed ball around the truth.

Theorem 1 (Convergence for fixed design)**.** *Suppose for any fixed realization of X and E, the estimates $\left\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\right\}_{k=1}^{\infty}$ obtained by implemeting the alternating search step satisfy the following bound for some $R > 0$ that only depends on p_1, p_2 and n:*

$$\left\|(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}) - (B^*, \Theta^*)\right\|_F \leq R(p_1, p_2, n), \qquad \forall k \geq 1.$$

Then any limit point $(B^\infty, \Theta_\epsilon^\infty)$ of the iterative algorithm is a stationary point of f.

Remark 4. Recall that in classical parametric statistics, MLE-type asymptotics are derived after establishing that with probability tending to 1 as the sample size n goes to infinity, the likelihood equation has a sequence of roots (hence stationary points of the likelihood function) that converges in probability to the true value. Any such sequence of roots is shown to be asymptotically normal and efficient. Note that such (a sequence of) roots may not be global maximizers since parametric likelihoods are not globally log-concave (see Chapter 6 Lehmann and Casella, 1998). Here we show that the $(B^\infty, \Theta_\epsilon^\infty)$ obtained by the iterative algorithm is a stationary point which satisfies the first-order condition for being a maximizer of the penalized log-likelihood function (which is just the negative of the penalized least-squares function). Moreover, if we let n go to infinity, $(B^\infty, \Theta_\epsilon^\infty)$ converges to the true value in probability (shown in Theorem 4), and therefore behaves the same as the sequence of roots in the classical parametric problem alluded to above. Thus, while $(B^\infty, \Theta_\epsilon^\infty)$ may not be the global maximizer, it can, nevertheless, to all intents and purposes, be deemed as the MLE.

Remark 5. The above convergence result is based upon solving the optimization problem on the "entire" space, that is, we don't restrict B to live in any subspace. However, when actually implementing the proposed computational procedure, the optimization of the B coordinate is restricted to $\mathcal{B}_1 \times \cdots \times \mathcal{B}_{p_2}$ (as defined in eqn.4). It should be noted that the

same convergence property still holds, since for all $k \geq 1$, the following bound holds, for some $R' > 0$:

$$\left\| \left(\widehat{B}^{(k)}_{\text{restricted}}, \widehat{\Theta}^{(k)}_{\epsilon} \right) - (B^*, \Theta^*_{\epsilon}) \right\|_{\text{F}} \leq R'(p_1, p_2, n). \tag{11}$$

Consequently, the rest of the derivation in Theorem 1 follows, leading to the convergence property. The bound in eqn (11) will be shown at the end of Section 3.2.

3.2 Estimation Consistency.

In this subsection, we show that given a random realization of X and E, with high probability, the sequence $\left\{ (\widehat{B}^{(k)}, \widehat{\Theta}^{(k)}_{\epsilon}) \right\}_{k=1}^{\infty}$ lies in a non-expanding ball around $(B^*, \Theta^*_{\epsilon})$, thus satisfying the condition of Theorem 1 for convergence of the alternating algorithm.

It should be noted that for the alternating search procedure, we restrict our estimation on a subspace identified by the screening step. However, for the remaining of this subsection, the main propositions and theorems are based on the procedure without such restriction, i.e., we consider "generic" regressions on the entire space of dimension $p_1 \times p_2$. Notwithstanding, it can be easily shown that the theoretical results for the regression parameters on a restricted domain follow easily from the generic case, as explained in Remark 9.

Before providing the details of the main theorem statements and proofs, we first introduce additional notations. Let $\beta = \text{vec}(B)$ be the vectorized version of the regression coefficient matrix. Correspondingly, we have $\widehat{\beta}^{(k)} = \text{vec}(\widehat{B}^{(k)})$ and $\beta^* = \text{vec}(B^*)$. Moreover, we drop the superscripts and use $\widehat{\beta}$ and $\widehat{\Theta}_{\epsilon}$ to denote the generic estimators given by (12) and (13), as opposed to those obtained in any specific iteration:

$$\widehat{\beta} \equiv \underset{\beta \in \mathbb{R}^{p_1 p_2}}{\text{argmin}} \left\{ -2\beta' \widehat{\gamma} + \beta' \widehat{\Gamma} \beta + \lambda_n \|\beta\|_1 \right\}, \tag{12}$$

$$\widehat{\Theta}_{\epsilon} \equiv \underset{\Theta_{\epsilon} \in \mathbb{S}^{p_2 \times p_2}_{++}}{\text{argmin}} \left\{ -\log\det \Theta_{\epsilon} + \text{tr}\left(\widehat{S} \Theta_{\epsilon} \right) + \rho_n \|\Theta_{\epsilon}\|_{1,\text{off}} \right\}, \tag{13}$$

where

$$\widehat{\Gamma} = \left(\widehat{\Theta}_{\epsilon} \otimes \frac{X'X}{n} \right), \ \widehat{\gamma} = \left(\widehat{\Theta}_{\epsilon} \otimes X' \right) \text{vec}(Y)/n, \ \widehat{S} = \frac{1}{n} \left(Y - X\widehat{B} \right)' \left(Y - X\widehat{B} \right).$$

Remark 6. As opposed to (12) and (13), if $\widehat{\gamma}$ and $\widehat{\Gamma}$ are replaced by plugging in the true values of the parameters, the two problems in (12) and (13) become

$$\bar{\beta} \equiv \underset{\beta \in \mathbb{R}^{p_1 p_2}}{\text{argmin}} \left\{ -2\beta' \bar{\gamma} + \beta' \bar{\Gamma} \beta + \lambda_n \|\beta\|_1 \right\}, \tag{14}$$

$$\bar{\Theta}_{\epsilon} \equiv \underset{\Theta_{\epsilon} \in \mathbb{S}^{p_2 \times p_2}_{++}}{\text{argmin}} \left\{ -\log\det \Theta_{\epsilon} + \text{tr}\left(S \Theta_{\epsilon} \right) + \rho_n \|\Theta_{\epsilon}\|_{1,\text{off}} \right\}, \tag{15}$$

where

$$\bar{\Gamma} = \left(\Theta^*_{\epsilon} \otimes \frac{X'X}{n} \right), \ \bar{\gamma} = \left(\Theta^*_{\epsilon} \otimes X' \right) \text{vec}(Y)/n, \ \ S = \frac{1}{n} \left(Y - XB^* \right)' \left(Y - XB^* \right) \equiv \widehat{\Sigma}_{\epsilon}.$$

In (14), we obtain β using a penalized maximum likelihood regression estimate, and (15) corresponds to the generic setting for using the graphical Lasso. A key difference between

the estimation problems in (12) and (13) versus those in (14) and (15) is that to obtain $\widehat{\beta}$ and $\widehat{\Theta}_\epsilon$ we use *estimated quantities* rather than the raw data. This is exactly how we implement our iterative algorithm, namely, we obtain $\widehat{\beta}^{(k)}$ using $\widehat{S}^{(k-1)}$ as a surrogate for the sample covariance of the true error (which is unavailable), then estimate $\widehat{\Theta}_\epsilon^{(k)}$ using the information in $\widehat{\beta}^{(k)}$. This adds complication for establishing the consistency results. Original consistency results for the estimation problem in (14) and (15) are available in Basu and Michailidis (2015) and Ravikumar et al. (2011), respectively. Here we borrow ideas from corresponding theorems in those two papers, but need to tackle concentration bounds of relevant quantities with additional care. This part of the result and its proof are shown in Theorem 4.

As a road map toward our desired result established in Theorem 4, we first show in Theorem 2 that for any fixed realization of X and E, under a number of conditions on (or related to) X and E, when $\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\|_\infty$ is small (up to a certain order), the error of $\widehat{\beta}$ is well-bounded. We then verify in Proposition 1 and 2 that for random X and E, the above-mentioned conditions hold with high probability. Similarly in Theorem 3, we show that for fixed realizations in X and E, under certain conditions (verified for random X and E in Proposition 3), the error of $\widehat{\Theta}_\epsilon$ is also well-bounded, given $\|\widehat{\beta} - \beta^*\|_1$ being small. Finally in Theorem 4, we show that for random X and E, with high probability, the iterative algorithm gives $\{(\widehat{\beta}^{(k)}, \Theta_\epsilon^{(k)})\}$ that lies in a small ball centered at $(\beta^*, \Theta_\epsilon^*)$, whose radius depends on p_1, p_2, n and the sparsity levels.

Next, for establishing the main propositions and theorems, we introduce some additional notations.

- Sparsity level of β^*: $s^{**} := \|\beta^*\|_0 = \sum_{j=1}^{p_2} \|B_j^*\|_0 = \sum_{j=1}^{p_2} s_j^*$. As a reminder of the previous notation, we have $s^* = \max_{j=1,\cdots,p_2} s_j^*$.
- True edge set of Θ_ϵ^*: S_ϵ^*, and let $s_\epsilon^* := |S_\epsilon^*|$ be its cardinality.
- Hessian of the log-determinant barrier $\log\det\Theta$ evaluated at Θ_ϵ^*:

$$H^* := \frac{\mathrm{d}^2}{\mathrm{d}\Theta^2} \log\Theta\big|_{\Theta_\epsilon^*} = \Theta_\epsilon^{*-1} \otimes \Theta_\epsilon^{*-1}.$$

- Matrix infinity norm of the true error covariance matrix Σ_ϵ^*:

$$\kappa_{\Sigma_\epsilon^*} := |||\Sigma_\epsilon^*|||_\infty = \max_{i=1,2,\cdots,p_2} \sum_{j=1}^{p_2} |\Sigma_{\epsilon,ij}^*|.$$

- Matrix infinity norm of the Hessian restricted to the true edge set:

$$\kappa_{H^*} := \left|\left|\left|(H^*_{S_\epsilon^* S_\epsilon^*})\right|\right|\right|_\infty = \max_{i=1,2,\cdots,p_2} \sum_{j=1}^{p_2} \left|H^*_{S_\epsilon^* S_\epsilon^*, ij}\right|.$$

- Maximum degree of Θ_ϵ^*: $d := \max_{i=1,2,\cdots,p_2} \|\Theta^*_{\epsilon,i\cdot}\|_0$.
- We write $A \gtrsim B$ if there exists some absolute constant c that is independent of the model parameters such that $A \geq cB$.

Definition 4 (Incoherence condition (Ravikumar et al., 2011)). Θ^*_ϵ satisfies the incoherence condition if:

$$\max_{e\in(S^*_\epsilon)^c} \|H^*_{eS^*_\epsilon}(H^*_{S^*_\epsilon S^*_\epsilon})^{-1}\|_1 \leq 1-\xi, \quad \text{for some } \xi\in(0,1).$$

Definition 5 (Restricted eigenvalue (RE) condition (Loh and Wainwright, 2012)). A symmetric matrix $A\in\mathbb{R}^{m\times m}$ satisfies the RE condition with curvature $\varphi>0$ and tolerance $\phi>0$, denoted by $A\sim RE(\varphi,\phi)$ if

$$\theta' A\theta \geq \varphi\|\theta\|^2 - \phi\|\theta\|_1^2, \quad \forall\theta\in\mathbb{R}^m.$$

Definition 6 (Diagonal dominance). A matrix $A\in\mathbb{R}^{m\times m}$ is strictly diagonally dominant if

$$|a_{ii}| > \sum_{j\neq i}|a_{ij}|, \quad \forall i=1,\cdots,m.$$

Based on the model in Section 2.1, since we are assuming $\boldsymbol{X}=(X_1,\cdots,X_{p_1})'$ and $\boldsymbol{\epsilon}=(\epsilon_1,\cdots,\epsilon_{p_2})$ come from zero-mean Gaussian distributions, it follows that $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are zero-mean sub-Gaussian random vectors with parameters (Σ_X,σ_x^2) and $(\Sigma^*_\epsilon,\sigma^2_\epsilon)$, respectively. Moreover, throughout this section, all results are based on the assumption that Θ^*_ϵ is diagonally dominant.

Remark 7. Before moving on to the main statements of Theorem 2, we would like to point out that with a slight abuse of notation, for Theorem 2 and its related propositions and corollaries, the statements and analyses are based on equation (12) only, with *any deterministic symmetric matrix* $\widehat{\Theta}_\epsilon$ within a small ball around Θ^*_ϵ. Similarly in Theorem 3, Proposition 3 and Corollary 2, the analyses are based on equation (13) only, for *any given deterministic* $\widehat{\beta}$ within a small ball around β^*. The randomness of $\widehat{\beta}$ and $\widehat{\Theta}_\epsilon$ during the iterative procedure will be taken into consideration comprehensively in Theorem 4.

Theorem 2 (Error bound for $\widehat{\beta}$ with fixed realizations of X and E). *Consider $\widehat{\beta}$ given by (12). For any fixed pair of realizations of X and E , assume the following:*

A1. *$\widehat{\Theta}_\epsilon$ is a deterministic matrix satisfying the bound $\|\widehat{\Theta}_\epsilon-\Theta^*_\epsilon\|_\infty\leq\nu_\Theta$ where $\nu_\Theta=\eta_\Theta\left(\sqrt{\frac{\log p_2}{n}}\right)$ and η_Θ is some constant depending only on Θ^*_ϵ;*

A2. *$\widehat{\Gamma}\sim RE(\varphi,\phi)$, with $s^{**}\phi\leq\varphi/32$;*

A3. *$(\widehat{\Gamma},\widehat{\gamma})$ satisfies the deviation bound*

$$\|\widehat{\gamma}-\widehat{\Gamma}\beta^*\|_\infty \leq \mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}},$$

where $\mathbb{Q}(\nu_\Theta)$ is some quantity depending on ν_Θ.

Then, for any $\lambda_n\geq 4\mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}}$, the following bound holds:

$$\|\widehat{\beta}-\beta^*\|_1 \leq 64 s^{**}\lambda_n/\varphi.$$

The following two propositions verify the RE condition for $\widehat{\Gamma}$ and deviation bound for $(\widehat{\Gamma}, \widehat{\gamma})$ hold with high probability for a random pair (X, E), given any symmetric, matrix $\widehat{\Theta}_\epsilon$ satisfying (A1).

Proposition 1 (Verification of RE condition for random X and E)**.** *Consider any deterministic matrix $\widehat{\Theta}_\epsilon$ satisfying (A1). Let the sample size satisfy $n \gtrsim \max\{s^{**} \log p_1, d^2 \log p_2\}$. With probability at least $1 - 2\exp(-c_3 n)$ for some constant $c_3 > 0$, $\widehat{\Gamma}$ satisfies the following RE condition:*

$$\widehat{\Gamma} \equiv \widehat{\Theta}_\epsilon \otimes (X'X/n) \sim RE\left(\varphi^*(\min_i \psi^i - d\nu_\Theta), \phi^* \max_i(\psi^i + d\nu_\Theta)\right),$$

where $\varphi^* = \frac{\Lambda_{\min}(\Sigma_X^*)}{2}$, $\phi^* = (\varphi^* \log p_1)/n$, and ψ^i is defined as:

$$\psi^i := \sigma_\epsilon^{ii} - \sum_{j \neq i}^{p_2} \sigma_\epsilon^{ij},$$

where σ_ϵ^{ij}'s denote the entries in Θ_ϵ^* hence ψ^i is the gap between its diagonal entry and the sum of off-diagonal entries for row i.

Proposition 2 (Deviation bound for $(\widehat{\Gamma}, \widehat{\gamma})$ for random X and E)**.** *Consider any deterministic matrix $\widehat{\Theta}_\epsilon$ satisfying (A1). Let sample size n satisfy $n \gtrsim \log(p_1 p_2)$. With probability at least*

$$1 - 12c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)] \quad \text{for some } c_1 > 0, c_2 > 1,$$

the following bound holds:

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty = \frac{1}{n}\left\|X'E\widehat{\Theta}_\epsilon\right\|_\infty \leq \mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1 p_2)}{n}},$$

where

$$\mathbb{Q}(\nu_\Theta) = c_2\left\{d\nu_\Theta\left[\Lambda_{\max}(\Sigma_X^*)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2} + \left[\frac{\Lambda_{\max}(\Sigma_X^*)}{\Lambda_{\min}(\Sigma_\epsilon^*)}\right]^{1/2}\right\}. \tag{16}$$

Remark 8. In Proposition 1, the quantity $d^2 \log p_2$ that shows up in the sample size requirement is a result of $\nu_\Theta = O(\sqrt{\log p_2/n})$, which is the common order of error in a generic graphical Lasso problem. Hence here we explicitly list it for the purpose of showing results for the generic graphical Lasso estimation problem. In our iterative algorithm, the order of $\nu_\Theta^{(k)}$ depends on the relative order of p_1 and p_2, which may potentially make the sample size requirement more stringent. This will be discussed in more detail in the proof of Theorem 4.

Given the results in Theorem 2, Proposition 1 and Proposition 2, next we provide Corollary 1, which gives the error bound for $\widehat{\beta}$ for random realizations of X and E.

Corollary 1 (Error Bound for $\widehat{\beta}$ for random X and E)**.** *Consider any determinisitic $\widehat{\Theta}_\epsilon$ satisfying the following element-wise ℓ_∞-bound:*

$$\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\|_\infty \leq \nu_\Theta,$$

with $\nu_\Theta = \eta_\Theta\sqrt{\frac{\log p_2}{n}}$. *Then for sample size* $n \gtrsim \log(p_1p_2)$ *and for any regularization parameter* $\lambda_n \geq 4\mathbb{Q}(\nu_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}}$ *with the expression of* $\mathbb{Q}(\cdot)$ *given in (16), there exists* $c_1 > 0$ *and* $c_2 > 1$ *such that with probability at least:*

$$1 - 12c_1\exp[-(c_2^2-1)\log(p_1p_2)] - 2\exp(-c_3 n),$$

the following bound holds:

$$\|\widehat{\beta} - \beta^*\|_1 \leq 64 s^{**}\lambda_n/\varphi, \tag{17}$$

where $\varphi = \frac{1}{2}\Lambda_{\min}(\Sigma_\epsilon^*)(\min_i \psi^i - d\nu_\Theta)$.

Next, we move onto analyzing the error bound of the other component, for a fixed given $\widehat{\beta}$.

Theorem 3 (Error bound for $\widehat{\Theta}_\epsilon$ for fixed realizations of X and E)**.** *Consider* $\widehat{\Theta}_\epsilon$ *given by (13). For any fixed pair of realization* (X, E)*, assume the following:*

B1. $\widehat{\beta}$ *is a deterministic vector satisfying* $\|\widehat{\beta}-\beta^*\|_1 \leq \nu_\beta$*, where* $\nu_\beta = \eta_\beta\left(\sqrt{\frac{\log(p_1p_2)}{n}}\right)$*, with*$\eta_\beta$ *being some constant depending only on* β^**;*

B2. $\|\widehat{S} - \Sigma_\epsilon^*\|_\infty \leq g(\nu_\beta)$ *where*

$$\widehat{S} = \frac{1}{n}(Y - X\widehat{B})'(Y - X\widehat{B}),$$

and $g(\nu_\beta)$ *is some quantity depending on* ν_β*;*

B3. *Incoherence condition holds for* Θ_ϵ^*.

Then, for $\rho_n = (8/\xi)g(\nu_\beta)$ *and sample size* n *satisfying* $n \gtrsim \log(p_1p_2)$*, the following error bound for* $\widehat{\Theta}_\epsilon$ *holds:*

$$\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\|_\infty \leq \{2(1 + 8\xi^{-1})\kappa_{H^*}\}g(\nu_\beta), \tag{18}$$

where ξ *is the incoherence parameter as defined in Definition 4.*

Proposition 3 gives an explicit expression for $g(\nu_\beta)$ under condition (B1). Specifically, it shows how well $\widehat{S}$ concentrates around Σ_ϵ^* for random X and E, given some $\widehat{B}$ exhibiting a small error from its true value (or $\widehat{\beta}$, equivalently),

Proposition 3. *Consider any determinisitc* $\widehat{\beta}$ *satisfying (B1). Then for sample size* n *satisfying* $n \gtrsim \log(p_1p_2)$*, with probability at least*

$$1 - 1/p_1^{\tau_1-2} - 1/p_2^{\tau_2-2} - 6c_1\exp[-(c_2^2-1)\log(p_1p_2)], \quad \text{for some } c_1 > 0, c_2 > 1, \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{S} - \Sigma_\epsilon^*\|_\infty \leq g(\nu_\beta), \tag{19}$$

where

$$g(\nu_\beta) = \sqrt{\frac{\log 4 + \tau_2 \log p_2}{c_\epsilon^* n}} + \nu_\beta^2 \left(\sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}} + \max_i(\Sigma^*_{X,ii}) \right) + 2c_2\nu_\beta \left[\Lambda_{\max}(\Sigma_X^*)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2} \sqrt{\frac{\log(p_1p_2)}{n}}, \tag{20}$$

c_ϵ^* and c_X^* are population quantities given in (57) and (62), respectively.

Given Theorem 3 and Proposition 3, we provide Corollary 2, which gives the error bound for $\widehat{\Theta}_\epsilon$ for random realizations of X and E:

Corollary 2 (Error bound for $\widehat{\Theta}$ for random X and E)**.** *Consider any deterministic $\widehat{\beta}$ satisfying the following bound*

$$\|\widehat{\beta} - \beta^*\|_1 \le \nu_\beta,$$

with $\nu_\beta = \eta_\beta \sqrt{\frac{\log(p_1p_2)}{n}}$. Also suppose the incoherence condition (B3) is satisfied. Then, for sample size $n \gtrsim \log(p_1p_2)$ and regularization parameter $\rho_n = (8/\xi)g(\nu_\beta)$ with $g(\nu_\beta)$ given in (20), with probability at least

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 6c_1 \exp[-(c_2^2 - 1)\log(p_1p_2)], \quad \text{for some } c_1 > 0, c_2 > 1, \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\|_\infty \le \{2(1 + 8\xi^{-1})\kappa_{H^*}\} g(\nu_\beta).$$

After providing the error bound for (12) and (13), in Theorem 4 we establish that with high probability, the error of the sequence of estimates obtained in the alternating search step of the algorithm described in Section 2.2 is *uniformly* bounded; that is, the sequence of estimates lie in a non-expanding ball around the true value of the parameters uniformly with a radius that does not depend on the iteration number k.

Theorem 4 (Error bound for $\{\widehat{\beta}^{(k)}\}$ and $\{\widehat{\Theta}_\epsilon^{(k)}\}$)**.** *Consider the iterative algorithm given in Section 2.2 that gives rise to sequences of $\{\widehat{\beta}^{(k)}\}$ and $\{\widehat{\Theta}_\epsilon^{(k)}\}$ alternately. For random realization of X and E, we assume the following:*

- ***C1.*** *The incoherence condition holds for Θ_ϵ^*.*
- ***C2.*** *Θ_ϵ^* is diagonally dominant.*
- ***C3.*** *The maximum sparsity level for all p_2 regression s^* satisfies $s^* = o(n/\log p_1)$.*

(I) For sample size satisfying $n \gtrsim \log(p_1p_2)$, there exist constants $c_1 > 0, c_2 > 1, c_3 > 0$ such that for any

$$\lambda_n^0 \ge 4c_2 \left[\Lambda_{\max}(\Sigma_X^*)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2} \sqrt{\frac{\log(p_1p_2)}{n}},$$

with probability at least $1 - 2\exp(-c_3 n) - 6c_1\exp[-(c_2^2 - 1)\log(p_1p_2)]$, the initial estimate $\widehat{\beta}^{(0)} \equiv vec(\widehat{B}^{(0)})$ satisfies the following bound

$$\|\widehat{\beta}^{(0)} - \beta^*\|_1 \le 64 s^{**} \lambda_n^0/\varphi^* \equiv \nu_\beta^{(0)}, \tag{21}$$

where $\varphi^* = \Lambda_{\min}(\Sigma_X^*)/2$. *Moreover, by choosing* $\rho_n^0 = (\frac{8}{\xi})g(\nu_\beta^{(0)})$ *where the expression for* $g(\cdot)$ *is given in (20), with probability at least*

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 2\exp(-c_3 n) - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)], \quad \text{for some } \tau_1, \tau_2 > 2,$$

the following bound holds:

$$\|\widehat{\Theta}_\epsilon^{(0)} - \Theta_\epsilon^*\|_\infty \leq \{2(1 + 8\xi^{-1})\kappa_{H^*}\} g(\nu_\beta^{(0)}) \equiv \nu_\Theta^{(0)}. \tag{22}$$

(II) For sample size satisfying $n \gtrsim d^2 \log(p_1 p_2)$, *for any iteration* $k \geq 1$, *with probability at least*

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 12c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)] - 2\exp[-c_3 n],$$

the following bounds hold for all $\widehat{\beta}^{(k)}$ *and* $\widehat{\Theta}_\epsilon^{(k)}$*:*

$$\begin{aligned} \|\widehat{\beta}^{(k)} - \beta^*\|_1 &\leq C_\beta \left(s^{**} \sqrt{\frac{\log(p_1 p_2)}{n}} \right), \\ \|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^*\|_\infty &\leq C_\Theta \left(\sqrt{\frac{\log(p_1 p_2)}{n}} \right). \end{aligned}$$

where s^{**} *is the sparsity of* β^*, C_β *and* C_Θ *are constants depending only on* β^* *and* Θ_ϵ^*, *respectively.*

As a direct result of Proposition 1 in Basu and Michailidis (2015) and Corollary 3 in Ravikumar et al. (2011), the following bound also holds:

Corollary 3. *Under the same set of conditions C1, C2 and C3 in Theorem 4, there exists* $\tau_1, \tau_2 > 2$, $c_1 > 0, c_2 > 1, c_3 > 0$ *and constants* C'_β *and* C'_Θ *such that for all iterations* k, *the following bound holds:*

$$\begin{aligned} \|\widehat{\beta}^{(k)} - \beta^*\|_F &\leq C'_\beta \left(\sqrt{\frac{s^{**} \log(p_1 p_2)}{n}} \right), \\ \|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^*\|_F &\leq C'_\Theta \sqrt{\frac{(s_\epsilon^* + p_2)\log(p_1 p_2)}{n}}, \end{aligned}$$

with probability at least

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 12c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)] - 2\exp[-c_3 n],$$

where s^{**} *and* s_ϵ^* *are the sparsity for* β^* *and* Θ_ϵ^*, *respectively.*

Remark 9. As mentioned earlier in this subsection, the actual implementation of the alternating search step is restricted to a subspace of $\mathbb{R}^{p_1 \times p_2}$. Next, we outline the corresponding theoretical results for this specific scenario in which for each regression j, some *fixed superset* of the indices of true covariates is given, and the regressions are restricted to these supersets, respectively. Note that we need to make sure that the restricted subspace contains all the true covariates for the results below to be valid.

Let S_j denote the given *fixed superset* for each regression j, and we consider regressing the response on X_{S_j}. We use $\widehat{\beta}_{\mathrm{R}}^{(k)}$ to denote the corresponding vectorized estimator of iteration k, that is,

$$\widehat{\beta}_{\mathrm{R}}^{(k)} = (\widehat{B}_{1,\mathrm{Restricted}}^{(k)\prime}, \cdots, \widehat{B}_{p_2,\mathrm{Restricted}}^{(k)\prime})',$$

where $\widehat{B}_{j,\mathrm{Restricted}}^{(k)\prime}$ is obtained by doing the regression in (7), however with the indices of covariates restricted to S_j. Also, we let β_{R}^* be the corresponding true value of $\widehat{\beta}_{\mathrm{R}}^{(k)}$. Note that always holds that

$$\|\widehat{\beta}_{\mathrm{R}}^{(k)} - \beta_{\mathrm{R}}^*\| = \|\widehat{\beta}^{(k)} - \beta^*\|.$$

Now let

$$\bar{S} = \bigcup_{j\in\{1,\cdots,p_2\}} S_j,$$

and let $\bar{s}$ be its cardinality. It can be shown that the best achievable error bound for $\widehat{\beta}_{\mathrm{R}}^{(k)}$ is identical to $\widehat{\beta}_{\bar{S}}^{(k)}$, where $\widehat{\beta}_{\bar{S}}^{(k)}$ is obtained by considering covariates $X_{\bar{S}}$ for all p_2 regressions, instead of the entire X. For this specific reason, formally, we state the theoretical results for the case where we consider regressing on $X_{\bar{S}}$, which is almost identical to the generic case.

Suppose conditions C1, C2 and C3 in Theorem 4 hold, then there exists constants $c_1 > 0, c_2 > 1, c_3 > 0, \tau_1 > 2, \tau_2 > 2$ such that: (I) for sample size satisfying $n \gtrsim \log(\bar{s}p_2)$, w.p. at least $1 - 2\exp(-c_3 n) - 6c_1\exp[-(c_2^2-1)\log(\bar{s}p_2)]$, for any

$$\lambda_n^0 \geq 4c_2\left[\Lambda_{\max}(\Sigma^*_{X_{\bar{S}}})\Lambda_{\max}(\Sigma^*_\epsilon)\right]^{1/2}\sqrt{\frac{\log(\bar{s}p_2)}{n}},$$

the initial estimate $\widehat{\beta}_{\bar{S}}^{(0)}$ satisfies the following bound:

$$\|\widehat{\beta}_{\bar{S}}^{(0)} - \beta_{\bar{S}}^*\|_1 \leq 64 s^{**}\lambda_n^0/\varphi_{\bar{S}}^* \equiv \nu_{\beta_{\bar{S}}}^{(0)},$$

where $\varphi_{\bar{S}}^* = \Lambda_{\min}(\Sigma^*_{X_{\bar{S}}})/2$. Moreover, by choosing $\rho_n^0 = (\frac{8}{\xi})g(\nu_{\beta_{\bar{S}}}^{(0)})$ where the expression for $g(\cdot)$ is given in (20), with probability at least

$$1 - 1/\bar{s}^{\tau_1-2} - 1/p_2^{\tau_2-2} - 2\exp(-c_3 n) - 6c_1\exp[-(c_2^2-1)\log(\bar{s}p_2)],$$

the following bound holds:

$$\|\widehat{\Theta}_\epsilon^{(0)} - \Theta_\epsilon^*\|_\infty \leq \{2(1+8\xi^{-1})\kappa_{H^*}\}g(\nu_{\beta_{\bar{S}}}^{(0)}) \equiv \nu_\Theta^{(0)}.$$

(II) For sample size satisfying $n \gtrsim d^2\log(\bar{s}p_2)$, for any iteration $k \geq 1$, with probability at least

$$1 - 1/\bar{s}^{\tau_1-2} - 1/p_2^{\tau_2-2} - 12c_1\exp[-(c_2^2>1)\log(\bar{s}p_2)] - 2\exp[-c_3 n],$$

the following bound hold for all $\widehat{\beta}_{\bar{S}}^{(k)}$ and $\widehat{\Theta}_\epsilon^{(k)}$:

$$\|\widehat{\beta}_{\bar{S}}^{(k)} - \beta^*\|_1 \leq C_\beta\left(s^{**}\sqrt{\frac{\log(\bar{s}p_2)}{n}}\right), \quad \|\widehat{\beta}_{\bar{S}}^{(k)} - \beta^*\|_F \leq C'_\beta\left(\sqrt{\frac{s^{**}\log(\bar{s}p_2)}{n}}\right),$$

$$\|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^*\|_\infty \leq C_\Theta\left(\sqrt{\frac{\log(\bar{s}p_2)}{n}}\right), \quad \|\widehat{\Theta}_\epsilon^{(k)} - \Theta_\epsilon^*\|_F \leq C'_\Theta\sqrt{\frac{(s_\epsilon^* + p_2)\log(\bar{s}p_2)}{n}},$$

where s^{**} is the sparsity of β^*, C_β, C'_β, C_Θ and C'_θ are all constants that do not depend on $n, \bar{S}, p_2$.

3.3 Family-wise Error Rate Control of the Screening Step.

As mentioned in the Introduction, for the iterative algorithm to work effectively, it is crucial to initialize from points that are close to the true parameters. Our screening step provides such guarantees *asymptotically*. Based on the screening step described in Section 2.2, initial estimates for each column of the regression matrix are obtained by Lasso or Ridge regression with the support set restricted to the one identified by the screening step. It is desirable for the screening step to correctly identify the true support set. In particular, we would like to retain as many true positive predictor variables as possible without discovering too many false positive ones. The following theorem states that as long as $\log(p_1 p_2)/n = o(1)$ and the sparsity is not beyond a specified level, the screening step will be able to recover all true positive predictors, while keeping the family-wise type I error under control.

Theorem 5. *Let S_j^* denote the true support set of the jth regression and s_j^* be its cardinality. Suppose that $\log(p_1 p_2)/n \to 0$ and the following condition for sparsity holds:*

$$\max\{s_j^*, j = 1, \cdots, p_2\} = o(\sqrt{n}/\log p_1).$$

Then, the screening step described in Section 2.2 will correctly recover S_j^* for all $j = 1, \cdots, p_2$ with probability approaching to 1, while keeping the family-wise type I error rate under the prespecified level α.

Remark 10. The specified level for sparsity is necessary for the de-biased Lasso procedure in Javanmard and Montanari (2014) to produce unbiased estimates for the regression coefficients. In terms of support recovery for the screening step, with $\log(p_1 p_2)/n = o(1)$, we only require $s^* = o(p_1)$, which is much weaker and easily satisfied.

The following corollary connects the screening step with the alternating search step, under the discussed asymptotic regime :

Corollary 4. *Consider the model set-up given in Section 2.1. Let s^* denote the maximum sparsity for all $B_j^*, j = 2, \cdots, p_2$, and d denote the maximum degree of Θ_ϵ^*. Also, let s^{**} denote the sparsity for β^* and s_ϵ^* denote the sparsity for Θ_ϵ^*. Assume there exist positive constants $c_{s^*}, c_{s^{**}}, c_d, c_{\bar{s}}, c_{p_2}$ satisfying*

$$0 < c_{s^*} + c_{\bar{s}} < 1/2; \quad 0 < c_{s^{**}} + c_{\bar{s}} < 1; \quad 0 < 2c_d + c_{\bar{s}} < 1; \quad 0 < \max\{c_{s_\epsilon^*}, c_{p_2}\} + c_{\bar{s}} < 1$$

such that

$$s^* = O(n^{c_s}); \quad s^{**} = O(n^{c_{s^{**}}}); \quad s_\epsilon^* = O(n^{c_{s_\epsilon^*}}); \quad d = O(n^{c_d}); \quad \bar{s} = O(e^{n^{c_{p_1}}}); \quad p_2 = O(n^{c_{p_2}}).$$

As $n \to \infty$,

$\mathbb{P}(\{$*The screening step correctly recovers the true support set for all $B_j, j = 1, \cdots, p$*$\}) \to 1$,

and for all iterations k:

$$\max_{k \geq 1} \left\| (\widehat{\beta}_{\mathrm{R}}, \widehat{\Theta}_\epsilon^{(k)}) - (\beta_{\mathrm{R}}^*, \Theta_\epsilon^*) \right\| \xrightarrow{p} 0.$$

The proof of this corollary follows along the same lines as Theorem 4, and we leave the details to the reader.

3.4 Estimation Error and Identifiability.

In this subsection, we discuss in detail the conditions needed for the parameters in our multi-layered network to be identifiable (estimable). We focus the presentation for ease of exposition on a three-layer network and then discuss the general M-layer case.

Consider a 3-layer graphical model. Let $\widetilde{X} = [(X^1)', (X^2)']'$ be the (p_1+p_2) dimensional random variable, which represents the "super"-layer on which we regress X^3 to estimate B^{13}, B^{23} and Σ^3. As shown in Theorem 2, the estimation error for $\widehat{\beta}$ takes the following form:

$$\|\widehat{\beta} - \beta^*\|_1 \leq 64 s^{**} \lambda_n / \varphi,$$

where φ is the curvature parameter for RE condition that scales with $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ (see Proposition 1). Therefore, the error of estimating these regression parameters is higher when $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ is smaller. In this section, we derive a lower bound on this quantity to demonstrate how the estimation error depends on the underlying structure of the graph.

For the undirected subgraph within a layer k, we denote its maximum node capacity by $\mathbf{v}(\Theta^k) := \max_{1\leq i\leq p_k} \sum_{j=1}^{p_k} |\Theta_{ij}|$. For the directed bipartite subgraph consisting of Layer $s \to t$ edges ($s < t$), we similarly define the maximum incoming and outgoing node capacities by $\mathbf{v}_{in}(B^{st}) := \max_{1\leq j\leq p_t} \sum_{i=1}^{p_s} |B^{st}_{ij}|$ and $\mathbf{v}_{out}(B^{st}) := \max_{1\leq i\leq p_s} \sum_{j=1}^{p_t} |B^{st}_{ij}|$. The following proposition establishes the lower bound in terms of these node capacities

Proposition 4.

$$\Lambda_{\min}(\Sigma_{\widetilde{X}}) \geq \mathbf{v}(\Theta^1)^{-1}\mathbf{v}(\Theta^2)^{-1} \left[1 + \left(\mathbf{v}_{in}(B^{12}) + \mathbf{v}_{out}(B^{12})\right)/2\right]^{-2}$$

The three components in the lower bound demonstrate how the structure of Layers 1 and 2 impact the accurate estimation of directed edges to Layer 3. Essentially, the bound suggests that accurate estimation is possible when the total effect (incoming and outgoing edges) at every node of each of the three subgraphs is not very large.

This is inherently related to the identifiability of the multi-layered graphical models and our ability to distinguish between the parents from different layers. For instance, if a node in Layer 2 has high partial correlation with nodes of Layer 1, i.e., a node in Layer 2 has parents from many nodes in Layer 1 and yields a large $\mathbf{v}_{in}(B^{12})$; or similarly, a node in Layer 1 is the parent of many nodes in Layer 2, yielding a large $\mathbf{v}_{out}(B^{12})$. In either case, we end up with some large lower bound for $\Lambda_{\min}(\Sigma_{\widetilde{X}})$ and it can be hard to distinguish Layer $1 \to 3$ edges from Layer $2 \to 3$ edges.

For a general M-layer network, the argument in the proof of Proposition 4 (see Section B for details) can be generalized in a straightforward manner. In the 2-layer network setting, with the notation defined in Section 2, by setting $\epsilon^1 = X^1$, we have

$$\begin{bmatrix} \epsilon^1 \\ \epsilon^2 \end{bmatrix} = P \begin{bmatrix} X^1 \\ X^2 \end{bmatrix}, \quad \text{where} \quad P = \begin{bmatrix} I & 0 \\ -(B^{12})' & I \end{bmatrix}.$$

For an M-layer network, a modified P is given in the following form:

$$P = \begin{bmatrix} I & 0 & \dots & 0 \\ -(B^{12})' & I & \dots & 0 \\ \vdots & \vdots & \vdots & 0 \\ -(B^{1,M-1})' & -(B^{2,M-1})' & \dots & I \end{bmatrix}$$

and combines node capacities for different layers. The conclusion is qualitatively similar, i.e., the estimation error of an M-layer graphical model is small as long as the maximum node capacities of different inter-layer and intra-layer subgraphs are not too large.

4. Performance Evaluation and Implementation Issues.

In this section, we present selected simulation results for our proposed method, in two-layer and three-layer network settings. Further, we introduce some acceleration techniques that can speed up the algorithm and reduce computing time.

4.1 Simulation Results.

For the 2-layer network, as mentioned in Section 2.1, since the main target of our proposed algorithm is to provide estimates for B^* and Θ^*_ϵ (since Θ_X can be estimated separately), we only present evaluation results for B^* and Θ^*_ϵ estimates. Similarly, for the three-layer network, we only present evaluation results involving Layer 3, using the notation in Section 3.4, that is, B^*_{XZ}, B^*_{YZ} and $\Theta^*_{\epsilon,Z}$ estimates, which is sufficient to show how our proposed algorithm works in the presence of a "super" - layer, taking advantage of the separability of the log-likelihood.

2-layered Network. To compare the proposed method with the most recent methodology that also provides estimates for the regression parameters and the precision matrix (CAPME, Cai et al. (2012)), we use the exact same model settings that have been used in that paper. Specifically, we consider the following two models:

- Model A: Each entry in B^* is nonzero with probability $5/p_1$, and off-diagonal entries for Θ^*_ϵ are nonzero with probability $5/p_2$.
- Model B: Each entry in B^* is nonzero with probability $30/p_1$, and off-diagonal entries for Θ^*_ϵ are nonzero with probability $5/p_2$.

As in Cai et al. (2012), for both models, nonzero entries of B^* and Θ^*_ϵ are generated from $\mathsf{Unif}\left[(-1,-0.5)\cup(0.5,1)\right]$, and diagonals of Θ^*_ϵ are set identical such that the condition number of Θ^*_ϵ is p_2.

	(p_1, p_2, n)
Model A	$p_1 = 30, p_2 = 60, n = 100$
	$p_1 = 60, p_2 = 30, n = 100$
	$p_1 = 200, p_2 = 200, n = 150$
	$p_1 = 300, p_2 = 300, n = 150$
Model B	$p_1 = 200, p_2 = 200, n = 100$
	$p_1 = 200, p_2 = 200, n = 200$

Table 1: Model Dimensions for Model A and B

To evaluate the selection performance of the algorithm, we use sensitivity (SEN), specificity (SPE) and Mathews Correlation Coefficient (MCC) as criteria:

$$\text{SEN} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \quad \text{SPE} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}.$$

Further, to evaluate the accuracy of the magnitude of the estimates, we use the relative error in Frobenius norm:

$$\text{rel-Fnorm} = \frac{\|\widetilde{B} - B^*\|_{\mathrm{F}}}{\|B^*\|_{\mathrm{F}}} \quad \text{or} \quad \frac{\|\widetilde{\Theta}_\epsilon - \Theta^*_\epsilon\|_{\mathrm{F}}}{\|\Theta^*_\epsilon\|_{\mathrm{F}}}.$$

Tables 2 and 3 show the results for both the regression matrix and the precision matrix. For the precision matrix estimation, we compare our result with those available in Cai et al. (2012), denoted as CAPME.

	(p_1, p_2, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(30,60,100)	0.96(0.018)	0.99(0.004)	0.93(0.014)	0.22(0.029)
	(60,30,100)	0.99(0.009)	0.99(0.003)	0.93(0.017)	0.18(0.021)
	(200,200,150)	0.99(0.001)	0.99(0.001)	0.88(0.009)	0.18(.007)
	(300,300,150)	1.00(0.001)	0.99(0.001)	0.84(0.010)	0.21(0.007)
Model B	(200,200,200)	0.970(0.004)	0.982(0.001)	0.927(0.002)	0.194 (0.009)
	(200,200,100)	0.32(0.010)	0.99(0.001)	0.49(0.009)	0.85(0.006)

Table 2: Performance evalution for the estimated regression matrix, over 50 replications

	(p_1, p_2, n)		SEN	SPE	MCC	rel-Fnorm
Model A	(30,60,100)		0.77(0.031)	0.92(0.007)	0.56(0.030)	0.51(0.017)
		CAPME	0.58(0.03)	0.89(0.01)	0.45(0.03)	
	(60,30,100)		0.76(0.041)	0.89(0.015)	0.59(0.039)	0.49(0.014)
	(200,200,150)		0.78(0.019)	0.97(0.001)	0.55(0.012)	0.60(0.007)
	(300,300,150)		0.71(0.017)	0.98(0.001)	0.51(0.011)	0.59(0.005)
Model B	(200,200,200)		0.73(0.023)	0.94(0.003)	0.39(0.017)	0.62(0.011)
		CAPME	0.36(0.02)	0.97(0.00)	0.35(0.01)	
	(200,200,100)		0.57(0.027)	0.44(0.007)	0.04(0.008)	0.84(0.002)
		CAPME	0.19(0.01)	0.87(0.00)	0.04(0.01)	

Table 3: Performance evaluation for estimated precision matrix, over 50 replications

As it can be seen from Tables 2 and 3, the sample size is a key factor that affects the performance. Our proposed algorithm performs extremely well in its selection properties on B and strikes a good balance between sensitivity and specificity in estimating Θ_ϵ. [3] For most settings, it provides substantial improvements over the CAPME estimator.

3-layer Network. For a 3-layer network, we consider the following data generation mechanism: for all three models A, B and C, each entry in B_{XY} is nonzero with probability

3. In practice, for the debias Lasso procedure, we use the default choice of tuning parameters suggested in the implementation of the code provided in Javanmard and Montanari (2014); for FWER, we suggest using $\alpha = 0.1$ as the thresholding level; for tuning parameter selection, we suggest doing a grid search for (λ_n, ρ_n) on $[0, 0.5\sqrt{\log p_1/n}] \times [0, 0.5\sqrt{\log p_2/n}]$ with BIC.

$5/p_1$, each entry in B_{XZ} and B_{YZ} is nonzero with probability $5/(p_1+p_2)$, and off-diagonal entries in $\Theta_{\epsilon,Z}$ are nonzero with probability $5/p_3$. Similar to the 2-layered set-up, the nonzero entries in $\Theta_{\epsilon,Z}$ are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$ with its diagnals set identical such that its condition number is p_3. For the regression matrices in the three models, nonzeros in B_{XY} are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$, and nonzeros in B_{XZ} and B_{YZ} are generated from $\{\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]*\text{Signal.Strength}\}$, where the signal strength in the three models are given by 1, 1.5 and 2, respectively. More specifically, for Model A, B and C, nonzeros in B_{XZ} or B_{YZ} are generated from $\mathsf{Unif}[(-1,-0.5)\cup(0.5,1)]$, $\mathsf{Unif}[(-1.5,-0.75)\cup(0.75,1.5)]$ and $\mathsf{Unif}[(-2,-1)\cup(1,2)]$, respectively.

	Layer 3 Signal.Strength	(p_1,p_2,p_3,n)
Model A	1	(50,50,50,200)
Model B	1.5	(50,50,50,200)
Model C	2	(50,50,50,200)
		(20,80,50,200)
		(80,20,50,200)
		(100,100,100,200)

Table 4: Model Dimensions and Signal Strength for Model A, B and C

As mentioned in the beginning of this subsection, we only evaluate the algorithm's performance on B_{XZ}, B_{YZ} and $\Theta_{\epsilon,Z}$.

Based on the results shown in Tables 5, 6 and 7, the signal strength across layers affects the accuracy of the estimation, which is in accordance with what has been discussed regarding identifiability. Overall, the MLE estimator performs satisfactorily across a fairly wide range of settings and in many cases achieving very high values for the MCC criterion.

4.1.1 Simulation Results for non-Gaussian data

In many applications, the data may not be exactly Gaussian, but approximately Gaussian. Next, we present selected simulation results when the data comes from some distribution that deviates from Gaussian. Specifically, we consider two types of deviations based on the following transformations: (i) a truncated empirical cumulative distribution function and (ii) a shrunken empirical cumulative distribution functions as discussed in Zhao et al. (2015). In both simulation settings, we consider Model A with $(p_1,p_2,n)=(30,60,100)$ under the two-layer setting, and the transformation is applied to errors in Layer 2. Table 8 shows the simulation results for these two scenarios over 50 replications.

Based on the results in Table 8, relatively small deviations from the Gaussian distribution do not affect the performance of the MLE estimates under the examined settings that are comparable to those obtained with Gaussian distributed data.

4.2 A comparison with the two-step estimator in Cai et al. (2012)

Next, we present a comparison between the MLE estimator and the two-step estimator of Cai et al. (2012). Specifically, we use the CAPME estimate as an initializer for the

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.51(0.065)	0.99(0.001)	0.69(0.049)	0.68(0.050)
Model B	(50,50,50,200)	0.85(0.043)	0.99(0.001)	0.898(0.025)	0.36(0.056)
Model C	(50,50,50,200)	0.97(0.018)	0.99(0.002)	0.96(0.016)	0.16(0.040)
	(20,80,50,200)	0.55(0.078)	0.99(0.001)	0.72(0.059)	0.63(0.066)
	(80,20,50,200)	0.99(0.006)	0.99(0.002)	0.94(0.017)	0.076(0.032)
	(100,100,100,200)	1.00(0.001)	0.99(0.001)	0.87(0.016)	0.07(0.007)

Table 5: Performance evaluation for estimated regression matrix B_{XZ} over 50 replications

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.53(0.051)	1.00(0.000)	0.72(0.036)	0.65(0.041)
Model B	(50,50,50,200)	0.90(0.033)	1.00(0.000)	0.95(0.019)	0.25(0.049)
Model C	(50,50,50,200)	0.98(0.013)	1.00(0.000)	0.99(0.007)	0.12(0.042)
	(20,80,50,200)	0.95(0.013)	1.00(0.000)	0.98(0.007)	0.19(0.030)
	(80,20,50,200)	0.96(0.027)	0.99(0.001)	0.97(0.022)	0.14(0.063)
	(100,100,100,200)	1.00(0.000)	1.00(0.000)	0.99(0.002)	0.025(0.002)

Table 6: Performance evaluation for estimated regression matrix B_{YZ} over 50 replications

	(p_1, p_2, p_3, n)	SEN	SPE	MCC	rel-Fnorm
Model A	(50,50,50,200)	0.69(0.044)	0.638(0.032)	0.20(0.036)	0.82(0.017)
Model B	(50,50,50,200)	0.77(0.050)	0.82(0.036)	0.42(0.071)	0.68(0.040)
Model C	(50,50,50,200)	0.88(0.041)	0.91(0.019)	0.63(0.059)	0.56(0.034)
	(20,80,50,200)	0.72(0.041)	0.80(0.028)	0.36(0.050)	0.72(0.021)
	(80,20,50,200)	0.90(0.028)	0.92(0.011)	0.68(0.039)	0.58(0.018)
	(100,100,100,200)	0.96(0.014)	0.96(0.003)	0.68(0.016)	0.049(0.010)

Table 7: Performance evaluation for estimated precision matrix $\Theta_{\epsilon,Z}$ over 50 replications

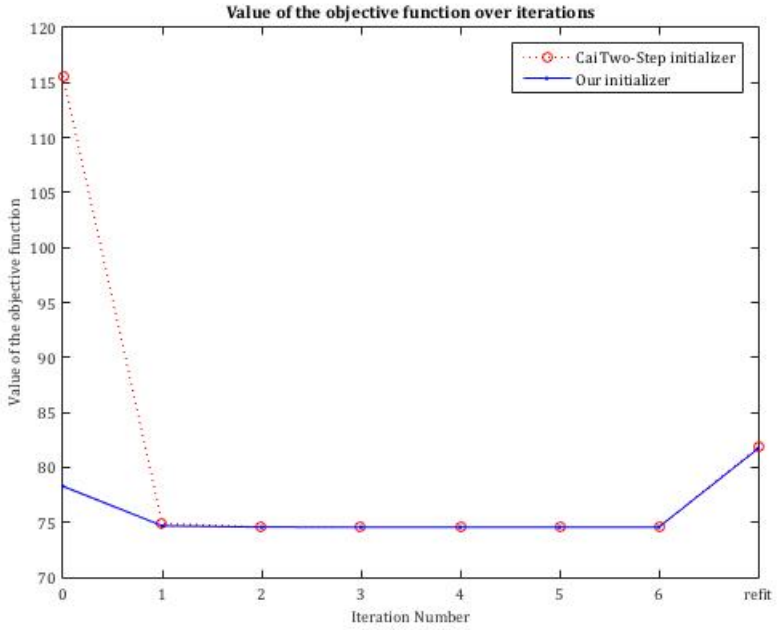
Setting	Parameter	SEN	SPE	MCC	rel-Fnorm
Model A $(30, 60, 100)$	B	0.96(0.017)	0.99(0.003)	0.94(0.012)	0.20(0.028)
shrunken	Θ_ϵ	0.76(0.031)	0.91(0.008)	0.55(0.030)	0.51(0.019)
Model A $(30, 60, 100)$	B	0.96(0.021)	0.98(0.004)	0.93(0.015)	0.21(0.034)
truncation	Θ_ϵ	0.76(0.033)	0.92(0.008)	0.56(0.035)	0.52(0.023)

Table 8: Simulation results for B and Θ_ϵ over 50 replications under npn transformation

MLE procedure and examine its evolution over successive iterations. We evaluate the value of the objective function at each iteration, and also compare it to the value of the objective function evaluated at our initializer (screening + Lasso/Ridge) and the estimates

afterward. For illustration purposes, we only show the results for a single realization under Model A with $p_1 = 30, p_2 = 60, n = 100$, although similar results were obtained in other simulation settings. Figure 2 shows the value of the objective function as a function of the iteration under both initialization procedures, while Table 9 shows how the cardinality of the estimates changes over iterations for both initializers. It can be seen that the iterative MLE algorithm significantly improves the value of the objective function over the CAPME initialization and also that the set of directed and undirected edges stabilizes after a couple iterations.

Figure 2: Comparison between Cai's estimate and our estimate



		0	1	2	3	4	5	6	refit
Our initializer	$\widehat{B}^{(k)}$	275	275	275	275	275	275	275	275
	$\widehat{\Theta}_\epsilon^{(k)}$	282	255	247	247	248	248	248	260
CAPME initializer	$\widehat{B}^{(k)}$	433	275	275	275	275	275	275	275
	$\widehat{\Theta}_\epsilon^{(k)}$	979	267	250	249	249	248	248	260

Table 9: Change in cardinality over iterations for B and Θ_ϵ

Based on Figure 2 and Table 9, we notice that Cai et. al's two-step estimator yields larger value of the objective function compared with our initializer that is obtained through screening followed by Lasso. However, over subsequent iterations, both initializers yield the same value in the objective function, which keeps decreasing according to the nature of block-coordinate descent.

4.3 Implementation issues

Next, we introduce some acceleration techniques for the MLE algorithm aiming to reduce computing time, yet maintaining estimation accuracy over iterations.

(p_2+1)-block update. In Section 2, we update B and Θ_ϵ by (6) and (8), respectively, and within each iteration, the updated B is obtained by an application of cyclic p_2-block coordinate descent with respect to each of its columns until convergence. As shown in Section 3.1, the outer 2-block update guarantees the MLE iterative algorithm to converge to a stationary point. However in practice, we can speed up the algorithm by updating B without waiting for it to reach the minimizer for every iteration other than the first one. More precisely, for the alternating search step, we take the following steps when actually implementing the proposed algorithm with initializer $\widehat{B}^{(0)}$ and $\widehat{\Theta}_\epsilon^{(0)}$:

– Iteration 1: update B and Θ_ϵ as follows, respectively:

$$\widehat{B}^{(1)} = \underset{B\in\mathcal{B}_1\times\cdots\times\mathcal{B}_{p_2}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2}(\sigma_\epsilon^{ij})^{(0)}(Y_i - XB_i)^\top(Y_j - XB_j) + \lambda_n \sum_{j=1}^{p_2}\|B_j\|_1 \right\},$$

and

$$\widehat{\Theta}_\epsilon^{(1)} = \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\operatorname{argmin}} \left\{ \log\det\Theta_\epsilon - \operatorname{tr}(\widehat{S}^{(1)}\Theta_\epsilon) + \rho_n\|\Theta_\epsilon\|_{1,\text{off}} \right\},$$

where $\widehat{S}^{(1)}$ is the sample covariance matrix of $\widehat{E}^{(1)} \equiv Y - X\widehat{B}^{(1)}$.

– For iteration $k \geq 2$, while not converged:

· For $j = 1, \cdots, p_2$, update B_j once by

$$\widehat{B}_j^{(k)} = \underset{B_j\in\mathcal{B}_j}{\operatorname{argmin}} \left\{ \frac{(\sigma_\epsilon^{jj})^{(k-1)}}{n}\|Y_j + r_j^{(k)} - XB_j\|_2^2 + \lambda_n\|B_j\|_1 \right\},$$

where

$$r_j^{(k)} = \frac{1}{(\sigma_\epsilon^{jj})^{(k-1)}} \left[\sum_{i=1}^{j-1}(\sigma_\epsilon^{ij})^{(k-1)}(Y_i - X\widehat{B}_i^{(k)}) + \sum_{i=j+1}^{p_2}(\sigma_\epsilon^{ij})^{(k-1)}(Y_i - X\widehat{B}_i^{(k-1)}) \right]. \tag{23}$$

· Update Θ_ϵ by

$$\widehat{\Theta}_\epsilon^{(k)} = \underset{\Theta_\epsilon\in\mathbb{S}_{++}^{p_2\times p_2}}{\operatorname{argmin}} \left\{ \log\det\Theta_\epsilon - \operatorname{tr}(\widehat{S}^{(k)}\Theta_\epsilon) + \rho_n\|\Theta_\epsilon\|_{1,\text{off}} \right\},$$

where $\widehat{S}^{(k)}$ is defined similarly.

Intuitively, for the first iteration, we wait for the algorithm to complete the whole cyclic p_2 block-coordinate descent step, as the first iteration usually achieves a big improvement in the value of the objective function compared to the initialization values, as depicted in

Figure 2. However, in subsequent iterations, the changes in the objective function become relatively small, so we do (p_2+1) successive block-updates in every iteration, and start to update Θ_ϵ once a full p_2 block update in B is completed, instead of waiting for the update in B proceeds cyclically until convergence. In practice, this way of updating B and Θ_ϵ leads to faster convergence in terms of total computing time, yet yields the same estimates compared with the exact 2-block update shown in Section 2.

Parallelization. A number of steps of the MLE algorithm is parallelizable. In the screening step, when applying the de-biased Lasso procedure (Javanmard and Montanari, 2014) to obtain the p-values, we need to implement p_2 separate regressions, which can be distributed to different compute nodes and carried out in parallel. So does the refitting step, in which we refit each column in B in parallel.

Moreover, according to Bradley et al. (2011); Richtárik and Takáč (2012); Scherrer et al. (2012) and a series of similar studies, though the block update in the alternating search step is supposed to be carried out sequentially, we can implement the update in parallel to speed up convergence, yet empirically yield identical estimates. This parallelization can be applied to either the minimization with respect to B within the 2-block update method, or the minimization with respect to each column of B for the (p_2+1)-block update method. Either way, $r_j^{(k)}$ in (23) is substituted by

$$r_{j,\text{parallel}}^{(k)} = \frac{1}{(\sigma_\epsilon^{jj})^{(k-1)}} \sum_{i \neq j}^{p_2} (\sigma_\epsilon^{ij})^{(k-1)} (Y_i - X\widehat{B}_i^{(k-1)}),$$

which is not updated until we have updated B_j's once for all $j = 1, \cdots, p_2$ in parallel.

Table 10 shows the elapsed time for carrying out our proposed algorithm using 2-block/(p_2+1) -block update with/without parallelization, under the simulation setting where we have $p_1 = p_2 = 200, n = 150$. The screening step and refitting step are both carried out in parallel for all four different implementations. [4]

	2-block	(p_2+1)-block	2-block in parallel	(p_2+1)-block in parallel
elapsed time (sec)	5074	2556	848	763

Table 10: Computing time with different update methods

As shown in the table, using (p_2+1)-block update and parallelization both can speed up convergence and reduce computing time, which takes only 1/7 of the computing time compared with using 2-block update without parallelization.

Remark 11. The total computing time depends largely on the number of bootstrapped samples we choose for the stability selection step. For the above displayed results, we used 50 bootstrapped samples to obtain the weight matrix. Nevertheless, one can select the number of bootstrap samples judiciously and reduce them if performance would not be seriously impacted.

4. For parallelization, we distribute the computation on 8 cores.

5. Discussion.

In this paper, we examined multi-layered Gaussian networks, proposed a provably converging algorithm for obtaining the estimates of the key model parameters and established their theoretical properties in high-dimensional settings. Note that we focused on ℓ_1 penalties for both the directed and undirected edges, since it was assumed that the multi-layer network was sparse both between layers and within layers. In many scientific applications, external information may require imposing group penalties, primarily on the directed edge parameters (B). For example, in a gene-protein 2-layer network, genes can be grouped according to their function in pathways and one may be interested in assessing the pathway's impact on proteins. In that case, a group lasso penalty can be imposed. In general, the proposed framework can easily accommodate other types of penalties in accordance to the underlying data generating procedure. The exact form of the error bounds established would be different, depending on the exact choice of penalty selected. Nevertheless, as long as the penalty is convex, all arguments regarding bi-convexity and convergence follow, and we can use similar strategies to bound the statistical error of the estimators, obtained via the iterative algorithm.

Next, we discuss connections of this work to that in Sohn and Kim (2012); Yuan and Zhang (2014); McCarter and Kim (2014). In these papapers, an alternative parameterization of the 2-layer network is adopted. Specifically, all nodes in layersd 1 and 2 are considered jointly and assumed to be drawn from the following Gaussian distribution:

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{YX} & \Omega_Y \end{pmatrix}^{-1}\right),$$

and by conditioning $\boldsymbol{Y}$ on $\boldsymbol{X}$, one obtains

$$\boldsymbol{Y}|\boldsymbol{X} \sim \mathcal{N}\left(-\Omega_Y^{-1}\Omega_{XY}'\boldsymbol{X}, \Omega_Y^{-1}\right). \tag{24}$$

Compare (24) with our model set-up in Section 2.1, the following correspondence holds:

$$B = -\Omega_{XY}\Omega_Y^{-1}, \qquad \Omega_Y = \Theta_\epsilon. \tag{25}$$

Note that the correspondence in (25) is only guaranteed to hold in selective settings. Specifically, at the population level, the correspondence between (Ω_{XY}, Ω_Y) and (B, Θ_ϵ) holds in the absence of any sparsity penalization. Further, in a low-dimensional data setting without penalty terms in the objective function, the estimates from the two parameterizations would be similar provided that the problem is well-conditioned and the sample size reasonably large.

However, the situation is different in high-dimensional settings and in the presence of sparsity penalties. Specifically, given data X and Y, instead of parametrizing the model in terms of (B, Θ_ϵ), the authors in Sohn and Kim (2012); Yuan and Zhang (2014); McCarter and Kim (2014) consider the following optimization problem, parametrized in (Ω_{XY}, Ω_Y):

$$\min_{\Omega_{XY}, \Omega_Y} g(\Omega_{XY}, \Omega_Y) \equiv g_0(\Omega_{XY}, \Omega_Y), + \mathcal{R}(\Omega_{XY}, \Omega_Y) \tag{26}$$

where $g_0(\Omega_{XY}, \Omega_Y) = -\log\det\Omega_Y + \frac{1}{n}\text{tr}\left[(Y + \Omega_{XY}\Omega_Y^{-1}X)'\Omega_Y(Y + \Omega_{XY}\Omega_Y^{-1}X)\right]$ is jointly convex in (Ω_{XY}, Ω_Y), and $\mathcal{R}(\Omega_{XY}, \Omega_Y)$ is some regularization term. In particular, the

element-wise ℓ_1 norm on Ω_Y, and the element-wise ℓ_1 or column-wise ℓ_1 norm (matrix $2,1$ norm) on Ω_{XY} are the main penalties under consideration in those papers.

Despite the convex formulation in (26), we would like to point out that in general, the sparsity pattern in B and Ω_{XY} are not transferable through the regularization term, which underlies a major difference between the formulation in (26) and the one presented in this paper. Given the correspondence in (25), there are two cases where B and Ω_{XY} share the same sparsity pattern: 1) Ω_Y (or Θ_ϵ, equivalently) is diagonal, or 2) both the i^{th} row in B and Ω_{XY} are identically zero, for an arbitrary $i = 1, \cdots, p_1$. However, both settings are fairly restrictive and unlike to occur in many applications.

Note that the linear model represents a natural modeling tool for a number of problems and the regression coefficients have a specific scientific interpretation. This is easily accomplished through the (B, Θ_ϵ)-parametrization, by adding proper regularization to B (e.g., penalty which enforces element-wise sparsity or group-Lasso type of sparsity, etc) if necessary. However, with the (Ω_{XY}, Ω_Y)-parametrization, the underlying sparsity in the true data generating procedure, encoded by B, will not be easily incorporated, and to add a regularization term on Ω_{XY} may lose the scientific interpretability, and may also lead to an estimated B whose sparsity pattern is completely mis-specified, obtained from (25) with $\widehat{\Omega}_{XY}, \widehat{\Omega}_Y$ plugged in.

Another difference we would like to point out is that once we add penalty terms to the objective function in the low dimensional setting, or switch to the high dimensional setting (as considered in Sohn and Kim (2012) and Yuan and Zhang (2014)), the correspondence between the optimizer(s) of (1) and the optimizer(s) of (26) become difficult to connect analytically.

Acknowledgments

George Michailidis was supported by NSF awards DMS-1228164 and DMS-1545277 and NIH award 7R21GM10171903. Moulinath Banerjee was supported by NSF award DMS-1308890.

Appendix A. Proofs for Main Theorems

Proof of Theorem 1. We initialize the algorithm at $(\widehat{B}^{(0)}, \widehat{\Theta}_\epsilon^{(0)}) \in \mathrm{dom}(f)$. Then for all $k \geq 1$,

$$\widehat{B}^{(k)} = \underset{B}{\mathrm{argmin}}\, f(B, \widehat{\Theta}_\epsilon^{(k-1)}), \tag{27}$$

$$\widehat{\Theta}_\epsilon^{(k)} = \underset{\Theta_\epsilon}{\mathrm{argmin}}\, f(\widehat{B}^k, \Theta_\epsilon). \tag{28}$$

Now, consider a limit point $(B^\infty, \Theta_\epsilon^\infty)$ of the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\geq 1}$. Note that such limit point exists by Bolzano-Weierstrass theorem since the sequence $\{(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\geq 1}$ is bounded. Consider a subsequence $\mathcal{K} \subseteq \{1, 2, \cdots\}$ such that $(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)})_{k\in\mathcal{K}}$ converges to $(B^\infty, \Theta_\epsilon^\infty)$. Now for the bounded sequence $\{(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\in\mathcal{K}}$, without loss of generality,[5] we can say that

$$\{(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)})\}_{k\in\mathcal{K}} \to (\widetilde{B}^\infty, \widetilde{\Theta}_\epsilon^\infty), \quad \text{for some } (\widetilde{B}^\infty, \widetilde{\Theta}_\epsilon^\infty) \in \mathrm{dom}(f).$$

By (27) it follows immediately that $\widetilde{\Theta}_\epsilon^\infty = \Theta_\epsilon^\infty$. Also, the following inequality holds:

$$f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k+1)}) \leq f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)}) \leq f(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}).$$

Thus, by letting $k \to \infty$ over $\mathcal{K}$, we have

$$f(B^\infty, \Theta_\epsilon^\infty) \leq f(\widetilde{B}^\infty, \Theta_\epsilon^\infty) \leq f(B^\infty, \Theta_\epsilon^\infty),$$

since f is continuous. This implies that

$$f(\widetilde{B}^\infty, \Theta_\epsilon^\infty) = f(B^\infty, \Theta_\epsilon^\infty). \tag{29}$$

Next, since $f(\widehat{B}^{(k+1)}, \widehat{\Theta}_\epsilon^{(k)}) \leq f(B, \widehat{\Theta}_\epsilon^{(k)})$, for all $B \in \mathbb{R}^{p_1\times p_2}$, let k grow along $\mathcal{K}$, and we obtain the following:

$$f(\widetilde{B}^\infty, \Theta_\epsilon^\infty) \leq f(B, \Theta_\epsilon^\infty), \quad \forall B \in \mathbb{R}^{p_1\times p_2}.$$

It then follows from (29) that

$$f(B^\infty, \Theta_\epsilon^\infty) \leq f(B, \Theta_\epsilon^\infty), \quad \forall B \in \mathbb{R}^{p_1\times p_2}. \tag{30}$$

Finally, note that $f(\widehat{B}^{(k)}, \widehat{\Theta}_\epsilon^{(k)}) \leq f(\widehat{B}^{(k)}, \Theta_\epsilon)$, for all $\Theta \in \mathbb{S}_{++}^{p_2\times p_2}$. As before, let k grow along $\mathcal{K}$ and with the continuity of f, we obtain:

$$f(B^\infty, \Theta_\epsilon^\infty) \leq f(B^\infty, \Theta_\epsilon), \quad \forall \Theta_\epsilon \in \mathbb{S}_{++}^{p_2\times p_2}. \tag{31}$$

Now, (30) and (31) together imply that $(B^\infty, \Theta_\epsilon^\infty)$ is a coordinate-wise minimum of f and by Fact 1, also a stationary point of f. □

5. switching to some further subsequence of $\mathcal{K}$ if necessary.

Proof of Theorem 2. The statement of Theorem 2 is a variation of Proposition 4.1 in Basu and Michailidis (2015), and its proof follows directly from the proof of the proposition in Basu and Michailidis (2015, Appendix B). We only outline how the statement differs. In the original statement of Proposition 4.1 in Basu and Michailidis (2015), the authors provide the error bound for $\bar{\beta}$, obtained as per (14) whose dimension is qp^2 with q denoting the true lag of the vector-autoregressive process, under an RE condition for $\bar{\Gamma}$ and a deviation bound for $(\bar{\gamma}, \bar{\Gamma})$. For our problem, we impose a similar RE condition on $\widehat{\Gamma}$ and deviation bound on $(\widehat{\gamma}, \widehat{\Gamma})$, so as to yield a bound on $\widehat{\beta}$ that lies in a p_1p_2-dimensional space. □

Proof of Theorem 3. The statement of this theorem is a variation of Theorem 1 in Ravikumar et al. (2011), so here, instead of providing a complete proof of the theorem, we only outline how the estimation problem differs in our setting, as well as the required changes in its proof.

In Ravikumar et al. (2011), the authors consider the optimization problem in (15), and show that for a random realization, with certain sample size requirement and choice of the regularization parameter, the following bound for $\bar{\Theta}_\epsilon$ holds with probability at least $1-1/p_2^\tau$ for some $\tau > 2$:

$$\|\bar{\Theta}_\epsilon - \Theta^*_\epsilon\|_\infty \leq \{2(1+8\xi^{-1})\kappa_{H^*}\}\bar{\delta}_f(p_2^\tau, n), \tag{32}$$

where $\bar{\delta}(r,n)$ is defined as

$$\bar{\delta}(r,n) := 8(1+4\sigma^2)\max_i(\Sigma^*_{\epsilon,ii})\sqrt{\frac{2\log(4r)}{n}}. \tag{33}$$

The quantity $\bar{\delta}(p_2^\tau, n)$ that shows up in expression (32) is the bound for $\|S - \Sigma^*_\epsilon\|_\infty \equiv \|\widehat{\Sigma}_\epsilon - \Sigma^*_\epsilon\|_\infty$. In particular, in Lemma 8 (Ravikumar et al., 2011), they show that with probability at least $1 - 1/p_2^\tau$, $\tau > 2$, the following bound holds:

$$\|S - \Sigma^*_\epsilon\|_\infty \leq \bar{\delta}(p_2^\tau, n).$$

In our optimization problem (13), we are using $\widehat{S}$ instead of S, hence a bound for $\|\widehat{S} - \Sigma^*_\epsilon\|_\infty$ is necessary, and the remaining argument in the proof of Theorem 1 (Ravikumar et al., 2011) will follow through.

Therefore in our theorem statement, we use $g(\nu_\beta)$ as a bound for $\|\widehat{S} - \Sigma^*_\epsilon\|_\infty$ then yield the bound for $\|\widehat{\Theta}_\epsilon - \Theta^*_\epsilon\|_\infty$, since we are using the surrogate error $\widehat{E} = Y - X\widehat{B}$ in the estimation, instead of the true error E. □

Proof of Theorem 4. We first consider part (I) of the theorem. Note that by (5), $\widehat{\beta}^{(0)}$ can be equivalently written as

$$\widehat{\beta}^{(0)} \equiv \underset{\beta \in \mathbb{R}^{p_1 \times p_2}}{\operatorname{argmin}} \left\{ -2\beta'\gamma^0 + \beta'\Gamma^0\beta + \lambda_n^0\|\beta\|_1 \right\}, \tag{34}$$

where

$$\Gamma^{(0)} = \mathrm{I} \otimes \frac{X'X}{n}, \quad \gamma^{(0)} = (\mathrm{I} \otimes X')\mathrm{vec}Y/n.$$

Consider the following events:

E1. $\left\{ \frac{X'X}{n} \sim RE(\varphi^*, \phi^*) \right\}$,

E2. $\left\{ \frac{1}{n} \left\| X'E \right\|_\infty \leq c_2 \left[\Lambda_{\max}(\Sigma_X^*) \Lambda_{\max}(\Sigma_\epsilon^*) \right]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}} \right\}$.

Note that **E1** $\cap$ **E2** implies the following events:

$$\Gamma^{(0)} \equiv \mathrm{I} \otimes \frac{X'X}{n} \sim RE(\varphi^*, \phi^*), \quad \text{where } \varphi^* = \Lambda_{\min}(\Sigma_X^*)/2.$$

and

$$\|\gamma^{(0)} - \Gamma^{(0)} \beta^*\|_\infty = \frac{1}{n} \left\| X'E \right\|_\infty \leq c_2 \left[\Lambda_{\max}(\Sigma_X^*) \Lambda_{\max}(\Sigma_\epsilon^*) \right]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}}. \tag{35}$$

Hence, by Proposition 4.1 of Basu and Michailidis (2015), the bound (21) holds on **E1** $\cap$ **E2**.

By Lemmas 1 and 2, $\mathbb{P}(\mathbf{E1})$ is at least $1 - 2\exp(-c_3 n)$, for some $c_3 > 0$. By Lemma 3, $\mathbb{P}(\mathbf{E2})$ is at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$ for some $c_1 > 0$, $c_2 > 1$. Hence, with probability at least

$$\mathbb{P}\left(\mathbf{E1} \cap \mathbf{E2}\right) \geq 1 - \mathbb{P}\left(\mathbf{E1}^c\right) - \mathbb{P}\left(\mathbf{E2}^c\right),$$

the bound in (21) holds, which proves the first part of (I). In particular, we have $\|\hat{\beta}^0 - \beta^*\|_1 \leq \nu_\beta^{(0)} \sim O(\sqrt{\log(p_1 p_2)/n})$ on **E1** $\cap$ **E2**.

To prove the second part of (I), note that by Theorem 3 the bound in (22) holds when B1-B3 are satisfied. Now, from the argument above, B1 holds on the event **E1** $\cap$ **E2**. Also, from the proof of Proposition 3, B2 is satisfied, i.e.,

$$\left\| \widehat{S}^{(0)} - \Sigma_\epsilon^* \right\|_\infty \leq g(\nu_\beta^{(0)}), \quad \text{where } \widehat{S}^{(0)} = \frac{1}{n}(Y - X\widehat{B}^{(0)})'(Y - X\widehat{B}^{(0)}), \tag{36}$$

on **E1** $\cap$ **E2** $\cap$ **E3** $\cap$ **E4**, where the events **E3** and **E4** are given by:

E3. $\left\{ \left\| \frac{E'E}{n} - \Sigma_\epsilon^* \right\|_\infty \leq \sqrt{\frac{\log 4 + \tau_2 \log p_2}{c_\epsilon^* n}} \right\}$ for some $\tau_2 > 2$ and $c_\epsilon^* > 0$ that depends on Σ_ϵ^*,

E4. $\left\{ \left\| \frac{X'X}{n} - \Sigma_X^* \right\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}} \right\}$ for some $\tau_1 > 2$ and $c_X * > 0$ that depends on Σ_X^*.

Therefore, the probability of the bound for $\widehat{\Theta}_\epsilon^{(0)}$ in (22) to hold is at least

$$\mathbb{P}\left(\mathbf{E1} \cap \mathbf{E2} \cap \mathbf{E3} \cap \mathbf{E4}\right), \tag{37}$$

By Lemma 2, Lemma 3 and the proof of Proposition 3, the probability in (37) is lower bounded by:

$$1 - 2\exp(-c_3 n) - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)] - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2}.$$

Consider the following two cases where the relative order of p_1 and p_2 differ. Case 1: $p_1 \prec p_2$, then $\nu_\Theta^{(0)} \sim O(\sqrt{\log p_2/n})$; case 2: $p_1 \succsim p_2$, then $\nu_\Theta^{(0)} \sim O\left(\log(p_1 p_2)/n\right)$. In either case, since we are assuming $\log(p_1 p_2)/n$ to be a small quantity and it follows that $\sqrt{\log(p_1 p_2)/n} \succsim \log(p_1 p_2)/n$, the following bound always holds:

$$\nu_\Theta^{(0)} \leq C_\Theta \sqrt{\frac{\log(p_1 p_2)}{n}} \equiv M_\Theta,$$

where C_Θ is some large fixed constant that bounds the constant terms in front of $\sqrt{\log(p_1p_2)/n}$.

Now we consider part (II) of the theorem. Note that for each $k \geq 1$, $\widehat{\beta}^{(k)}$ and $\widehat{\Theta}_\epsilon^{(k)}$ are obtained via solving the following two optimizations:

$$\widehat{\beta}^{(k)} = \underset{\beta \in \mathbb{R}^{p_1 \times p_2}}{\operatorname{argmin}} \left\{ -2\beta' \widehat{\gamma}^{(k-1)} + \beta' \widehat{\Gamma}^{(k-1)} \beta + \lambda_n \|\beta\|_1 \right\}, \tag{38}$$

$$\widehat{\Theta}_\epsilon^{(k)} = \underset{\Theta_\epsilon \in \mathbb{S}_{++}^{p_2 \times p_2}}{\operatorname{argmin}} \left\{ \log\det \Theta_\epsilon - \operatorname{tr}(\widehat{S}^{(k)} \Theta_\epsilon) + \rho_n \|\Theta_\epsilon\|_{1,\text{off}} \right\}, \tag{39}$$

where

$$\widehat{\gamma}^{(k)} = \widehat{\Theta}^{(k)} \otimes \frac{X'Y}{n}, \quad \widehat{\Gamma}^{(k)} = \widehat{\Theta}^{(k)} \otimes \frac{X'X}{n}, \quad \widehat{S}^{(k)} = \frac{1}{n}(Y - X\widehat{B}^{(k)})'(Y - X\widehat{B}^{(k)}).$$

Consider the bound on $\hat{\beta}^{(k)}$ for $k = 1$. The argument is similar to that of $\hat{\beta}^{(0)}$, with appropriate modifications to account for the fact that the objective function now involves log likelihood instead of least squares. Formally, we consider the event $\mathbf{E1} \cap \mathbf{E2} \cap \mathbf{E3} \cap \mathbf{E4} \cap \mathbf{E5}$, where

$$\mathbf{E5.} \ \left\{ \tfrac{1}{n} \|X'E\Theta_\epsilon^*\|_\infty \leq c_2 \left[\tfrac{\Lambda_{\max}(\Sigma_X^*)}{\Lambda_{\min}(\Sigma_\epsilon^*)} \right]^{1/2} \sqrt{\tfrac{\log(p_1p_2)}{n}} \right\}.$$

Note that $\{\|\widehat{\Theta}_\epsilon^{(0)} - \Theta_\epsilon^*\|_\infty \leq \nu_\Theta^{(0)}\}$ holds on this event. By Lemma 3, $\mathbb{P}(\mathbf{E5}) \geq 1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1p_2)]$. Combining this with the lower bound on (37) and the sample size requirement (note this sample size requirement can be relaxed to $n \gtrsim \log(p_1p_2)$ if $p_1 \prec p_2$), we obtain that with probability at least

$$1 - 1/p_1^{\tau_1 - 2} - 1/p_2^{\tau_2 - 2} - 12c_1 \exp[-(c_2^2 - 1)\log(p_1p_2)] - 2\exp[-c_3 n],$$

the following three events hold simultaneously:

A1' $\|\widehat{\Theta}_\epsilon^{(0)} - \Theta_\epsilon^*\|_\infty \leq \nu_\Theta^{(0)} \precsim O(\sqrt{\log(p_1p_2)/n})$;

A2' $\widehat{\Gamma}^{(0)} \sim RE(\varphi^{(0)}, \phi^{(0)})$ where

$$\varphi^{(0)} \geq \frac{\Lambda_{\min}(\Sigma_X^*)}{2}(\min_i \psi^i - dM_\Theta) \quad \text{and} \quad \phi^{(0)} \leq \frac{\log p_1}{n} \frac{\Lambda_{\min}(\Sigma_X^*)}{2}(\max_j \psi^j + dM_\Theta);$$

A3' $\|\widehat{\gamma}^{(0)} - \widehat{\Gamma}^{(0)}\beta^*\|_\infty \leq \mathbb{Q}(\nu_\Theta^{(0)})\sqrt{\frac{\log(p_1p_2)}{n}}$ with the expression for $\mathbb{Q}(\cdot)$ given in (16).

By Theorem 2, by choosing $\lambda_n \geq 4\mathbb{Q}(M_\Theta)\sqrt{\frac{\log(p_1p_2)}{n}}$, the following bound holds:

$$\|\widehat{\beta}^{(1)} - \beta^*\|_1 \leq 64 s^{**} \lambda_n / \varphi^{(0)}. \tag{40}$$

The error bound for $\widehat{\Theta}_\epsilon^{(1)}$ can now be established using the same argument for $\hat{\Theta}_\epsilon^{(0)}$, with the only difference that now we consider the event $\mathbf{E1} \cap \ldots \cap \mathbf{E5}$ instead of $\mathbf{E1} \cap \ldots \cap \mathbf{E4}$ and use (40) instead of (21).

Note that an upper bound for the leading term of the right hand side of (40) is at most of the order $O(\sqrt{\log(p_1p_2)/n})$, and can be written as

$$C_\beta \left(s^{**} \sqrt{\frac{\log(p_1p_2)}{n}} \right) \equiv M_\beta,$$

with C_β being some potentially large number that bounds the constant term. Notice that M_β is of the same order as $\nu_\beta^{(0)}$; thus, for $\widehat{\Theta}_\epsilon^{(1)}$, we can also achieve the following bound:

$$\|\widehat{\Theta}_\epsilon^{(1)} - \Theta_\epsilon^*\|_\infty \leq M_\Theta,$$

with high probability since we are assuming C_Θ to be some potentially large number.

Note that the events $\mathbf{E1}, \ldots, \mathbf{E5}$ rely only on the parameters and not on the estimated quantities, and on their intersection we have uniform upper bounds on the errors of $\widehat{\beta}^{(k)}$ and $\widehat{\Theta}_\epsilon^k$ for $k = 0, 1$. Hence the error bounds for $k = 1$ can be used to invoke Theorems 2 and 3 inductively on realizations X and E from the set $\mathbf{E1} \cap \ldots \cap \mathbf{E5}$ to provide high probability error bounds for all subsequent iterates as well. This leads to the uniform error bounds of part (II) with the desired probability.

□

Proof of Theorem 5. First, we note that with a Bonferroni correction, the family-wise type I error will be automatically controlled at level α. Hence, we will focus on the power of the screening step. Also, from Theorem 7 of Javanmard and Montanari (2014), it is easy to see that all the arguments below hold for a large set of random realizations of X, whose probability approaches 1 under the specified asymptotic regime when the eigenvalues of Σ_X are bounded away from 0 and infinity.

Let $B^* = \begin{bmatrix} B_1^* & \cdots & B_{p_2}^* \end{bmatrix}$ denote the true value of the regression coefficients and $\check{B}_j, j = 1, \cdots, p_2$ denote the estimates given by the de-biased Lasso procedure in Javanmard and Montanari (2014). With the given level for sparsity, by Theorem 8 in Javanmard and Montanari (2014), each $\check{B}_j$ satisfies the following:

$$\sqrt{n}(\check{B}_j - B_j^*) = Z + \Delta,$$

where $Z \sim \mathcal{N}\left(0, \sigma^2 M_j \widehat{\Sigma}_X M_j'\right)$ and Δ vanishes asymptically. Here $\widehat{\Sigma}_X$ is the sample covariance matrix of the predictors X, σ is the population noise level of the error term ϵ_j, and M_j is the matrix corresponding to the jth regression, produced by the procedure described in Javanmard and Montanari (2014)[6]. Let $\check{B}_{j,i}$ denote the ith coordinate of the jth regression coefficient vector $\check{B}_j$ and $\check{\Sigma}_j$ be the covariance matrix of the estimator $\check{B}_j$, then

$$\check{\Sigma}_j = \frac{\sigma^2}{n} M_j \widehat{\Sigma}_X M_j',$$

and in particular, the variance of $\check{B}_{j,i}$ is $\check{\Sigma}_{j,ii} := \check{\sigma}_{ii}^j$. Using these notations, for a prespecified level α, the test statistics for testing $H_0^{ji} : B_{j,i}^* = 0$ vs. $H_A^{ji} : B_{j,i}^* \neq 0$, for all $i =$

6. Details of the procedure is described in p.2871 in Javanmard and Montanari (2014), with M being an intermediate quantity obtained by solving an optimization problem.

$1, \cdots, p_1; j = 1, \cdots, p_2$ can be equivalently written as

$$\widehat{T}_{j,i} = \begin{cases} 1 & \text{if } |\check{B}_{j,i}|/\check{\sigma}_{ii}^{j} > z_{\alpha/(2p_1p_2)}, \\ 0 & \text{otherwise.} \end{cases}$$

where z_α denotes the upper α quantiles of $\mathcal{N}(0,1)$.

Define the "family-wise" power as follows:

$$\begin{aligned} \mathbb{P}\left(\text{all true alternatives are detected}\right) &= \mathbb{P}\left(\bigcap_{1\le j\le p_2}\bigcap_{k\in S_j^*}\{\widehat{T}_{j,k}=1\}\right) \\ &= 1-\mathbb{P}\left(\bigcup_{1\le j\le p_2}\bigcup_{k\in S_j^*}\{\widehat{T}_{j,k}=0\}\right). \end{aligned}$$

Correspondingly, the family-wise type II error can be written as

$$\mathbb{P}\left(\bigcup_{1\le j\le p_2}\bigcup_{k\in S_j^*}\{\widehat{T}_{j,k}=0\}\right) \le \sum_{j=1}^{p_2}\sum_{k\in S_j^*}\mathbb{P}\left(\widehat{T}_{j,k}=0\right). \tag{41}$$

By Theorem 16 in Javanmard and Montanari (2014), asymptotically, $\forall k \in S_j, j = 1, \cdots, p_2$,

$$\mathbb{P}\left(\widehat{T}_{j,k}=0\right) \le 1 - G\left(\frac{\alpha}{p_1p_2}, \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\right); \qquad 0<\gamma\le \min|B^*_{j,k}|, \quad \forall k\in S_j, j=1,\cdots,p_2. \tag{42}$$

Here

$$G(\alpha,u) \equiv 2-\mathbb{P}(\Phi < z_{\alpha/2}+u) - \mathbb{P}(\Phi < z_{\alpha/2}-u),$$

where we use Φ to denote the random variable following a standard Gaussian distribution and the choice of n in (42) doesn't depend on k. Hence, (42) can be rewritten as

$$\begin{aligned} \mathbb{P}\left(\widehat{T}_{j,k}=0\right) &\le 1 - G\left(\frac{\alpha}{p_1p_2}, \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\right) \\ &= \mathbb{P}\left(\Phi < z_{\alpha/(2p_1p_2)} - \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\right) - \mathbb{P}\left(\Phi > z_{\alpha/(2p_1p_2)} + \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}}\right) \\ &\le \mathbb{P}\left(\Phi > \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - z_{\alpha/(2p_1p_2)}\right), \end{aligned} \tag{43}$$

where we use Φ to denote the random variable following a standard Gaussian distribution.

Note that the following inequality holds for standard Normal percentiles:

$$2e^{-t^2} \le \mathbb{P}(|\Phi| > t) \le e^{-t^2/2},$$

and by taking the inverse function, the following inequality holds:

$$\sqrt{-\log\frac{y}{2}} \leq z_{y/2} \leq \sqrt{-2\log y}.$$

Letting $y = \frac{\alpha}{p_1 p_2}$, it follows that

$$\left(-\log\frac{\alpha}{2p_1p_2}\right)^{1/2} \leq z_{\alpha/(2p_1p_2)} \leq \left(-2\log\frac{\alpha}{p_1p_2}\right)^{1/2},$$

hence

$$\mathbb{P}\left(\Phi > \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - z_{\alpha/(2p_1p_2)}\right) \leq \mathbb{P}\left(\Phi > \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - \sqrt{-2\log\frac{\alpha}{p_1p_2}}\right).$$

Now given

$$\frac{\log(p_1p_2)}{n} \to 0,$$

it follows that

$$\frac{\sqrt{2\log\left(\frac{p_1p_2}{\alpha}\right)}}{\sqrt{n}/\sigma[\Sigma^{-1}_{k,k}]^{1/2}} \to 0,$$

indicating that for sufficiently large n, the following lower bound holds for some constant $c_0 > 0$:

$$\left(\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - \sqrt{-2\log\frac{\alpha}{p_1p_2}}\right) \geq c_0\sqrt{n}.$$

Note that c_0 is univeral for all choices of k, since this lower bound can be achieved by substituting $\Sigma^{-1}_{k,k}$ by $(1/\Lambda_{\min}(\Sigma_X))$, which is assumed to be bounded away from infinity. Combined with the fact that $\mathbb{P}(\Phi > t) \leq e^{-t^2/2}$, the last expression in (43) can thus be bounded by

$$\mathbb{P}\left(\Phi > \frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - z_{\alpha/(2p_1p_2)}\right) \leq \exp\left[-\frac{1}{2}\left(\frac{\sqrt{n}\gamma}{\sigma[\Sigma^{-1}_{k,k}]^{1/2}} - \sqrt{-2\log\frac{\alpha}{p_1p_2}}\right)^2\right] \leq e^{-c_1 n}, \tag{44}$$

for some universal constant $c_1 > 0$, and the bound in (44) holds uniformly for all $k \in S_j, \forall j$. Combine (41), (42) and (44), it follows that

$$\mathbb{P}\left(\bigcup_{1\leq j\leq p_2}\bigcup_{k\in S_j^*}\{\widehat{T}_{j,k} = 0\}\right) \leq s^* p_2 \exp(-c_1 n). \tag{45}$$

Now with $\log(p_1p_2)/n = o(1)$ and the given sparsity level, that is, $s^* = o(\sqrt{n}/\log p_1)$, it follows that

$$s^* p_2 \exp(-c_1 n) = o(1),$$

and by (45), we have:

$$\mathbb{P}\left(\text{family-wise type II error}\right) \to 0, \quad \Leftrightarrow \quad \mathbb{P}\left(\text{family-wise power}\right) \to 1.$$

This is equivalent to establishing that, given $\log(p_1 p_2)/n \to 0$, the screening step recovers the true support sets S_j^* for all $j = 1, 2, \cdots, p_2$ with high probability, while keeping the family-wise type I error rate under control. □

Appendix B. Proofs for Propositions and Auxillary Lemmas

In this subsection, we provide proofs for the propositions presented in Section 3, which requires several auxillary lemmas, whose proofs are presented along the context.

To prove Proposition 1, we need the following two lemmas. Lemma 1 was originally provided as Lemma B.1 in Basu and Michailidis (2015), which states that if the sample covariance matrix of X satisfies the RE condition and Θ is diagonally dominant, then $(X'X/n) \otimes \Theta$ also satisfies the RE condition. Here we omit its proof and only state the main result. Lemma 2 verifies that with high probability, the sample covariance matrix of the design matrix X satisfies the RE condition.

Lemma 1. *If $X'X/n \sim RE(\varphi^*, \phi^*)$, and Θ is diagonally dominant, that is, $\psi^i := \sigma^{ii} - \sum_{j \neq i} \sigma^{ij} > 0$ for all $i = 1, 2, \cdots, p_2$, where σ^{ij} is the ijth entry in Θ, then*

$$\Theta \otimes X'X/n \sim RE\left(\varphi^* \min_i \psi^i, \phi^* \max_i \psi^i\right).$$

Lemma 2. *With probability at least $1 - 2\exp(-c_3 n)$, for a zero-mean sub-Gaussian random design matrix $X \in \mathbb{R}^{n \times p_1}$, its sample covariance matrix $\widehat{\Sigma}_X$ satisfies the RE condition with parameter φ^* and ϕ^*, i.e.,*

$$\widehat{\Sigma}_X \sim RE(\varphi^*, \phi^*), \tag{46}$$

where $\widehat{\Sigma}_X = X'X/n$, $\varphi^* = \Lambda_{\min}(\Sigma_X^*)/2$, $\phi^* = \varphi^* \log p_1/n$.

Proof. To prove this lemma, we first use Lemma 15 in Loh and Wainwright (2012), which states that if $X \in \mathbb{R}^{n \times p}$ is zero-mean sub-Gaussian with parameter (Σ, σ^2), then there exists a universal constant $c > 0$ such that

$$\mathbb{P}\left(\sup_{v \in \mathbb{K}(2s)} \left| \frac{\|Xv\|_2^2}{n} - \mathbb{E}\left[\frac{\|Xv\|_2^2}{n}\right] \right| \geq t \right) \leq 2\exp\left(-cn \min\left(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right) + 2s \log p\right), \tag{47}$$

where $\mathbb{K}(2s)$ is a set of $2s$ sparse vectors, defined as

$$\mathbb{K}(2s) := \{v \in \mathbb{R}^p : \|v\| \leq 1, \|v\|_0 \leq 2s\}.$$

By taking $t = \frac{\Lambda_{\min}(\Sigma_X^*)}{54}$, with probability at least $1 - 2\exp\left(-c'n + 2s \log p_1\right)$ for some $c' > 0$, the following bound holds:

$$|v'(\widehat{\Sigma}_X - \Sigma_X^*)v| \leq \frac{\Lambda_{\min}(\Sigma_X^*)}{54}, \quad \forall v \in \mathbb{K}(2s). \tag{48}$$

Then applying supplementary Lemma 13 in Loh and Wainwright (2012), for an estimator $\widehat{\Sigma}_X$ of Σ_X^* satisfying the deviation condition in (48), the following RE condition holds:

$$v' S_x v \geq \frac{\Lambda_{\min}(\Sigma_X^*)}{2}\|v\|_2^2 - \frac{\Lambda_{\min}(\Sigma_X^*)}{2s}\|v\|_1^2.$$

Finally, set $s = c''n/4\log p_1$, then with probability at least $1 - 2\exp(-c_3 n)$ ($c_3 > 0$), $\widehat{\Sigma}_X \sim RE(\varphi^*, \phi^*)$ with $\varphi^* = \Lambda_{\min}(\Sigma_X^*)/2$, $\phi^* = \varphi^* \log p_1/n$. □

With the above two lemmas, we are ready to prove Proposition 1.

Proof of Proposition 1. We first show that if Θ_ϵ^* is diagonally dominant, then $\widehat{\Theta}_\epsilon$ is also diagonally dominant provided that the error of $\widehat{\Theta}_\epsilon$ is of the given order and n is sufficiently large. Define

$$\widehat{\psi}^i = \widehat{\sigma}_\epsilon^{ii} - \sum_{j\neq i} \widehat{\sigma}_\epsilon^{ij},$$

where $\widehat{\sigma}_\epsilon^{ij}$ is the ijth entry of $\widehat{\Theta}_\epsilon$, then $\widehat{\psi}^i$ is the gap between the diagonal entry and the off-diagonal entries of row i in matrix $\widehat{\Theta}_\epsilon$. We can decompose $\widehat{\psi}^i$ into the following:

$$\widehat{\psi}^i = \left[\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}\right] + \left[(\widehat{\sigma}_\epsilon^{ii} - \sigma_\epsilon^{ii}) + \sum_{j\neq i}(\sigma_\epsilon^{ij} - \widehat{\sigma}_\epsilon^{ij})\right].$$

Recall that we define ψ_i as $\psi^i = \sigma_\epsilon^{ii} - \sum_{j\neq i}^{p_2}\sigma_\epsilon^{ij}$. Hence

$$\begin{aligned}
\min \widehat{\psi}^i &\geq \min_i \psi^i - \left|\left|\left|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\right|\right|\right|_\infty \geq \min_i(\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}) - d\nu_\Theta = \min \psi^i - d\nu_\Theta, \\
\max \widehat{\psi}^i &\leq \max_i \psi^i + \left|\left|\left|\widehat{\Theta}_\epsilon - \Theta_\epsilon^*\right|\right|\right|_\infty \leq \max_i(\sigma_\epsilon^{ii} - \sum_{j\neq i}\sigma_\epsilon^{ij}) + d\nu_\Theta = \max \psi^i + d\nu_\Theta.
\end{aligned} \tag{49}$$

Now given $\nu_\Theta = \eta_\Theta \frac{\log p_2}{n} = O(\sqrt{\log p_2/n})$, with $n \gtrsim d^2 \log p_2$, $d\nu_\Theta = o(1)$, and it follows that

$$\min_i \psi^i - d\nu_\Theta \geq 0.$$

Now by Lemma 2, $X'X/n \sim RE(\varphi^*, \phi^*)$ with high probability. Combine with Lemma 1 and inequality (49), with probability at least $1 - 2\exp(-c_3 n)$ for some $c_3 > 0$, $\widehat{\Gamma}$ satisfies the following RE condition:

$$\widehat{\Gamma} = \widehat{\Theta}_\epsilon \otimes (X'X/n) \sim RE\left(\varphi^*(\min_i \psi^i - d\nu_\Theta), \phi^* \max_i(\psi^i + d\nu_\Theta)\right), \tag{50}$$

where $\varphi^* = \Lambda_{\min}(\Sigma_X^*)/2$, $\phi^* = \varphi^* \log p_1/n$. □

To prove Proposition 2, we first prove Lemma 3.

Lemma 3. *Let $X \in \mathbb{R}^{n\times p}$ be a zero-mean sub-Gaussian matrix with parameter (Σ_X, σ_X^2) and $E \in \mathbb{R}^{n\times p_2}$ be a zero-mean sub-Gaussian matrix with parameters $(\Sigma_\epsilon, \sigma_\epsilon^2)$. Moreover, X and E are independent. Let $\Theta_\epsilon := \Sigma_\epsilon^{-1}$, then if $n \gtrsim \log(p_1p_2)$, the following two expressions hold with probability at least $1 - 6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$ for some $c_1 > 0, c_2 > 1$, respectively:*

$$\frac{1}{n}\left\|X'E\right\|_\infty \leq c_2\left[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon)\right]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}, \tag{51}$$

and

$$\frac{1}{n}\left\|X'E\Theta_\epsilon\right\|_\infty \leq c_2\left[\frac{\Lambda_{\max}(\Sigma_X)}{\Lambda_{\min}(\Sigma_\epsilon)}\right]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}. \tag{52}$$

Proof. The proof of this lemma uses Lemma 14 in Loh and Wainwright (2012), in which they show that if $X \in \mathbb{R}^{n\times p_1}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2) and $Y \in \mathbb{R}^{n\times p_2}$ is a zero-mean sub-Gaussian matrix with parameters (Σ_y, σ_y^2), then if $n \gtrsim \log(p_1p_2)$,

$$\mathbb{P}\left(\left\|\frac{Y'X}{n} - cov(y_i, x_i)\right\|_\infty \geq t\right) \leq 6p_1p_2\exp\left(-cn\min\left\{\frac{t^2}{(\sigma_x\sigma_y)^2}, \frac{t}{\sigma_x\sigma_y}\right\}\right),$$

where X_i and Y_i are the ith row of X and Y, respectively.

Here, we replace Y by E, and since E and X are independent, $cov(X_i, E_i) = 0$. Let $t = c_2\sigma_X\sigma_\epsilon\sqrt{\log(p_1p_2)/n}$, $c_2 > 1$ we get

$$\mathbb{P}\left(\left\|\frac{X'E}{n}\right\|_\infty \geq c_2\sigma_X\sigma_\epsilon\sqrt{\frac{\log(p_1p_2)}{n}}\right) \leq 6c_1(p_1p_2)^{1-c_2^2} = 6c_1\exp\left[-(c_2^2-2)\log(p_1p_2)\right].$$

Note that the sub-Gaussian parameter satisfies $\sigma_X^2 \leq \max_i(\Sigma_{X,ii}) \leq \Lambda_{\max}(\Sigma_X)$. This directly gives the bound in (51).

To obtain the bound in (52), we note that if E is sub-Gaussian with parameters $(\Sigma_\epsilon, \sigma_\epsilon^2)$, then $E\Theta$ is sub-Gaussian with parameter $(\Theta, \theta_\epsilon^2)$, where

$$\theta_\epsilon^2 \leq \max_i(\Theta_{\epsilon,ii}) \leq \Lambda_{\max}(\Theta_\epsilon) = \frac{1}{\Lambda_{\min}(\Sigma_\epsilon)}.$$

Then we replace Y by $E\Theta$ and yield the bound in (52). □

As a remark, here we note that the event in (51) and (52) may not be independent. However, the two events hold simultaneously with probability at least $1-2c_2\exp[-c_2\log(p_1p_2)]$, with this crude bound for probability hold for sure.

Now we are ready to prove Proposition 2.

Proof of Proposition 2. First we note that

$$X'E\widehat{\Theta}_\epsilon = X'E\Theta_\epsilon + X'E(\widehat{\Theta}_\epsilon - \Theta_\epsilon^*),$$

which directly gives the following inequality:

$$\|\widehat{\gamma} - \widehat{\Gamma}\beta^*\|_\infty = \frac{1}{n}\left\|X'E\widehat{\Theta}_\epsilon\right\|_\infty \leq \frac{1}{n}\left\|X'E\Theta_\epsilon^*\right\|_\infty + \frac{1}{n}\left\|X'E(\widehat{\Theta}_\epsilon - \Theta_\epsilon^*)\right\|_\infty. \tag{53}$$

Now we would like to bound the two terms separately.

The first term can be bounded by (52) in Lemma 3, that is,

$$\frac{1}{n}\left\|X'E\Theta_\epsilon^*\right\|_\infty \le c_2\left[\frac{\Lambda_{\max}(\Sigma_X)}{\Lambda_{\min}(\Sigma_\epsilon^*)}\right]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}.$$

w.p. at least $1-6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$.

For the second term, first we note that

$$\begin{aligned}\frac{1}{n}\left\|X'E(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)\right\|_\infty &= \frac{1}{n}\max_{\substack{1\le i\le p_1\\1\le j\le p_2}}\left|e_i'X'E(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)e_j\right| \\ &\le \frac{1}{n}\max_i\left\|e_i'X'E\right\|_\infty\max_j\left\|(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)e_j\right\|_1,\end{aligned} \tag{54}$$

where we have $e_i\in\mathbb{R}^{p_1}$ and $e_j\in\mathbb{R}^{p_2}$, and the inequality comes from the fact that $|a'b|\le\|a\|_\infty\|b\|_1$. Note that

$$\max_i\left\|e_i'X'E\right\|_\infty = \|X'E\|_\infty,$$

since $\|e_i'X'E\|_\infty$ gives the largest element (in absolute value) of the ith row of $X'E$, and taking the maximum over all i's gives the largest element of $X'E$ over all entries. And for $\max_j\left\|(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)e_j\right\|_1$, it holds that

$$\max_j\left\|(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)e_j\right\|_1 = \left|\!\left|\!\left|\widehat{\Theta}_\epsilon-\Theta_\epsilon^*\right|\!\right|\!\right|_1 = \left|\!\left|\!\left|\widehat{\Theta}_\epsilon-\Theta_\epsilon^*\right|\!\right|\!\right|_\infty,$$

where $|||A|||_1 := \max_{\|x\|_1=1}\|Ax\|_1$ is the ℓ_1-operator norm, and the last equality follows from the fact that $|||A|||_1 = |||A'|||_\infty$. As a result, (54) can be re-written as:

$$\frac{1}{n}\left\|X'E(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)\right\|_\infty \le \left(\frac{1}{n}\|X'E\|_\infty\right)\left(\left|\!\left|\!\left|\widehat{\Theta}_\epsilon-\Theta_\epsilon^*\right|\!\right|\!\right|_\infty\right). \tag{55}$$

Now, using (51), w.p. at least $1-6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$, we have

$$\frac{1}{n}\left\|X'E\right\|_\infty \le c_2\left[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}},$$

and since $\|\widehat{\Theta}_\epsilon-\Theta_\epsilon^*\|_\infty\le\nu_\Theta$, it directly follows that $\left|\!\left|\!\left|\widehat{\Theta}_\epsilon-\Theta_\epsilon^*\right|\!\right|\!\right|_\infty \le d\nu_\Theta$. Therefore, with probability at least $1-6c_1\exp[-(c_2^2-1)\log(p_1p_2)]$,

$$\frac{1}{n}\left\|X'E(\widehat{\Theta}_\epsilon-\Theta_\epsilon^*)\right\|_\infty \le c_2d\nu_\Theta\left[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2}\sqrt{\frac{\log(p_1p_2)}{n}}. \tag{56}$$

Combine the two terms, we obtain the conclusion in Proposition 2. □

Proof of Proposition 3. First we note the following decomposition:

$$\|\widehat{S} - \Sigma_\epsilon^*\|_\infty \leq \|S - \Sigma_\epsilon\|_\infty + \|\widehat{S} - S\|_\infty := \|W_1\|_\infty + \|W_2\|_\infty,$$

where S is the sample covariance matrix of the true errors E.

For W_1, by Lemma 8 in Ravikumar et al. (2011), for sample size

$$n \geq 512(1+4\sigma_\epsilon^2)^4 \max_i(\Sigma^*_{\epsilon,ii})^4 \log(4p_2^{\tau_2}),$$

the following bound holds w.p. at least $1 - 1/p_2^{\tau_2 - 2}(\tau_2 > 2)$,

$$\|W_1\|_\infty \leq \sqrt{\frac{\log 4 + \tau_2 \log p_2}{c_\epsilon^* n}}, \quad \text{where } c_\epsilon^* = \left[128(1+4\sigma_\epsilon^2)^2 \max_i(\Sigma^*_{\epsilon,ii})^2\right]^{-1}. \tag{57}$$

For W_2, rewrite it as:

$$W_2 = \frac{2}{n} E'X(B^* - \widehat{B}) + (B^* - \widehat{B})' \left(\frac{X'X}{n}\right)(B^* - \widehat{B}). \tag{58}$$

The first term in (58) can be bounded as:

$$\left\|\frac{2}{n}E'X(B^* - \widehat{B})\right\|_\infty \leq 2\left\|\left\|B^* - \widehat{B}\right\|\right\|_1 \left\|\frac{1}{n}X'E\right\|_\infty \leq 2\|\beta^* - \widehat{\beta}\|_1 \cdot \left\|\frac{1}{n}X'E\right\|_\infty. \tag{59}$$

By Lemma 3, with probability at least $1 - 6c_1 \exp[-(c_2^2 - 1)\log(p_1 p_2)]$, the following bound holds:

$$\left\|\frac{2}{n}E'X(B^* - \widehat{B})\right\|_\infty \leq 2c_2\nu_\beta \left[\Lambda_{\max}(\Sigma_X)\Lambda_{\max}(\Sigma_\epsilon^*)\right]^{1/2} \sqrt{\frac{\log(p_1 p_2)}{n}}, \tag{60}$$

with the sample size requirement being $n \gtrsim \log(p_1 p_2)$.

For the second term in (58), we consider the following bound:

$$\begin{aligned} \|(B^* - \widehat{B})' \left(\frac{X'X}{n}\right)(B^* - \widehat{B})\|_\infty &\leq \left\|\left\|B^* - \widehat{B}\right\|\right\|_1 \left\|\left(\frac{X'X}{n}\right)(B^* - \widehat{B})\right\|_\infty \\ &\leq \left\|\left\|B^* - \widehat{B}\right\|\right\|_1^2 \left\|\left(\frac{X'X}{n}\right)\right\|_\infty. \end{aligned} \tag{61}$$

Here, we apply Lemma 8 in Ravikumar et al. (2011) to the design matrix X, for sample size

$$n \geq 512(1+4\sigma_x^2)^4 \max_i(\Sigma_{X,ii})^4 \log(4p_1^{\tau_1}),$$

the following bound holds w.p. at least $1 - 1/p_1^{\tau_1 - 2}(\tau_1 > 2)$,

$$\left\|\left(\frac{X'X}{n}\right) - \Sigma_X\right\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}}, \qquad \text{where } c_X^* = \left[128(1+4\sigma_x^2)^2 \max_i(\Sigma_{X,ii})^2\right]^{-1}. \tag{62}$$

This indicates that with this choice of n, the following bound holds with probability at least $1 - 1/p_1^{\tau_1 - 2}(\tau_1 > 2)$,

$$\left\| \left(\frac{X'X}{n} \right) \right\|_\infty \leq \sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}} + \max_i(\Sigma_{X,ii}).$$

Combine with the bound in (61), with probability at least $1 - 1/p_1^{\tau_1 - 2}(\tau_1 > 2)$, the following bound holds:

$$\|(B^* - \widehat{B})' \left(\frac{X'X}{n} \right) (B^* - \widehat{B})\|_\infty \leq \nu_\beta^2 \left(\sqrt{\frac{\log 4 + \tau_1 \log p_1}{c_X^* n}} + \max_i(\Sigma_{X,ii}) \right). \tag{63}$$

Now combine (59), (60) and (63), we reach the conclusion of Proposition 3, with the leading term in the sample size requirement being $n \gtrsim \log(p_1 p_2)$. □

Proof for Proposition 4. From the structural equations of a multi-layered graph introduced in Section 2.1, and setting $\epsilon^1 := X^1$, we can write

$$\begin{bmatrix} \epsilon^1 \\ \epsilon^2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ -(B^{12})' & I \end{bmatrix} \begin{bmatrix} X^1 \\ X^2 \end{bmatrix}. \tag{64}$$

Define $P = [I, 0; -(B^{12})', I]$. Then, $P\widetilde{X}$ is a centered Gaussian random vector with a block diagonal variance-covariance matrix $diag(\Sigma^1, \Sigma^2)$. Hence, the concentration matrix of $\widetilde{X}$ takes the form

$$\Theta_{\tilde{X}} = \Sigma_{\tilde{X}}^{-1} = \begin{bmatrix} I & -B^{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Theta^1 & 0 \\ 0 & \Theta^2 \end{bmatrix} \begin{bmatrix} I & 0 \\ -(B^{12})' & 0 \end{bmatrix}.$$

This leads to an upper bound

$$\|\Theta_{\tilde{X}}\| \leq \|\Theta^1\| \|\Theta^2\| \|P\|^2.$$

The result then follows by using the matrix norm inequality $\|A\| \leq \sqrt{\|A\|_1 \|A\|_\infty}$ (Golub and Van Loan, 2012), where $\|A\|_1$ and $\|A\|_\infty$ denote the maximum absolute row and column sums of A, and the fact that $\Lambda_{\min}(\Sigma_{\tilde{X}}) = \|\Theta_{\tilde{X}}\|^{-1}$. □

Appendix C. Numerical comparisons between different parametrizations.

In this subsection, we provide some numerical evidence to substantiate the point we made in Section 5, that the two parametrizations are not always equivalent. This is a point also mentioned in the original work on AMP graphs by Andersson et al. (2001), the framework adopted in this paper. The other parametrization which we referred to as the (Ω_{XY}, Ω_Y)-*parametrization* corresponds to the LWF framework (see Andersson et al., 2001, p.34-35). In the presence of sparsity penalization, a specific sparsity pattern for the (B, Θ_ϵ)-*parameterization* may not be recoverable through the (Ω_{XY}, Ω_Y)-*parametrization* and vice versa.

Consider the following two simulation settings, in which the data are generated from the AMP framework ((B, Θ_ϵ)-parameterization) and the LWF framework ((Ω_{XY}, Ω_Y)-*parametrization*) respectively.

- AMP framework. The data are generated according to the model $Y = XB^* + E$, similar to Model A described in Section 4; that is, each entry in B^* is nonzero with probability $5/p_1$, and off-diagonal entries for Θ^*_ϵ are nonzero with probability $5/p_2$. Nonzero entries of B^* and Θ^*_ϵ are generated from $\mathsf{Unif}\left[(-1,-0.5)\cup(0.5,1)\right]$, and diagonals of Θ^*_ϵ are set identical, such that the condition number of Θ^*_ϵ is p_2. Table 11 shows the performance of estimated B using different methods that are designed for different parameterizations: the node-conditional method (mixed MRF) and the proposed method in this study (PML).

Table 11: Performance for $\widehat{B}$ using different methods for different parameterizations

(p_1, p_2, n)	Method	SEN	SPC	MCC
$(30, 60, 100)$	mixed MRF (th)	0.86	0.71	0.45
	PML	0.96	0.99	0.93
$(60, 30, 100)$	mixed MRF (th)	0.96	0.76	0.70
	PML	0.99	0.99	0.93
$(200, 200, 150)$	mixed MRF (th)	0.80	0.99	0.70
	PML	0.99	0.99	0.88

- LWF framework. The data are generated based on the multivariate Gaussian specification:

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{YX} & \Omega_Y \end{pmatrix}^{-1}\right),$$

  Specifically, Ω_X is banded with 1 on the diagonal and 0.2 on the upper and lower first diagonal, Ω_Y is also banded with 1 on the diagonal and 0.3 on the upper and lower first diagonal. Each entry in $\Omega_X Y$ is nonzero with probability $5/p_1$, and the nonzero entries are generated from $\mathsf{Unif}\left[(-1,-0.8)\cup(0.8,1)\right]$. Further, we bump up the diagonal of the joint precision matrix $\left[\begin{smallmatrix} \Omega_X & \Omega_{XY} \\ \Omega'_{XY} & \Omega_Y \end{smallmatrix}\right]$ such that it is positive definite. Table 12 depicts the selection property of the estimated Ω_{XY} using different methods that are designed for different parameterizations.

Table 12: Performance for $\widehat{\Omega}_{XY}$ using different methods for different parameterizations

(p_1, p_2, n)	Method	SEN	SPC	MCC
$(30, 60, 100)$	mixed MRF	0.84	0.88	0.63
	PML-th	0.99	0.52	0.39
$(60, 30, 100)$	mixed MRF	0.847	0.95	0.70
	PML-th	1	0.80	0.52
$(200, 200, 150)$	mixed MRF	0.89	0.93	0.70
	PML-th	1	0.79	0.30

Note that to retrieve stable and meaningful results, for the AMP framework, the estimates using mixed MRF are thresholded at a proper level, and for the LWF framework, the estimates using PML are also thresholded.

It can be seen that the method compatible with the data generation mechanism exhibits superior performance, vis-a-vis its competitor that was designed for another parameterization. Further, the mixed MRF method suffers in terms of both sensititvity and specificity under the AMP parameterization, while the PML method suffers in terms of specificity only under the LWF parameterization.

Appendix D. An example for multi-layered network estimation.

As mentioned at the beginning of Section 2, the proposed methodology is designed for obtaining MLEs for multi-layer Gaussian networks, but the problem breaks down into a sequence of 2-layered estimation problems. Here we give an detailed example to illustrate how our proposed methodology proceeds for a 3-layered network.

Suppose there are p_1, p_2 and p_3 nodes in Layers 1, 2 and 3, respectively. This three-layered network is modeled as follows:

- $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma_X)$, $\boldsymbol{X} \in \mathbb{R}^{p_1}$.
- For $j = 1, \cdots, p_2$: $Y_j = \boldsymbol{X}' B_j^{xy} + \epsilon_j^Y$, $B_j^{xy} \in \mathbb{R}^{p_1}$. $(\epsilon_1^Y \cdots \epsilon_{p_2}^Y)' \sim \mathcal{N}(0, \Sigma_{\epsilon,Y})$.
- For $l = 1, 2, \cdots, p_3$: $Z_l = \boldsymbol{X}' B_l^{xz} + \boldsymbol{Y}' B_l^{yz} + \epsilon_l^Z$, $B_l^{xz} \in \mathbb{R}^{p_1}$ and $B_l^{yz} \in \mathbb{R}^{p_2}$. $(\epsilon_1^Z \cdots \epsilon_{p_3}^Z)' \sim \mathcal{N}(0, \Sigma_{\epsilon,Z})$.

The parameters of interest are : Θ_X, $\Theta_{\epsilon,Y} := \Sigma_{\epsilon,Y}^{-1}$, $\Theta_{\epsilon,Z} := \Sigma_{\epsilon,Z}^{-1}$, which denote the within-layer conditional dependencies, and

$$B_{XY} = \begin{bmatrix} B_1^{xy} & \cdots & B_{p_2}^{xy} \end{bmatrix}, \quad B_{XZ} = \begin{bmatrix} B_1^{xz} & \cdots & B_{p_3}^{xz} \end{bmatrix} \text{ and } B_{YZ} = \begin{bmatrix} B_1^{yz} & \cdots & B_{p_3}^{yz} \end{bmatrix},$$

which encode the across-layer dependencies.

Now given data $X \in \mathbb{R}^{n \times p_1}$, $Y \in \mathbb{R}^{n \times p_2}$ and $Z \in \mathbb{R}^{n \times p_3}$, all centered, the full log-likelihood can be written as:

$$\ell(Z, Y, X) = \ell(Z|Y, X; \Theta_{\epsilon,Z}, B_{YZ}, B_{XZ}) + \ell(Y|X; \Theta_{\epsilon,Y}, B_{XY}) + \ell(X; \Theta_X). \tag{65}$$

The separability of the log-likelihood enables us to ignore the inner structure of the combined layer $\widetilde{X} := (X, Y)$ when trying to estimate the dependencies between Layer 1 and Layer 3, Layer 2 and Layer 3, as well as the conditional dependencies within Layer 3. As a consequence, the optimization problem minimizing the negative log-likelihood can be decomposed into three separate problems, i.e., solving for $\{\Theta_{\epsilon,Z}, B_{XZ}, B_{YZ}\}$, $\{\Theta_{\epsilon,Y}, B_{XY}\}$ and $\{\Theta_X\}$, respectively.

The estimation procedure described in Section 2.2 can thus be carried out in a recursive way in a sense of what follows. To obtain estimates for $\{B_{XZ}, B_{YZ}, \Theta_{\epsilon,Z}\}$, based on the formulation in (2), we solve the following opmization problem:

$$\min_{\substack{\Theta_{\epsilon,Z} \in \mathbb{S}_{++}^{p_3 \times p_3} \\ B_{XZ}, B_{YZ}}} \left\{ \begin{array}{c} -\log\det\Theta_{\epsilon,Z} + \frac{1}{n}\sum_{j=1}^{p_3}\sum_{i=1}^{p_3} \sigma_Z^{ij} (Z_i - XB_i^{xz} - YB_i^{yz})^\top (Z_j - XB_j^{xz} - YB_j^{yz}) \\ +\lambda_n(\|B_{XZ}\|_1 + \|B_{YZ}\|_1) + \rho_n \|\Theta_{\epsilon,Z}\|_{1,\text{off}} \end{array} \right\},$$

which can be solved by treating the combined design matrix $\widetilde{X} = (X, Y)$ as a single super layer and Z as the response layer, then apply each step described in Section 2.2. To obtain estimates for B_{XY} and $\Theta_{\epsilon,Y}$, we can ignore the 3rd layer for now and apply the exact procedure all over again, by treating Y as the response layer and X as the design layer. The estimate for the precision matrix of the bottom layer Θ_X can be obtained by graphical lasso (Friedman et al., 2008) or the nodewise regression (Meinshausen and Bühlmann, 2006).

References

Steen A Andersson, David Madigan, and Michael D Perlman. Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28(1):33–85, 2001.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

Joseph K Bradley, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. Parallel coordinate descent for ℓ_1-regularized loss minimization. *arXiv preprint arXiv:1105.5379*, 2011.

Peter Bühlmann and Sara Van De Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2012.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Mathias Drton and Michael D Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Morten Frydenberg. The chain graph Markov property. *Scandinavian Journal of Statistics*, 17(4):333–353, 1990.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU Press, 2012.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Steffen L Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Steffen L Lauritzen and Nanny Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17 (1):31–57, 1989.

Wonyul Lee and Yufeng Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis*, 111:241–255, 2012.

Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer Science & Business Media, 1998.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Calvin McCarter and Seyoung Kim. On sparse Gaussian chain graph models. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bi Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, pages 1–52, 2012.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Chad Scherrer, Mahantesh Halappanavar, Ambuj Tewari, and David Haglin. Scaling up coordinate descent algorithms for large ℓ_1 regularization problems. *arXiv preprint arXiv:1206.6409*, 2012.

Michael E Sobel. Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450):647–651, 2000.

Kyung-Ah Sohn and Seyoung Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 1081–1089, 2012.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *Seventh IEEE International Conference on Data Mining*, pages 322–331. IEEE, 2007.

Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed graphical models via exponential families. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 1042–1050, 2014.

Jianxin Yin and Hongzhe Li. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630, 2011.

Xiao-Tong Yuan and Tong Zhang. Partial Gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687, 2014.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Tuo Zhao, Xingguo Li, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. *huge: High-dimensional undirected graph estimation*, 2015. URL http://CRAN.R-project.org/package=huge. R package version 1.2.7.