The Annals of Statistics
2014, Vol. 42, No. 2, 532–562
DOI: 10.1214/13-AOS1187

GEMINI: GRAPH ESTIMATION WITH MATRIX VARIATE NORMAL INSTANCES

BY SHUHENG ZHOU[1]

University of Michigan

Undirected graphs can be used to describe matrix variate distributions. In this paper, we develop new methods for estimating the graphical structures and underlying parameters, namely, the row and column covariance and inverse covariance matrices from the matrix variate data. Under sparsity conditions, we show that one is able to recover the graphs and covariance matrices with a single random matrix from the matrix variate normal distribution. Our method extends, with suitable adaptation, to the general setting where replicates are available. We establish consistency and obtain the rates of convergence in the operator and the Frobenius norm. We show that having replicates will allow one to estimate more complicated graphical structures and achieve faster rates of convergence. We provide simulation evidence showing that we can recover graphical structures as well as estimating the precision matrices, as predicted by theory.

1. Introduction. The matrix variate normal model has a long history in psychology and social sciences, and is becoming increasingly popular in biology and genetics, econometric theory, image and signal processing and machine learning in recent years. In this paper, we present a theoretical framework to show that one can estimate the covariance and inverse covariance matrices well using only one matrix from the matrix-variate normal distribution. The motivation for this problem comes from many applications in statistics and machine learning. For example, in microarray studies, a single $f \times m$ data matrix X represents expression levels for m genes on f microarrays; one needs to find out simultaneously the correlations and partial correlations between genes, as well as between microarrays. Another example concerns observations from a spatiotemporal stochastic process which can be described with a matrix normal distribution with a separable covariance matrix $S \otimes T$, where typically, S is called spatial covariance, T is called the temporal covariance and $\otimes$ is the Kronecker product. When the stochastic process is spatial–temporal, some structures can be assumed for one or both of the matrices in the Kronecker product. However, typically one has only one observational matrix.

Received August 2013; revised November 2013.
[1]Supported in part by NSF Grant DMS-13-16731.
MSC2010 subject classifications. Primary 62F12; secondary 62F30.
Key words and phrases. Graphical model selection, covariance estimation, inverse covariance estimation, graphical Lasso, matrix variate normal distribution.

We call the random matrix X which contains f rows and m columns a single data matrix, or one instance from the matrix variate normal distribution. We say that an $f \times m$ random matrix X follows a matrix normal distribution with a separable covariance matrix $\Sigma = A \otimes B$, which we write

$$X_{f\times m} \sim \mathcal{N}_{f,m}(M, A_{m\times m} \otimes B_{f\times f}). \tag{1}$$

This is equivalent to say $\mathrm{vec}\{X\}$ follows a multivariate normal distribution with mean $\mathrm{vec}\{M\}$ and covariance $\Sigma = A \otimes B$. Here, $\mathrm{vec}\{X\}$ is formed by stacking the columns of X into a vector in $\mathbf{R}^{mf}$. Intuitively, A describes the covariance between columns of X while B describes the covariance between rows of X. See [4, 10] for characterization and examples. Note that we can only estimate A and B up to a scaled factor, as $A\eta \otimes \frac{1}{\eta}B = A \otimes B$ for any $\eta > 0$, and hence this will be our goal of the paper, and precisely what we mean, when we say we are interested in estimating covariances A and B.

Undirected graphical models are often used to describe high dimensional distributions. We will use such descriptions in the present work to encode structural assumptions on the inverse of the row and column covariance matrices. A common structural assumption is that the inverse covariance matrices, also known as the precision matrices, are sparse, which means that the number of nonzero entries (sparsity levels) in one or both of them are bounded. Under sparsity assumptions, a popular approach to obtain a sparse estimate for the precision matrix is given by the ℓ_1-norm regularized maximum-likelihood function, also known as the GLasso [2, 9, 18, 26]. All these methods and their analysis assume that one is given independent samples and the estimation of A or B alone is their primary goal, as they all assume that X has either independent rows or independent columns. A direct application of the GLasso estimator to estimate $A \otimes B$ with no regard for its separable structure will lead to computational misery, as the cost will become prohibitive for f, m in the order of 100. Various work [5, 14, 23] focused on algorithms and convergence properties on estimating Σ using a large number of samples $X(1), \ldots, X(n)$. A mean-restricted matrix-variate normal model was considered in [1], where they proposed placing additive penalties on estimated inverse covariance matrices in order to obtain regularized row and column covariance/precision matrices. Other recent work with an iterative approach for solving the graphical model selection problem in the context of matrix variate normal distribution include [11, 13, 19, 24, 28]. None of these works was able to show convergence in the operator norm which works in case $n = 1$ and $f, m \to \infty$ as in our work.

1.1. *Our approach and contributions.* In this work, we take a penalized approach and show from a theoretical point of view, the advantages of estimating covariance matrices A, B and the graphs corresponding to their inverses simultaneously albeit via separable optimization functions. The key observation and starting point of our work is: although A and B are not identifiable given the separable

representation as in (1), their correlation matrices $\rho(A)$ and $\rho(B)$, and the graphical structures corresponding to their inverses are identifiable, and can indeed be efficiently estimated for a given matrix $X \sim \mathcal{N}_{f,m}(0, A \otimes B)$. Moreover, $\rho(A)^{-1}$ and $\rho(B)^{-1}$ encode the same structural information as A^{-1} and B^{-1} do, in the sense that they share an identical set of nonzero edges. Therefore, we propose estimating the overall $\Sigma = A \otimes B$ and its inverse by (i) first estimating correlation matrices $\rho(A)$ and $\rho(B)$ (and their inverses) using a pair of ℓ_1-norm penalized estimators for an instance $X \sim \mathcal{N}_{f,m}(0, A \otimes B)$, (ii) and then combining these two estimators with the estimated variances to form an estimator for Σ.

Toward this end, we develop Gemini (Graph estimation with matrix variate normal instances), a new method for estimating graphical structures, and the underlying parameters A and B. We will answer the following question: how sparse does A^{-1} or B^{-1} need to be in order for us to obtain statistical convergence rates for estimating A and B (up to a scaled factor) simultaneously with one data matrix X? Our estimators extend, with suitable adaptation, to the general setting where n replicates of X are available. Our method is computational efficient. The dominating cost involves in estimating $\rho(A)^{-1}$ and $\rho(B)^{-1}$: the total cost is in the order of $O(f^3 + m^3)$ for sparse graphs or $O(f^4 + m^4)$ for general graphs.

In summary, we make the following theoretical contributions: (i) consistency and rates of convergence in the operator and the Frobenius norm of the covariance matrices and their inverses, (ii) large deviation results for the sample correlation estimators which we propose for estimating both the row and column correlation and covariance matrices given a single matrix or multiple replicates of the matrix-normal data, (iii) conditions that guarantee simultaneous estimation of the graphs for both rows and columns. We note that with all other parameters hold invariant, the rates of convergence in all metrics in (i) and (ii) in estimating A, B (and their inverses) will be proportional to $n^{-1/2}$. To the best of our knowledge, these are the first such results on the matrix-variate normal distributions in the high dimensional setting for finite and small sample instances, by which we mean $n < \log \max(m, f)$. We provide simulation evidence and a real data example showing that we can recover graphical structures as well as estimate the precision matrices effectively.

There is no known closed-form solution for the maximum of the likelihood function for the matrix-variate normal distribution. There has been a line of work in the literature which suggested using iterative algorithms, namely, the Flip-Flop methods to estimate the covariance matrix with the Kronecker structure; see, for example, [5, 14, 23] and references therein. In the present work, building upon the baseline Gemini estimators, we also propose a three-step penalized variant of the Flip-Flop algorithms in Section 5. We show that under an additional condition, this approach yields certain improvements upon the baseline Gemini estimators.

The rest of the paper is organized as follows. In Section 2, we will define our model and the method. Section 3 presents the main theoretical results in this paper on estimating $A \otimes B$, as well as discussions on our method and results; moreover,

we review the related work to place our work in context. Section 4 provides large deviation inequalities for the sample correlation coefficients in approximating the underlying parameters of $\rho(A)$ and $\rho(B)$; more general bounds of this nature are derived in Section 13 in the supplementary material [29]. Convergence rates in the Frobenius norm for estimating the inverse correlation matrices are also derived. We propose a Noniterative Penalized Flip-Flop algorithm and study its convergence properties in Sections 5 and 6. Section 7 shows our numerical results. We conclude in Section 8. We place all technical proofs in the supplementary material [29].

1.2. *Notation*. For a matrix $A = (a_{ij})_{1\le i,j\le m}$, let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ denote the entry-wise max norm; let $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ denote the matrix ℓ_1 norm. The Frobenius norm is given by $\|A\|_F^2 = \sum_i \sum_j a_{ij}^2$. Let $|A|$ denote the determinant and $\operatorname{tr}(A)$ be the trace of A. Let $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ be the largest and smallest eigenvalues, and $\kappa(A)$ be the condition number for matrix A. The operator or ℓ_2 norm $\|A\|_2^2$ is given by $\varphi_{\max}(AA^T)$. Let $r(A) = \|A\|_F^2/\|A\|_2^2$ denote the stable rank for matrix A. We write $|\cdot|_1$ for the ℓ_1 norm of a matrix vectorized, that is, $|A|_1 = \|\operatorname{vec}\{A\}\|_1 = \sum_i \sum_j |a_{ij}|$. Let $|A|_{1,\text{off}} = \sum_{i\neq j} |A_{ij}|$, and $|A|_{0,\text{off}}$ be the number of nonzero nondiagonal entries in the matrix. We use A^{-T} to denote $(A^{-1})^T$. We write $\operatorname{diag}(A)$ for a diagonal matrix with the same diagonal as A. For a symmetric matrix A, let $\Upsilon(A) = (\upsilon_{ij})$ where $\upsilon_{ij} = \mathbb{I}(a_{ij} \neq 0)$, where $\mathbb{I}(\cdot)$ is the indicator function. Let I be the identity matrix. We let C be a constant which may change from line to line. For two numbers a, b, $a \wedge b := \min(a, b)$, and $a \vee b := \max(a, b)$. We write $a \asymp b$ if $ca \le b \le Ca$ for some positive absolute constants c, C which are independent of n, f, m or sparsity parameters.

2. The model and the method. In the matrix variate normal setting, we aim to estimate the row and column covariance (correlation) matrices, from which we can obtain an estimate for Σ. The problem of covariance estimation in the context of matrix variate normal distribution is intimately connected to the problem of graphical model selection, where the graphs corresponding to the column and the row vectors are determined by the sparsity patterns (or the zeros) of B^{-1} and A^{-1}, respectively. Graph estimation in this work means precisely the estimation of the zeros, as well as the nonzero entries in A^{-1} and B^{-1}. We formulate such correspondence precisely in Section 2.1. We define our estimators in Sections 2.2 and 2.3.

2.1. *Problem definition*: *The matrix normal graphical model*. We show in Figure 1 the data matrix X and its column vectors: $x^1, x^2, \ldots, x^k, \ldots, x^m$, and row vectors $y^1, y^2, \ldots, y^f$.

This notation is followed throughout the rest of the paper. First recall the following definition concerning the classical Gaussian graphical model for a random vector.

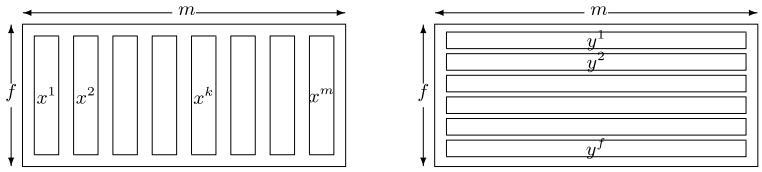

FIG. 1. *Column and row vectors of matrix X, where $X \sim \mathcal{N}_{f,m}(0, A \otimes B)$. Let $A = (a_{ij})$ and $B = (b_{ij})$. The normalized column vectors $x^1/\sqrt{a_{11}}, \ldots, x^m/\sqrt{a_{mm}}$, where $a_{ii} > 0$, follow a multivariate normal distribution $\mathcal{N}_f(0, B)$ while normalized row vectors $y^1/\sqrt{b_{11}}, \ldots, y^f/\sqrt{b_{ff}}$, where $b_{jj} > 0$, follow $\mathcal{N}_m(0, A)$.*

DEFINITION 2.1. Let $V = (V_1, \ldots, V_f)^T$ be a random Gaussian vector, which we represent by an undirected graph $G = (\mathcal{V}, F)$. The vertex set $\mathcal{V} := \{1, \ldots, f\}$ has one vertex for each component of the vector V. The edge set F consists of pairs (j, k) that are joined by an edge. If V_j is independent of V_k given the other variables, then $(j, k) \notin F$.

Now let $\mathcal{V} = \{1, \ldots, f\}$ be an index set which enumerates rows of X according to a fixed order. For all $i = 1, \ldots, m$, we assign to each variable of a column vector x^i exactly one element of the set $\mathcal{V}$ by a rule of correspondence $g : x^i \to \mathcal{V}$ such that $g(x^i_j) = j$, $j = 1, \ldots, f$. The graphs $G_i(\mathcal{V}, F)$ constructed for each random column vector $x^i, i = 1, \ldots, m$ according to Definition 2.1 will share an identical edge set F, because the normalized column vectors $x^1/\sqrt{a_{11}}, \ldots, x^m/\sqrt{a_{mm}}$ follow the same multivariate normal distribution $\mathcal{N}_f(0, B)$. Hence, graphs $G_1, \ldots, G_m$ are isomorphic and we write $G_i \simeq G_j, \forall i, j$. Due to the isomorphism, we use $G(\mathcal{V}, F)$ to represent the family of graphs $G_1, \ldots, G_m$. Hence, a pair (ℓ, k) which is absent in F encodes conditional independence between the ℓth row and the kth row give all other rows. Similarly, let $\Gamma = \{1, \ldots, m\}$ be the index set which enumerates columns of X according to a fixed order. We use $H(\Gamma, E)$ to represent the family of graphs $H_1, \ldots, H_f$, where H_i is constructed for row vector y^i, and $H_i \simeq H_j, \forall i, j$. Now $H(\Gamma, E)$ is a graph with adjacency matrix $\Upsilon(H) = \Upsilon(A^{-1})$ as edges in E encode nonzeros in A^{-1}. And $G(\mathcal{V}, F)$ is a graph with adjacency matrix $\Upsilon(G) = \Upsilon(B^{-1})$. The Kronecker product, $H \otimes G$, is defined as the graph with adjacency matrix $\Upsilon(H) \otimes \Upsilon(G)$ [22], where clearly missing edges correspond to zeros in the inverse covariance $A^{-1} \otimes B^{-1}$, and $H \otimes G$ represents the graph of the p-variate Gaussian random vector $\mathrm{vec}\{X\}$, where $p = mf$. In the present work, we aim to estimate $\Upsilon(H)$ and $\Upsilon(G)$ separately. Estimating their Kronecker product directly following the classical p-variate Gaussian graphical modeling approach will be costly in terms of both computation and the sample requirements.

2.2. *The Gemini estimators.* We start with the one-matrix case. We note that between A and B, the dimension of one matrix is the same as the number of samples available for estimating parameters in the other matrix in case $n = 1$. Therefore, m and f are allowed to grow so long as they grow with respect to each other. The first hurdle we need to deal with, besides the simultaneous row and column correlations, is the fact that between the two covariance matrices A and B (as well as their inverses), the one with the higher dimension, which contains more canonical parameters, is always left with a smaller number of correlated samples in order to achieve its inference tasks. The remedy comes from the following observation. Although ambient dimension f, m cannot be both bounded by the other unless $f = m$, the sparsity over nondiagonal entries of each precision matrix can be assumed to be bounded by the ambient dimension of the other.

Under such sparsity assumptions, we first provide a pair of separable regularized estimators for the correlation matrices $\rho(A) = (a_{ij}/\sqrt{a_{ii}a_{jj}})$ and $\rho(B) = (b_{ij}/\sqrt{b_{ii}b_{jj}})$,

$$\widehat{A}_\rho = \operatorname*{arg\,min}_{A_\rho \succ 0} \{\operatorname{tr}(\widehat{\Gamma}(A)A_\rho^{-1}) + \log|A_\rho| + \lambda_B |A_\rho^{-1}|_{1,\text{off}}\}, \tag{2a}$$

$$\widehat{B}_\rho = \operatorname*{arg\,min}_{B_\rho \succ 0} \{\operatorname{tr}(\widehat{\Gamma}(B)B_\rho^{-1}) + \log|B_\rho| + \lambda_A |B_\rho^{-1}|_{1,\text{off}}\}, \tag{2b}$$

where the input are a pair of sample correlation matrices $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$

$$\widehat{\Gamma}_{ij}(A) := \frac{\langle x^i, x^j\rangle}{\|x^i\|_2\|x^j\|_2} \quad \text{and} \quad \widehat{\Gamma}_{ij}(B) := \frac{\langle y^i, y^j\rangle}{\|y^i\|_2\|y^j\|_2}, \tag{3}$$

and the ℓ_1 penalties are imposed on the off-diagonal entries of the inverse correlation estimates. Note that the population parameters A and B can be written as

$$A \otimes B := \big(W_1\rho(A)W_1\big) \otimes \big(W_2\rho(B)W_2\big)/\big(\operatorname{tr}(A)\operatorname{tr}(B)\big),$$

where $W_1/\sqrt{\operatorname{tr}(B)} = \operatorname{diag}(\sqrt{a_{11}}, \ldots, \sqrt{a_{mm}})$ and $W_2/\sqrt{\operatorname{tr}(A)} = \operatorname{diag}(\sqrt{b_{11}}, \ldots, \sqrt{b_{ff}})$. In order to get an estimate for $A \otimes B$, we multiply each of the two regularized estimators $\widehat{A}_\rho$ and $\widehat{B}_\rho$ by an estimated weight matrix $\widehat{W}_1$ or $\widehat{W}_2$, respectively,

$$\widehat{W}_1 = \operatorname{diag}\big(\|x^1\|_2, \|x^2\|_2, \ldots, \|x^m\|_2\big) = \operatorname{diag}\big(X^TX\big)^{1/2},$$
$$\widehat{W}_2 = \operatorname{diag}\big(\|y^1\|_2, \|y^2\|_2, \ldots, \|y^f\|_2\big) = \operatorname{diag}\big(XX^T\big)^{1/2}.$$

Up to a multiplicative factor $\operatorname{tr}(B)$ and $\operatorname{tr}(A)$, $\widehat{W}_1^2$ and $\widehat{W}_2^2$ will provide an estimate for $\operatorname{diag}(A)$ and $\operatorname{diag}(B)$, respectively; hence, to estimate $A \otimes B$, we compute the Kronecker product of our weighted estimators,

$$\widehat{A \otimes B} := (\widehat{W}_1\widehat{A}_\rho\widehat{W}_1) \otimes (\widehat{W}_2\widehat{B}_\rho\widehat{W}_2)/\|X\|_F^2$$

while adjusting the unknown multiplicative factors $\operatorname{tr}(B)\operatorname{tr}(A)$ by $\|X\|_F^2$.

Clearly, the sample correlation estimators (3) are obtained from the gram matrices X^TX and XX^T of the column and row vectors as follows:

$$\text{(4)}\qquad \widehat{\Gamma}(A) = \widehat{W}_1^{-1}(X^TX)\widehat{W}_1^{-1} \quad \text{and} \quad \widehat{\Gamma}(B) = \widehat{W}_2^{-1}(XX^T)\widehat{W}_2^{-1} \qquad \text{where}$$

$$\text{(5)}\quad \mathbb{E}X^TX = \sum_{i=1}^{f} \mathbb{E}y^i \otimes y^i = \operatorname{tr}(B)A, \qquad \mathbb{E}XX^T = \sum_{i=1}^{m} \mathbb{E}x^i \otimes x^i = \operatorname{tr}(A)B,$$

and the multiplicative factors $\operatorname{tr}(B)$ and $\operatorname{tr}(A)$ become irrelevant due to cancellation. By setting the gradient equations of objective functions (2a) and (2b) to zero, we see that the pair of estimators satisfy $\operatorname{diag}(\widehat{A}_\rho) = \operatorname{diag}(\widehat{\Gamma}(A))$ and $\operatorname{diag}(\widehat{B}_\rho) = \operatorname{diag}(\widehat{\Gamma}(B))$ as desired. Moreover, the penalty parameters λ_B and λ_A are chosen to dominate the maximum of entry-wise errors for estimating $\rho(A)$ and $\rho(B)$ with $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as characterized in Theorem 4.4 (cf. Remark 4.2 and the comments which follow immediately).

2.3. *Gemini for replicates of X.* We now adapt the Gemini estimators as defined in Section 2.2 to the general setting where we have multiple replicates of X. Suppose that we have n independently and identically distributed matrices $X(1), \dots, X(n) \sim \mathcal{N}_{f,m}(0, A \otimes B)$. For each t, we denote by

$$\text{(6)}\qquad X(t) = \begin{bmatrix} x(t)^1 & x(t)^2 & \cdots & x(t)^m \end{bmatrix} = \begin{bmatrix} y(t)^1 & y(t)^2 & \cdots & y(t)^f \end{bmatrix}^T$$

the matrix $X_{f\times m}(t)$ with $x(t)^1, \dots, x(t)^m \in \mathbf{R}^f$ being its columns vectors and $y^1(t), \dots, y^f(t)$ being its row vectors.

First, we update our sample correlation matrices, which we will plug in (2a) and (2b) to obtain the penalized correlation estimators $\widehat{A}_\rho$ and $\widehat{B}_\rho$.

$$\text{(7a)}\qquad \widehat{\Gamma}_{ij}(A) := \frac{\sum_{t=1}^n \langle x(t)^i, x(t)^j \rangle}{\sqrt{\sum_{t=1}^n \|x(t)^i\|_2^2}\sqrt{\sum_{t=1}^n \|x(t)^j\|_2^2}},$$

$$\text{(7b)}\qquad \widehat{\Gamma}_{ij}(B) := \frac{\sum_{t=1}^n \langle y(t)^i, y(t)^j \rangle}{\sqrt{\sum_{t=1}^n \|y(t)^i\|_2^2}\sqrt{\sum_{t=1}^n \|y(t)^j\|_2{}^2}}.$$

Next, we update the weight matrices $\widehat{W}_1$ and $\widehat{W}_2$ as follows:

$$\text{(8a)}\qquad \widehat{W}_1 = \operatorname{diag}\left(\sqrt{\frac{1}{n}\sum_{t=1}^n \|x(t)^i\|_2^2}, i = 1, \dots, m\right),$$

$$\text{(8b)}\qquad \widehat{W}_2 = \operatorname{diag}\left(\sqrt{\frac{1}{n}\sum_{t=1}^n \|y(t)^j\|_2^2}, j = 1, \dots, f\right).$$

We can then construct an estimator for $A \otimes B$ as before,

$$\text{(9)}\qquad \widehat{A \otimes B} := (\widehat{W}_1 \widehat{A}_\rho \widehat{W}_1) \otimes (\widehat{W}_2 \widehat{B}_\rho \widehat{W}_2) \Big/ \left(\frac{1}{n}\sum_{t=1}^n \|X(t)\|_F^2\right).$$

We will show in Theorem 4.1 large deviation bounds for estimating the correlation coefficients in $\rho(A)$ and $\rho(B)$ with entries in sample correlation $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ constructed above, which are crucial in proving the convergence rates for estimating $A \otimes B$ and its inverse with $\widehat{A \otimes B}$ and $\widehat{A \otimes B}^{-1}$.

3. Theoretical results. In this section, we present in Theorem 3.1 and Theorem 3.2 the convergence rates for estimating the row and column covariance matrices and their inverses with respect to the operator norm and the Frobenius norm, respectively. Our analysis is nonasymptotic in nature; however, we first formulate our results from an asymptotic point of view for simplicity. To do so, we consider an array of matrix variate normal data

$$X(1), \ldots, X(n) \text{ i.i.d.} \sim \mathcal{N}_{f,m}(0, A_0 \otimes B_0), \qquad n = 1, 2, \ldots, \tag{10}$$

where f, m may change with n. Let $|A_0^{-1}|_{0,\text{off}}$ and $|B_0^{-1}|_{0,\text{off}}$ be the number of nonzero nondiagonal entries in the inverse covariance matrices A_0^{-1} and B_0^{-1}, respectively. Recall, for matrix A, $r(A) = \|A\|_F^2/\|A\|_2^2$ and $\kappa(A)$ denote its stable rank and condition number, respectively.

We make the following assumptions.

(A1) The dimensions f and m are allowed to grow with respect to each other, and

$$|A_0^{-1}|_{0,\text{off}} = o\big(nf/\log(m \vee f)\big) \qquad (f, m \to \infty) \quad \text{and}$$
$$|B_0^{-1}|_{0,\text{off}} = o\big(nm/\log(m \vee f)\big) \qquad (f, m \to \infty).$$

(A2) The eigenvalues $\varphi_i(A_0), \varphi_j(B_0), \forall i, j$ of the positive definite covariance matrices A_0 and B_0 are bounded away from 0 and $+\infty$.

Moreover, we assume that the stable ranks $r(A_0)$ and $r(B_0)$ satisfy $r(A_0)$, $r(B_0) \geq 4\log(m \vee f)/n$, which holds trivially if $n \geq 4\log(m \vee f)$; otherwise, it is sufficient to require that $(m \vee f) = o(\exp(\frac{f}{\kappa^2(B_0)} \wedge \frac{m}{\kappa^2(A_0)}))$.

We now state the main results of this paper, which are new to the best of our knowledge. These bounds are stated in terms of the relative errors.

THEOREM 3.1. *Consider data generating random matrices in* (10). *Suppose that* (A1) *and* (A2) *hold, and the penalty parameters are chosen to be*

$$\lambda_A = \lambda_{A_0} \asymp \frac{\|A_0\|_F}{\operatorname{tr}(A_0)} \frac{\log^{1/2}(m \vee f)}{\sqrt{n}} \asymp \frac{\log^{1/2}(m \vee f)}{\sqrt{mn}} \to 0 \quad \textit{and}$$
$$\lambda_B = \lambda_{B_0} \asymp \frac{\|B_0\|_F}{\operatorname{tr}(B_0)} \frac{\log^{1/2}(m \vee f)}{\sqrt{n}} \asymp \frac{\log^{1/2}(m \vee f)}{\sqrt{fn}} \to 0.$$

Then with probability at least $1 - \frac{3}{(m\vee f)^2}$, *for* $\widehat{A\otimes B}$ *as defined in* (9),

$$\|\widehat{A\otimes B} - A_0\otimes B_0\|_2 \le \|A_0\|_2\|B_0\|_2\delta \quad \textit{and}$$
$$\|\widehat{A\otimes B}^{-1} - A_0^{-1}\otimes B_0^{-1}\|_2 \le \|A_0^{-1}\|_2\|B_0^{-1}\|_2\delta'$$
$$\textit{where } \delta,\delta' = O\big(\lambda_{A_0}\sqrt{|B_0^{-1}|_{0,\mathrm{off}}\vee 1} + \lambda_{B_0}\sqrt{|A_0^{-1}|_{0,\mathrm{off}}\vee 1}\big) = o(1).$$

THEOREM 3.2. *Consider data generating random matrices as in* (10). *Let* λ_{A_0} *and* λ_{B_0} *be chosen as in Theorem* 3.1. *Let* $\widehat{A\otimes B}$ *be as defined in* (9). *Under* (A1) *and* (A2),

$$\|\widehat{A\otimes B} - A_0\otimes B_0\|_F \le \delta\|A_0\|_F\|B_0\|_F$$
$$\textit{where } \delta = O\big(\lambda_{A_0}\sqrt{|B_0^{-1}|_{0,\mathrm{off}}\vee f}/\sqrt{f} + \lambda_{B_0}\sqrt{|A_0^{-1}|_{0,\mathrm{off}}\vee m}/\sqrt{m}\big) = o(1). \tag{11}$$

In particular, suppose (i) $1 \le n \le \log(m\vee f)$ *or* (ii) $|A_0^{-1}|_{0,\mathrm{off}} = O(m)$ *and* $|B_0^{-1}|_{0,\mathrm{off}} = O(f)$. *Then* $\delta = O(\lambda_{A_0} + \lambda_{B_0})$. *The same conclusions hold for the inverse estimate, with* δ *being bounded in the same order as in* (11).

The two summands in δ and δ' in Theorems 3.1 and 3.2 correspond to the rates of convergence in the operator and the Frobenius norm for estimating the row and column covariance matrices A_0, B_0, up to a scale factor, respectively. These rates are derived in Sections 4.2 and 10. We prove Theorems 3.1 and 3.2 in Section 11 in the supplementary material [29], where we examine the rate of (11) in case $n \ge 4\log(m\vee f)$ in Remark 11.3, and show the absolute error bounds in Theorems 11.1 and 11.2. There we also make the connection between the one-matrix and the multiple-matrix cases in order to understand the rates for $n > 1$.

3.1. *Discussion.* To put our discussions on the rates of convergence for covariance estimation in context, we first present an example from the classical multivariate analysis. Consider the case where we are given a single sample from the matrix variate normal distribution with $B_0 = I$, and the dimensions f, m increase to infinity, while the aspect ratio $f/m \to \mathsf{const} > 1$. The classical multivariate analysis focuses on estimating A_0 using data matrix X; the simplest way to estimate A_0 is to compute the sample covariance

$$\widetilde{A}_f = \frac{1}{f}X^TX = \frac{1}{f}\sum_{i=1}^{f} x_i\otimes x_i \qquad \text{where } x_1,\dots,x_m \text{ i.i.d.} \sim \mathcal{N}_m(0, A_0).$$

The problem here is to determine the minimal number of independent rows we need so that the sample covariance matrix $\widetilde{A}_f$ approximates A "well" in the operator norm. This concerns the classical "Bai–Yin law" in random matrix theory regarding the Wishart random matrix $\widetilde{A}_f$, which says that the spectrum of

$\widetilde{A}_f$ is almost surely contained in the interval $[a^2/f + o(1), b^2/f + o(1)]$ where $a = (\sqrt{f} - \sqrt{m})_+$ and $b = \sqrt{f} + \sqrt{m}$ in case $A_0 = I$. For general covariance matrix A_0, the following holds with high probability (cf. [21]):

$$\|\widetilde{A}_f - A_0\|_2 \le \big(2\sqrt{m/f} + (m/f) + o(1)\big)\|A_0\|_2. \tag{12}$$

While such results provide a satisfactory answer to the covariance estimation problem in the regime $f \ge m$ for general multivariate normal distributions, it remains challenging to answer the following questions: (a) how to estimate the covariance matrix which has the larger dimension of the two? That is, how can we approximate A_0 well in the operator norm when $f < m$? (b) how to estimate both A_0 and B_0 given both correlated rows and columns?

Our answer to the first question is to use the penalized methods. The operator norm bound in Theorem 3.1 illustrates the point that the combination of sparsity and spectral assumptions as in (A1), (A2) and ℓ_1-regularization ensures convergence on estimation of the covariance and precision matrices, even though their ambient dimensions may greatly exceed the given sample sizes. In particular, the ambient dimensions which appear in the numerator in (12) are replaced with the sparsity parameters (cf. Theorem 3.1):

$$\delta, \delta' = O\big(\log^{1/2}(m \vee f)\big(\sqrt{|A_0^{-1}|_{0,\mathrm{off}} \vee 1}/\sqrt{f} + \sqrt{|B_0^{-1}|_{0,\mathrm{off}} \vee 1}/\sqrt{m}\big)\big) = o(1),$$

which holds for $n = 1$ with high probability under (A1) and (A2), as (A1) implies that, up to a logarithmic factor, the number of nonzero off-diagonal entries in A_0^{-1} or B_0^{-1} must be bounded by the dimension of the other matrix. We will relax such sparsity conditions in Section 3.2.

To answer the second question, first recall that in the current setting, (4) suggests that $\widetilde{A}_f = X^T X/f$ and $\widetilde{B}_m = XX^T/m$ are good starting points for us to construct estimators for A_0, B_0, $\rho(A_0)$, and $\rho(B_0)$ despite the presence of dependence along the other dimension. The relationships between the row and column correlations of X are known to complicate the solution to the related problem of testing the hypothesis that microarrays are independent of each other given possibly correlated genes [6]. Taking these complex relationships into consideration, we construct covariance and correlation estimators based on $\widetilde{A}_f$ and $\widetilde{B}_m$, as well as the pair of functions in (29); we will develop concentration bounds which illustrate their interactions throughout the rest of the paper.

3.2. *Relaxing the sparsity assumptions.* While the rates in Theorem 3.2 are essentially tight, we can tighten those in Theorem 3.1 under an alternative set of sparsity conditions. In particular, relaxation of (A1) is feasible when we consider a restricted uniformity class of inverse covariance matrices whose matrix ℓ_1 norm is bounded by a parameter M: for $0 \le q < 1$,

$$\mathcal{U}_q\big(d_0(m), M\big) = \left\{\Theta = (\theta_{ij})_{1\le i,j\le m} : \max_i \sum_{j=1}^{m} |\theta_{ij}|^q \le d_0(m), \|\Theta\|_1 \le M, \Theta \succ 0\right\}.$$

It is to be understood that $d_0(m)$, M are positive numbers bounded away from 0 which are allowed to grow with m, f, n. We focus on the case when $q = 0$ and consider positive definite matrices with row/column sparsity constraints, upon which we obtain a more refined result on the ℓ_2 error bounds in Theorem 3.3. First, we replace (A1) with (A1′), where $\Theta_0 = \rho(A_0)^{-1}$ and $\Phi_0 = \rho(B_0)^{-1}$ denote the inverse correlation matrices.

(A1′) Suppose that $\Theta_0 \in \mathcal{U}_0(d_0(m), M)$ and $\Phi_0 \in \mathcal{U}_0(d_0(f), K)$, where $d_0(m)$, $d_0(f)$, M and K are positive and bounded away from 0. The dimensions f and m are allowed to grow with respect to each other while the number of nonzero elements in each row or column of Θ_0 and Φ_0 must be bounded by $d_0(m)$ and $d_0(f)$, respectively: as $f, m \to \infty$

$$d_0(m)\|\Theta_0\|_1^2 = o\biggl(\frac{\sqrt{nf}}{\log^{1/2}(m \vee f)}\biggr) \quad \text{and} \quad d_0(f)\|\Phi_0\|_1^2 = o\biggl(\frac{\sqrt{nm}}{\log^{1/2}(m \vee f)}\biggr).$$

We present Theorem 3.3 using the CLIME estimators [3], which are obtained by first solving the following optimization functions:

$$\widetilde{\Theta} = \arg\min_{\Theta \in \mathbf{R}^{m \times m}} |\Theta|_1 \qquad \text{subject to } \bigl\|\widehat{\Gamma}(A_0)\Theta - I\bigr\|_{\max} \le \lambda_M, \tag{13}$$

$$\widetilde{\Phi} = \arg\min_{\Phi \in \mathbf{R}^{f \times f}} |\Phi|_1 \qquad \text{subject t o} \bigl\|\widehat{\Gamma}(B_0)\Phi - I\bigr\|_{\max} \le \lambda_K \tag{14}$$

for λ_M and λ_K to be specified in Theorem 3.3; then a symmetrization step selects each entry for the estimators $\widehat{\Theta}^{\mathrm{CL}} = (\widehat{\theta}_{ij})$ and $\widehat{\Phi}^{\mathrm{CL}} = (\widehat{\phi}_{ij})$, as follows:

$$\widehat{\Theta}_{\mathrm{CL}} = (\widehat{\theta}_{ij}) \qquad \text{s.t. } \widehat{\theta}_{ij} = \widetilde{\theta}_{ij}\mathbb{I}\bigl(|\widetilde{\theta}_{ij}| \le |\widetilde{\theta}_{ji}|\bigr) + \widetilde{\theta}_{ji}\mathbb{I}\bigl(|\widetilde{\theta}_{ij}| > |\widetilde{\theta}_{ji}|\bigr), \tag{15}$$

$$\widehat{\Phi}_{\mathrm{CL}} = (\widehat{\phi}_{ij}) \qquad \text{s.t. } \widehat{\phi}_{ij} = \widetilde{\phi}_{ij}\mathbb{I}\bigl(|\widetilde{\phi}_{ij}| \le |\widetilde{\phi}_{ji}|\bigr) + \widetilde{\phi}_{ji}\mathbb{I}\bigl(|\widetilde{\phi}_{ij}| > |\widetilde{\phi}_{ji}|\bigr). \tag{16}$$

THEOREM 3.3. *Consider data generating random matrices as in* (10). *Suppose that* (A1′) *and* (A2) *hold. Let* λ_{A_0}, λ_{B_0} *be as in Theorem* 3.1, *and*

$$\lambda_M \asymp \|\Theta_0\|_1 \lambda_{B_0} \quad \textit{and} \quad \lambda_K \asymp \|\Phi_0\|_1 \lambda_{A_0}$$

for λ_M, λ_K *as in* (13) *and* (14). *Let* $\widehat{\Theta}_{\mathrm{CL}}$ *and* $\widehat{\Phi}_{\mathrm{CL}}$ *be as in* (15) *and* (16).

Then with $\mathbb{P}(\mathcal{X}_0) \ge 1 - \frac{3}{(m \vee f)^2}$, $\widehat{\Theta}_{\mathrm{CL}}$ *and* $\widehat{\Phi}_{\mathrm{CL}}$ *are positive definite*; *and for* $\widehat{A \otimes B}$ *as defined in* (9), *where* $\widehat{A}_\rho := \widehat{\Theta}_{\mathrm{CL}}^{-1}$ *and* $\widehat{B}_\rho := \widehat{\Phi}_{\mathrm{CL}}^{-1}$,

$$\|\widehat{A \otimes B} - A_0 \otimes B_0\|_2 \le \|A_0\|_2 \|B_0\|_2 \delta \quad \textit{and}$$

$$\|\widehat{A \otimes B}^{-1} - A_0^{-1} \otimes B_0^{-1}\|_2 \le \|B_0^{-1}\|_2 \|A_0^{-1}\|_2 \delta',$$

$$\textit{where } \delta, \delta' = O\bigl(\lambda_{A_0} d_0(f)\|\Phi_0\|_1^2 + \lambda_{B_0} d_0(m)\|\Theta_0\|_1^2\bigr) = o(1).$$

Proof of Theorem 3.3 appears in Section 16.

REMARKS. Suppose that f, m are sufficiently large, and $f < m$. We focus our discussions on Θ_0. Denote the maximum node degree by $\deg(\Theta_0) := \max_i \sum_j \mathbb{I}(\theta_{ij} \neq 0)$. We note that (A1′) imposes the bounded node degree constraint in that: $\deg(\Theta_0) = o(\sqrt{nf}/\log^{1/2}(m \vee f))$, while in (A1) a hub node alone can have up to $o(nf/\log(m \vee f))$ adjacent nodes. Suppose that (A2) holds, and $n < \log(m \vee f)$. In this case, (A1′) relaxes (A1) in the sense that it allows $\deg(\Theta_0) = \Omega(1)$, and hence $|A_0^{-1}|_{0,\text{off}} = \Omega(m)$, while (A1) does not. Thus, the graphs considered in (A1) can be those which contain a single or multiple disjoint components with some singleton nodes, while those in (A1′) are allowed to be fully connected graphs.

Theorem 3.3 improves upon Theorem 3.1 when M, K are slowly growing with respect to m, f, n, while $d_0(m)$ and $d_0(f)$ are of lower order relative to the total number edges in each graph. However, this improvement requires that we replace (2a) and (2b) with the CLIME or graphical Dantzig-type estimators [3, 25], for which we are able to obtain faster rates of convergence in the operator norm under (A1′) and (A2) in estimating each covariance/correlation matrix. The replacement is due to the lack of convergence bounds on the ℓ_2 errors which are tighter than those presented in Theorem 4.5, for the graphical Lasso estimators; as a consequence, the two summands in δ and δ' in Theorem 3.1 were obtained using the rates of convergence in the Frobenius norm, rather than the operator norm as we do in Theorem 3.3, for estimating the general (but sparse) inverse correlation matrices. To the best of our knowledge, comparable convergence bounds on the operator norm for the graphical Lasso-type estimators are available only under an *irrepresentability condition* as developed in [17]. We can indeed invoke their results in the present setting to relax the sparsity constraint on in (A1), and to prove faster rates of convergence in the operator norm in view of Theorem 4.1.

3.3. *Related work.* Algorithmic and theoretical properties of the graphical Lasso or Lasso-type estimators have been well studied in the Gaussian graphical model setting; see, for example, [2, 7, 9, 12, 15–18, 26, 30]. Under sparsity and neighborhood stability conditions, the work by [15] showed that the graph with p nodes can be estimated efficiently using the nodewise penalized regression approach using a very small sample size n in comparison to the maximum node degree and the ambient dimension p. The work of [3, 25, 31], using variants of the approach in [15], showed convergence rates in the operator and the Frobenius norm in estimating the precision matrix in case $p > n$, where independent samples are always assumed. It will be interesting to consider replacing the ℓ_1 penalties with the SCAD-type penalties or using the adaptive Lasso-type penalties as in [7, 12]. These approaches will reduce certain bias in the penalized estimators; see, for example, discussions in [8, 32]. The recent work of [1] focuses on missing value imputation, rather than estimation of the graphs or the underlying parameters. When f, m diverge as $n \to \infty$, the rates in [24] are significantly

slower than the corresponding ones in the present work. Following essentially the same methods as in [1], the same convergence rate as in (11) on estimating the covariance $\Sigma = A_0 \otimes B_0$ in the Frobenius norm is obtained in [13, 19], in case $|A_0^{-1}|_{0,\text{off}} = O(m)$, and $|B_0^{-1}|_{0,\text{off}} = O(f)$; however, this rate is obtained with the additional requirement that the number of replicates of X must be at least on the order of $n \geq \Omega((\frac{f}{m} \vee \frac{m}{f}) \log \max(f, m, n))$. These results exclude the case for $n = 1$ or for $n < \log(m \vee f)$, which is the main focus of the present paper.

4. Estimation of the correlation coefficients. In this section, we elaborate on two key technical results, namely, the concentration bounds for sample correlation estimates and the convergence bounds for the penalized inverse correlation estimates.

4.1. *Concentration bounds for sample correlations.* We now show the concentration bounds for estimating the parameters in $\rho(A_0)$ and $\rho(B_0)$. Theorem 4.1 covers the small sample settings, where the number of replications n are upper bounded by $\log(m \vee f)$, where $m \vee f := \max(m, f)$. We believe these are the first of such results to the best of our knowledge. For completeness, we also state the bounds when $n > \log(m \vee f)$ is large.

Let K be the ψ_2 norm of ξ for $\xi \sim \mathcal{N}(0, 1)$ defined as

$$K := \|\xi\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \big(\mathbb{E}|\xi|^p\big)^{1/p}; \qquad \text{thus, } K \text{ is the smallest } K_2, \tag{17}$$

$$\text{which satisfies } \big(\mathbb{E}|\xi|^p\big)^{1/p} \leq K_2 \sqrt{p} \qquad \forall p \geq 1; \text{ see [21]}. \tag{18}$$

THEOREM 4.1. *Consider data generating random matrices as in* (10). *Let C be some absolute constant to be defined in* (57),

$$\begin{aligned} &\tau_0 = 2C K^2 \log^{1/2}(m \vee f)/\sqrt{n} \qquad \textit{where } K \textit{ is defined as in } (17), \\ &\alpha_n := \|A_0\|_F \tau_0 / \operatorname{tr}(A_0) \quad \textit{and} \quad \beta_n := \|B_0\|_F \tau_0 / \operatorname{tr}(B_0). \end{aligned} \tag{19}$$

Let $m \vee f \geq 2$. Then with probability at least $1 - \frac{3}{(m \vee f)^2}$, for $\alpha_n, \beta_n < 1/3$, and $\widehat{\Gamma}(A_0)$ and $\widehat{\Gamma}(B_0)$ as in (7a) *and* (7b),

$$\forall i \neq j \qquad \big|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)\big| \leq \frac{\alpha_n}{1 - \alpha_n} + \big|\rho_{ij}(B_0)\big| \frac{\alpha_n}{1 - \alpha_n} \leq 3\alpha_n,$$

$$\forall i \neq j \qquad \big|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)\big| \leq \frac{\beta_n}{1 - \beta_n} + \big|\rho_{ij}(A_0)\big| \frac{\beta_n}{1 - \beta_n} \leq 3\beta_n$$

and

$$\left| \frac{1}{n} \sum_{t=1}^{n} \big\|X(t)\big\|_F^2 - \operatorname{tr}(A_0) \operatorname{tr}(B_0) \right| \leq \operatorname{tr}(A_0) \operatorname{tr}(B_0)(\alpha_n \wedge \beta_n). \tag{20}$$

REMARK 4.2. Note that under (A1) and (A2), we have $\alpha_n, \beta_n \to 0$ as $m, f \to \infty$, where

$$\alpha_n \asymp C_A \frac{\log^{1/2}(m \vee f)}{\sqrt{mn}} \quad \text{and} \quad \beta_n \asymp C_B \frac{\log^{1/2}(m \vee f)}{\sqrt{fn}}$$

$$\text{where } C_A := \frac{\sqrt{m}\|A_0\|_F}{\operatorname{tr}(A_0)} = \frac{\sqrt{m}\sqrt{\operatorname{tr}(A_0 A_0)}}{\operatorname{tr}(A_0)} \tag{21}$$

$$\text{and } C_B := \frac{\sqrt{f}\|B_0\|_F}{\operatorname{tr}(B_0)} = \frac{\sqrt{f}\sqrt{\operatorname{tr}(B_0 B_0)}}{\operatorname{tr}(B_0)}$$

are treated to be constants throughout this paper under the bounded spectrum assumptions in (A2). Their magnitudes reflect how eigenvalues of each component covariance matrix vary across its entire spectrum, and how much they affect the estimation of the other matrix.

The penalty parameters in Theorem 3.1 and 3.2, are chosen to dominate the dominate the maximum of entry-wise errors for estimating $\rho(A)$ and $\rho(B)$ with $\widehat{\Gamma}(A)$ and $\widehat{\Gamma}(B)$ as characterized in Theorem 4.1:

$$\lambda_{A_0} \asymp C_A \log^{1/2}(m \vee f)/\sqrt{mn} \quad \text{and} \quad \lambda_{B_0} \asymp C_B \log^{1/2}(m \vee f)/\sqrt{fn}.$$

The notation λ_{A_0} and λ_{B_0} thus reflect their dependencies on the eigenspectrum of A_0 and B_0, which in turn affects the rate of convergence in the Frobenius norm in estimating $\rho(B_0)$ and $\rho(A_0)$ with the penalized estimators. The following large deviation bounds in Lemma 4.3 are the key results in proving Theorem 4.1. We write it explicitly to denote by $\mathcal{X}_0$ the event that all large deviation inequalities as stated in Lemma 4.3 hold.

LEMMA 4.3. *Suppose that* (A2) *holds. Denote by* $\mathcal{X}_0$ *the event that the following inequalities hold simultaneously for* α_n, β_n *as defined in* (19)

$$\forall i, j \qquad \left| \frac{1}{n} \sum_{t=1}^{n} \langle y(t)^i, y(t)^j \rangle \Big/ \left(\operatorname{tr}(A_0) \sqrt{b_{ii} b_{jj}} \right) - \rho_{ij}(B_0) \right| \le \alpha_n,$$

$$\forall i, j \qquad \left| \frac{1}{n} \sum_{t=1}^{n} \langle x(t)^i, x(t)^j \rangle \Big/ \left(\operatorname{tr}(B_0) \sqrt{a_{ii} a_{jj}} \right) - \rho_{ij}(A_0) \right| \le \beta_n.$$

Suppose $m \vee f \ge 2$. *Then* $\mathbb{P}(\mathcal{X}_0) \ge 1 - \frac{3}{(m \vee f)^2}$.

The proofs for Theorem 4.1 and Lemma 4.3 appear in Section 13. We restate the first two inequalities Theorem 4.1 in case $n = 1$ in Theorem 4.4. Let C_A, C_B be as in Remark 4.2.

THEOREM 4.4. *Suppose* $m \vee f \geq 2$ *and* (A2) *holds. Let* $\widehat{\Gamma}(A_0)$ *and* $\widehat{\Gamma}(B_0)$ *be as in* (3). *Then with probability at least* $1 - \frac{3}{(m\vee f)^2}$, *for all* $i \neq j$

$$\big|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)\big| \leq 2CK^2C_A\big(1 + \big|\rho_{ij}(B_0)\big|\big)\frac{\log^{1/2}(m \vee f)}{\sqrt{m}}\big(1 + o(1)\big),$$

$$\big|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)\big| \leq 2CK^2C_B\big(1 + \big|\rho_{ij}(A_0)\big|\big)\frac{\log^{1/2}(m \vee f)}{\sqrt{f}}\big(1 + o(1)\big).$$

REMARKS. We next compare the concentration bounds for the matrix normal distribution as in Theorems 4.4 and 4.1 with those for the multivariate Gaussian. First suppose that $f \leq m$ and B_0 is an $f \times f$ identity matrix. That is, we are given independent rows in X. Then the rate of convergence for estimating $\rho_{ij}(A_0)$ with (3) is bounded in [30, 31] (cf. Lemma 13 and equation (43) in [31]) as follows: With probability at least $1 - 1/(f \vee m)^2$,

$$\big\|\widehat{\Gamma}(A_0) - \rho(A_0)\big\|_{\max} < 3C_3\sqrt{\log(m \vee f)/f} \tag{22}$$

for f large enough, so long as $m < e^{f/4C_3^2}$ for some constant $C_3 > 4\sqrt{5/3}$. Now suppose that B_0 follows an AR(1) model with parameter ρ, then the RHS of (22) is necessarily replaced with a slower rate of

$$\beta \asymp C_B\log^{1/2}(m \vee f)/\sqrt{f}. \tag{23}$$

We note that this rate as well as the rate of $\beta_n \asymp C_B \log^{1/2}(m \vee f)/\sqrt{nf}$ are at the same order as the classical rate of (22) as the effective sample size for estimating A_0 is nf (cf. Remark 11.3). However, both β and β_n are affected by the measure of C_B, which will increases as the parameter ρ increases; we illustrate this behavior in our numerical results in Section 7.3. We are able to remove the dependency on C_B in (23) in Section 6 under additional sparsity conditions.

4.2. *Bounds on estimating the inverse correlation matrices.* In this section, we show explicit nonasymptotic convergence rates in the Frobenius norm for estimating $\rho(A_0)$, $\rho(B_0)$ and their inverses in Theorem 4.5. In Section 14, we present in Corollary 14.1 a bound on the off-diagonal vectorized ℓ_1 norm on the error matrices for estimating $\Theta_0 = \rho(A_0)^{-1}$ and $\Phi_0 = \rho(B_0)^{-1}$, which may be of independent interests.

We say that event $\mathcal{T}(A_0)$ holds for sample correlation matrix $\widehat{\Gamma}(A_0)$ for some parameter $\delta_{n,f} \to 0$, if for all j, $\widehat{\Gamma}_{jj}(A_0) = \rho_{jj}(A_0) = 1$ and

$$\max_{j,k,j\neq k}\big|\widehat{\Gamma}_{jk}(A_0) - \rho_{jk}(A_0)\big| \leq \delta_{n,f}, \tag{24}$$

and the event $\mathcal{T}(B_0)$ holds for sample correlation matrix $\widehat{\Gamma}(B_0)$ for some parameter $\delta_{n,m} \to 0$, if for all j, $\widehat{\Gamma}_{jj}(B_0) = \rho_{jj}(B_0) = 1$ and

$$\max_{j,k,j\neq k}\big|\widehat{\Gamma}_{jk}(B_0) - \rho_{jk}(B_0)\big| \leq \delta_{n,m}. \tag{25}$$

THEOREM 4.5. *Suppose that* (A2) *holds. Let* $\widehat{A}_\rho$ *and* $\widehat{B}_\rho$ *be the unique minimizers defined by* (2a) *and* (2b) *with sample correlation matrices* $\widehat{\Gamma}(A_0)$ *and* $\widehat{\Gamma}(B_0)$ *as their input. Suppose that event* $\mathcal{T}(A_0)$ *holds for* $\widehat{\Gamma}(A_0)$ *for some* $\delta_{n,f}$ *and event* $\mathcal{T}(B_0)$ *holds for* $\widehat{\Gamma}(B_0)$ *for some* $\delta_{n,m}$, *such that*

$$\begin{aligned} &\delta_{n,f}\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1} = o(1) \quad \text{and} \quad \delta_{n,m}\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1} = o(1), \\ &\text{set for some } 0 < \epsilon, \varepsilon < 1, \lambda_B = \delta_{n,f}/\varepsilon \text{ and } \lambda_A = \delta_{n,m}/\epsilon. \end{aligned} \tag{26}$$

Then on event $\mathcal{T}(A_0) \cap \mathcal{T}(B_0)$, *we have for* $9 < C < 18$

$$\|\widehat{A}_\rho - \rho(A_0)\|_2 \le \|\widehat{A}_\rho - \rho(A_0)\|_F \le C\kappa\big(\rho(A_0)\big)^2 \lambda_B \sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1},$$

$$\|\widehat{B}_\rho - \rho(B_0)\|_2 \le \|\widehat{B}_\rho - \rho(B_0)\|_F \le C\kappa\big(\rho(B_0)\big)^2 \lambda_A \sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}$$

and

$$\|\widehat{A}_\rho^{-1} - \rho(A_0)^{-1}\|_2 \le \|\widehat{A}_\rho^{-1} - \rho(A_0)^{-1}\|_F < \frac{C\lambda_B\sqrt{|A_0^{-1}|_{0,\text{off}} \vee 1}}{2\varphi^2_{\min}(\rho(A_0))}, \tag{27}$$

$$\|\widehat{B}_\rho^{-1} - \rho(B_0)^{-1}\|_2 \le \|\widehat{B}_\rho^{-1} - \rho(B_0)^{-1}\|_F \le \frac{C\lambda_A\sqrt{|B_0^{-1}|_{0,\text{off}} \vee 1}}{2\varphi^2_{\min}(\rho(B_0))}. \tag{28}$$

Variants of Theorem 4.5 was shown in [18] in the context of Gaussian graphical models; our proof follows similar arguments, and hence is omitted. Lemma 4.6 justifies the choices of the penalty parameters λ_{A_0} and λ_{B_0}.

LEMMA 4.6. *Let* $\alpha_n, \beta_n < 1/3$ *be as defined in Theorem* 4.1. *Let*

$$\delta_{n,f} = \frac{2\beta_n}{1-\beta_n} = O\left(C_B \frac{\log^{1/2}(m \vee f)}{\sqrt{nf}}\right) \quad \text{and}$$

$$\delta_{n,m} = \frac{2\alpha_n}{1-\alpha_n} = O\left(C_A \frac{\log^{1/2}(m \vee f)}{\sqrt{nm}}\right).$$

Then, event $\mathcal{T}(A_0) \cap \mathcal{T}(B_0)$ *holds on* $\mathcal{X}_0$ *for the sample correlation matrices as defined in* (7a) *and* (7b), *respectively.*

By Theorem 4.1, *we have* $\mathbb{P}(\mathcal{T}(A_0) \cap \mathcal{T}(B_0)) \ge 1 - \frac{3}{(m \vee f)^2}$.

5. Variations on a theme. It is curious whether or not one can improve upon the Gemini sample covariance/correlation estimators using the Flip-Flop methods. Essentially the Flip-Flop methods [5, 14, 23] couple the estimation for A_0 and B_0 by feeding the current estimate for either of the two into the likelihood function (or the penalized variants to be defined) in order to optimize it with respect to the

other. Upon initialization of A in (29) with an identity matrix, they obtain the MLE for A_0 and B_0 by solving the following two equations alternately and iteratively:

$$\widetilde{B}(A) = \frac{1}{nm}\sum_{t=1}^{n} X(t)A^{-1}X(t)^T, \qquad \widetilde{A}(B) = \frac{1}{nf}\sum_{t=1}^{n} X(t)^T B^{-1}X(t) \tag{29}$$

such that the corresponding output $\widetilde{B}$, or $\widetilde{A}$ becomes the input as B, or A to the RHS of equations in (29); this process repeats until certain convergence criteria are reached. The baseline Gemini method, where we simultaneously optimize a pair of convex functions (2a) and (2b), can be seen as a single-step approximation of a penalized version of (29), where we simply set both B and A on the RHS of equations in (29) to be the identity matrix.

We now introduce a natural variation of the Gemini estimators as given by the Noniterative Penalized Flip-Flop (NiPFF) algorithm, where we construct more sophisticated covariance and correlation estimators based on the pair of functions in (29).

Noniterative Penalized Flip-Flop algorithm:

1. Assume $f \le m$. Initialize $A_{\text{init}} = I$. Compute $\widehat{\Gamma}(B_0)$ based on (7b) as before, and compute $\widehat{B}_\rho$ using GLasso (2b) with the penalty parameter λ_{A_0} to be chosen (cf. Lemma 6.1). Let $B_1 = \widetilde{W}_2\widehat{B}_\rho\widetilde{W}_2/m$.
2. Now compute the sample covariance $\widetilde{A}(B_1)$ using (29) and the sample correlation matrix $\widehat{\Gamma}(A_0)$ with

$$\widehat{\Gamma}(A_0) = \widetilde{W}_1^{-1}\widetilde{A}(B_1)\widetilde{W}_1^{-1} \qquad \text{where } \widetilde{W}_1 = \operatorname{diag}\bigl(\widetilde{A}(B_1)\bigr)^{1/2}. \tag{30}$$

   Obtain an estimate $\widehat{A}_\rho(B_1)$ using GLasso (2a) with $\widehat{\Gamma}(A_0)$ in (30) as its input, where $\lambda_B = \lambda_{B_1}$ is to be specified (cf. Remark 6.3).
   Let $A_1 = \widehat{A}_* = \widetilde{W}_1\widehat{A}_\rho(B_1)\widetilde{W}_1$.
3. Compute sample covariance matrix $\widetilde{B}(A_1)$ using (29), and the sample correlation matrix $\widehat{\Gamma}(B_0)$ with

$$\widehat{\Gamma}(B_0) = \widetilde{W}_2^{-1}\widetilde{B}(A_1)\widetilde{W}_2^{-1} \qquad \text{where } \widetilde{W}_2 := \operatorname{diag}\bigl(\widetilde{B}(A_1)\bigr)^{1/2}. \tag{31}$$

   Obtain an estimate $\widehat{B}_\rho(A_1)$ using (2b), with $\widehat{\Gamma}(B_0)$ in (31) as its input, where $\lambda_A = \lambda_{A_1}$ is to be specified (cf. Theorem 6.4 and Remark 17.8).
   Let $\widehat{B}_* = \widetilde{W}_2\widehat{B}_\rho(A_1)\widetilde{W}_2$.

6. Analysis for the penalized Flip-Flop algorithm. We illustrate the interactions between the row-wise and column-wise correlations and covariances via the large deviation bounds to be described in this section. To make our discussion concrete, suppose we aim to estimate $A_* = (a_{*,ij}) = mA_0/\operatorname{tr}(A_0)$ and $B_* = (b_{*,ij}) = B_0\operatorname{tr}(A_0)/m$ instead of A_0 and B_0. Note that A_* has been normalized to have $\operatorname{tr}(A_*) = m$ for identifiability. Let

$$\lambda_{f,n} = 2CK^2\frac{\log^{1/2}(m\vee f)}{\sqrt{fn}} \quad \text{and} \quad \lambda_{m,n} = 2CK^2\frac{\log^{1/2}(m\vee f)}{\sqrt{mn}}, \tag{32}$$

where C is as in (19) and K as in (17). In analyzing the Flip-Flop algorithm, we make the following additional assumption.

(A3) The inverse correlation matrices have bounded $|\rho(A_0)^{-1}|_1$ and $|\rho(B_0)^{-1}|_1$:

$$|\rho(A_0)^{-1}|_1 \asymp m \quad \text{and} \quad |\rho(B_0)^{-1}|_1 \asymp f.$$

First, we bound the entry-wise errors for the sample covariance and correlation matrices as defined in step 2. We note that the conclusions of Lemma 6.1 and Theorem 6.2 continue to hold even if ε is chosen outside of the interval $(0, 2/3]$, so long it is bounded away from 0 and 1.

LEMMA 6.1. *Suppose* $(m \vee f) = o(\exp(m \wedge f))$. *Suppose that* (A1), (A2) *and* (A3) *hold. Let* $\widehat{B}_\rho$ *and* B_1 *be obtained as in step* 1, *where we choose*

$$\lambda_{A_0} = \frac{2\alpha}{\varepsilon(1-\alpha)} \geq \frac{3\alpha}{1-\alpha} \qquad \text{for } \alpha = C_A \lambda_{m,n} \text{ where } C_A = \|A_0\|_F \sqrt{m}/\mathrm{tr}(A_0)$$

and $0 < \varepsilon < 2/3$. *Then on event* $\mathcal{A}_1$, *for* $\widetilde{A}(B_1)$ *as defined in* (29)

$$|(\widetilde{A}(B_1) - A_*)_{ij}| \leq \sqrt{a_{*,ii} a_{*,jj}}\lambda_{f,n}(1 + o(1)) + |a_{*,ij}|\widetilde{\mu}, \tag{33}$$

$$\text{where } \widetilde{\mu} = \lambda_{A_0}|\widehat{B}_\rho^{-1}|_{1,\mathrm{off}}/f + \frac{\alpha}{1-\alpha}|\widehat{B}_\rho^{-1}|_1/f \leq \mu \tag{34}$$

$$\text{for } \mu = \lambda_{A_0}|\rho(B_0)^{-1}|_{1,\mathrm{off}}/f + \frac{\alpha}{(1-\alpha)}|\rho(B_0)^{-1}|_1/f + o(\lambda_{A_0}). \tag{35}$$

Moreover, we have for some constant $d \leq 8$, $\mathbb{P}(\mathcal{A}_1) \geq 1 - \frac{d}{(m \vee f)^2}$.

THEOREM 6.2. *Suppose all conditions in Lemma* 6.1 *hold. Let* $\widehat{\Gamma}(A_0)$ *be as defined in* (30). *Then on event* $\mathcal{A}_1$, *for* $\widetilde{\eta} := \lambda_{f,n}(1 + o(1)) + \widetilde{\mu}$, *where* $\widetilde{\mu}$ *is as defined in* (34), $\forall i \neq j$

$$\begin{aligned} &|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)| \\ &\quad \leq (1 + o(1))\lambda_{f,n}(1 + |\rho_{ij}(A_0)|) + \frac{2|\rho_{ij}(A_0)|\widetilde{\mu}}{1 - \widetilde{\eta}} \end{aligned} \tag{36}$$

$$\leq \frac{2\eta}{1-\eta} \qquad \text{where } \eta = \lambda_{f,n}(1 + o(1)) + \mu \text{ for } \mu \text{ as in } (35). \tag{37}$$

REMARK 6.3. On event $\mathcal{A}_1$, the random quantities $\widetilde{\mu}$ and $\widetilde{\eta}$ are upper bounded by μ (35) and η (37), respectively, which can be rewritten as follows.

Define $C_f := |\rho(B_0)^{-1}|_1/f + \frac{2}{\varepsilon}|\rho(B_0)^{-1}|_{1,\mathrm{off}}/f$ so that

$$\mu = \frac{\alpha}{(1-\alpha)}(C_f + o(1)) \quad \text{and} \quad \eta = \left(\lambda_{f,n} + \frac{\alpha}{(1-\alpha)}C_f\right)(1 + o(1)),$$

which suggests that we set the penalty in step 2 in the order of η,

$$\lambda_{B_1} \asymp 2\eta/(1-\eta) \asymp \lambda_{f,n} + C_f\alpha/(1-\alpha)C_f \asymp \lambda_{f,n} + \lambda_{m,n}.$$

Clearly $C_f \asymp 1$ under (A3). Indeed, throughout this paper, we assume

$$\lambda_{B_1} = \frac{2\widetilde{\eta}}{\varepsilon_1(1-\widetilde{\eta})} \qquad \text{where } 0 < \varepsilon_1 < 1. \tag{38}$$

We compute the rates of convergence in the operator and the Frobenius norm for estimating A_* with $\widehat{A}_*$ in step 2 in Corollary 17.2 in Section 17.1. The rates we obtain in Corollary 17.2 correspond to exactly those in Corollary 10.1 for the baseline Gemini estimator, with slightly better leading constants.

Next, we bound the entry-wise errors for the sample correlation matrix as defined in step 3 in Theorem 6.4. The corresponding result for sample covariance is stated in Lemma 17.5.

THEOREM 6.4. *Suppose* $m \vee f = o(\exp(m \wedge f))$. *Suppose that* (A1), (A2), *and* (A3) *hold. Let* $\widehat{\Gamma}(B_0)$ *be as defined in* (31). *Let* $\zeta = \lambda_{m,n}(1+o(1)) + \xi$, *where* ξ *is as defined in* (41). *Then on event* $\mathcal{A}_1 \cap \mathcal{E}_2$, $\forall i \neq j$,

$$\begin{aligned} &\big|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)\big| \\ &\quad\le \frac{\lambda_{m,n}(1+o(1))}{1-\zeta} + \big|\rho_{ij}(B_0)\big|\frac{\zeta+\xi}{1-\zeta} \end{aligned} \tag{39}$$

$$\le \big(\lambda_{m,n}\big(1+\big|\rho_{ij}(B_0)\big|\big) + 2\big|\rho_{ij}(B_0)\big|\xi\big)\big(1+o(1)\big) \tag{40}$$

$$\text{for } \xi = \lambda_{B_1}\big|\rho(A_0)^{-1}\big|_{1,\mathrm{off}}/m + \frac{\eta}{1-\eta}\big|\rho(A_0)^{-1}\big|_1/m + o(\lambda_{B_1}). \tag{41}$$

Moreover, we have for some constant $d \le 10$, $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{E}_2) \ge 1 - \frac{d}{(m\vee f)^2}$.

6.1. *Discussion.* Throughout this discussion, the $O(\cdot)$ notation hides a constant no larger than $1 + o(1)$. We first compare the bound in (36) with that of Theorem 4.1, where on event $\mathcal{X}_0$, for $\widehat{\Gamma}(A_0)$ as defined in (7a),

$$\forall i \neq j \qquad \big|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)\big| = O\big(C_B\lambda_{f,n}\big(1+\big|\rho_{ij}(A_0)\big|\big)\big), \tag{42}$$

where $C_B = \|B_0\|_F\sqrt{f}/\mathrm{tr}(B_0)$. On the other hand, the influence of $\lambda_{A_0} \asymp \frac{2\alpha}{1-\alpha}$ on the entry-wise error for estimating $\rho_{ij}(A_0)$ in (36) is regulated through both C_f, which is a bounded constant under (A3) (see Remark 6.3), as well as the magnitude of $\rho_{ij}(A_0)$ itself; to see this, by (36), $\forall i \neq j$,

$$\big|\widehat{\Gamma}_{ij}(A_0) - \rho_{ij}(A_0)\big| = O\big(\lambda_{f,n}\big(1+\big|\rho_{ij}(A_0)\big|\big) + 2\big|\rho_{ij}(A_0)\big|C_fC_A\lambda_{m,n}\big).$$

This rate is in the same order as that in (42). However, when $\lambda_{m,n} \ll \lambda_{f,n}$, the second term is of smaller order compared to the first term. In this case, the upper

bound in (36) is dominated by the first term on the RHS, and one can perhaps obtain a slightly better bound with Theorem 6.2, as the leading term no longer depends on the constant C_B as displayed in (42).

We next compare the bound in (39) with that of Theorem 4.1. Before we proceed, we first define the following parameter:

$$C_m = |\rho(A_0)^{-1}|_1/m + \frac{2}{\varepsilon_1}|\rho(A_0)^{-1}|_{1,\text{off}}/m \qquad \text{so that}$$

$$\xi \le \frac{\eta}{1-\eta}\big(C_m + o(1)\big) \quad \text{and} \quad \zeta \le \left(\lambda_{m,n} + \frac{\eta}{1-\eta}C_m\right)\big(1+o(1)\big),$$

where $0 < \varepsilon_1 < 1$ is the same as in (38). Hence, we have on $\mathcal{A}_1 \cap \mathcal{E}_2$, by (39),

$$|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)| = O\left(\lambda_{m,n}\big(1 + |\rho_{ij}(B_0)|\big) + |\rho_{ij}(B_0)|C_m\frac{2\eta}{1-\eta}\right),$$

where $2\eta/(1-\eta) \asymp \lambda_{m,n} + \lambda_{f,n}$. Clearly, the influence of $\lambda_{B_1} \asymp \frac{2\eta}{1-\eta}$ on the entry-wise error for estimating $\rho_{ij}(B_0)$ is regulated through the quantity C_m which is a constant under (A3), as well as the magnitude of $\rho_{ij}(B_0)$.

We note that when $m \asymp f$, these rates are in the same order of $O(\lambda_{m,n}(1 + |\rho_{ij}(B_0)|))$ as those in Theorem 4.1 on event $\mathcal{X}_0$. Moreover, for pairs of (i, j) where $i \neq j$, such that $|\rho_{ij}(B_0)|$ is small, one can perhaps obtain a slightly better bound with Theorem 6.4, as the first (leading) term which involves $\lambda_{m,n}$ no longer depends on the constant $C_A \ge 1$ as needed in (21). In summary, for the following two cases, we expect that the sample correlation estimate $\widehat{\Gamma}(B_0)$ which we obtain in step 3 improves upon the initial estimate in step 1:

1. For all $i \neq j$, $\rho_{ij}(B_0)$ is bounded in magnitudes; for example, when $|\rho_{ij}(B_0)| = O(\sqrt{f/m})$, then $\zeta' \asymp \lambda_{m,n}$. In particular, for $\rho(B_0) = I$,

$$\forall i \neq j \qquad \big|\widehat{\Gamma}_{ij}(B_0) - \rho_{ij}(B_0)\big| \le 2\lambda_{m,n}\big(1 + o(1)\big).$$

Hence, the error in estimating A_0 is propagated into the estimate of $\rho_{ij}(B_0)$ only when $\rho_{ij}(B_0) \neq 0$.

2. When m and f are close to each other in that the ratio $m/f \to \mathsf{const} > 1$, and simultaneously, C_m, C_f, and $|\rho_{ij}(B_0)|$ are small for all $i \neq j$; then $2\zeta' = 2(\lambda_{m,n} + \max_{i\neq j}|\rho_{ij}(B_0)|\xi) \asymp \lambda_{m,n} + \lambda_{f,n}$ provides a tight upper bound for the RHS of (40).

Suppose that $m \gg f$. Then the original estimator in (7b) could be much better for pairs of (i, j) with a large $|\rho_{ij}(B_0)|$. As for such pairs, the second term is of larger order than the first term in (40). A refined analysis on the GLasso given the estimates in Theorem 6.4 is left as future work.

7. Numerical results. We demonstrate the effectiveness of the Gemini method as well as the Noniterative Penalized Flip-Flop method, which we refer to as the FF method, with simulated data. We also show an example of applying Gemini to a real data set, the EEG data, obtained from UCI Machine learning repository [20] in Section 7.4. For a penalty parameter $\lambda \geq 0$, the GLasso estimator is given by

$$\operatorname{glasso}(\widehat{\Gamma}, \lambda) = \underset{\Theta \succ 0}{\arg\min}\big(\operatorname{tr}(\widehat{\Gamma}\Theta) - \log|\Theta| + \lambda|\Theta|_{1,\text{off}}\big),$$

where $\widehat{\Gamma}$ is a sample correlation matrix. We use the R-package `glasso` [9] to compute the GLasso solution. For the two estimation methods, we have various tuning parameters, namely λ, ν for the baseline Gemini estimators, and ϕ, υ for the FF method. In our simulation study, we look at three different models from which A and B will be chosen. Let $\Omega = A^{-1} = (\omega_{ij})$ and $\Pi = B^{-1} = (\pi_{ij})$. Let E denote edges in Ω, and F denote edges in Π. We choose A from one of these two models:

- AR(1) model. In this model, the covariance matrix is of the form $A = \{\rho^{|i-j|}\}_{i,j}$. The graph corresponding to Ω is a chain.
- Star-Block model. In this model the covariance matrix is block-diagonal with equal-sized blocks whose inverses correspond to star structured graphs, where $A_{ii} = 1$, for all i. We have 20 subgraphs, where in each subgraph, 8 nodes are connected to a central hub node with no other connections. The rest of the nodes in the graph are singletons. Covariance matrix for each block S in A is generated as in [17]: $S_{ij} = \rho = 0.5$ if $(i, j) \in E$ and $S_{ij} = \rho^2$ otherwise.

For Π, we use the random concentration matrix model in [30]. The graph is generated according to a type of Erdős–Rényi random graph model. Initially, we set $\Pi = 0.25I_{f\times f}$, where $f = 80$. Then we randomly select d edges and update Π as follows: for each new edge (i, j), a weight $w > 0$ is chosen uniformly at random from $[w_{\min}, w_{\max}]$ where $w_{\max} > w_{\min} > 0$; we subtract w from π_{ij} and π_{ji}, and increase π_{ii} and π_{jj} by w. This keeps Π positive definite. For both models of A, we have $A_* = A\frac{m}{\operatorname{tr}(A)} = A = \rho(A)$. Let $\Omega_* = \frac{\operatorname{tr}(A)}{m}\Omega$ and $\Pi_* = \frac{m}{\operatorname{tr}(A)}\Pi$. Thus, we have $\Omega_* = \Omega$ and $\Pi_* = \Pi$ for all combinations of A and B in this section.

7.1. *Regularization paths and cross-validation.* We illustrate the behaviors of the Gemini estimators for each model combination of A, B with $m = 400$ and $f = 80$ over the full regularization paths. To evaluate consistency, we use relative errors in the operator and the Frobenius norm. For model selection consistency, we use false positive and false negative rates and Matthews correlation coefficient (MCC) as defined in Table 1. For each pair of covariance matrices, we do the following. First, we generate A and B, where A is $m \times m$ and B is $f \times f$. Let $A^{1/2}$ and $B^{1/2}$ be the unique square root of matrix A and B, respectively. Let T and T' be a set of values in $(0, 0.5]$. Now, repeat the following steps 100 times:

TABLE 1
Metrics for evaluating $\widehat{E}(\lambda)$

Metric	Definition
False positives (FPs)	# of incorrectly selected edges in $\widehat{E}(\lambda)$: $\lvert\widehat{E}(\lambda) \setminus E\rvert$
False negatives (FNs)	# of edges in E that are not selected in $\widehat{E}(\lambda)$: $\lvert E \setminus \widehat{E}(\lambda)\rvert$
True positives (TPs)	# of correctly selected edges: $\lvert\widehat{E}(\lambda) \cap E\rvert$
True negatives (TNs)	# of zeros in $\widehat{E}(\lambda)$ that are also zero in E
False positive rate (FPR)	$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN}) = \mathrm{FP}/(\binom{m}{2} - \lvert E\rvert)$
False negative rate (FNR)	$\mathrm{FNR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN}) = \mathrm{FN}/\lvert E\rvert$
MCC	$\mathrm{TP} \times \mathrm{TN}/\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}$

1. Sample random matrices $X^{(1)}, \ldots, X^{(n)}$ i.i.d. $\sim \mathcal{N}_{f,m}(0, A \otimes B)$:

$$X^{(t)} = B^{1/2} Z(t) A^{1/2} \qquad \text{where } Z_{ij}(t) \sim N(0,1)\ \forall i, j, \forall t = 1, \ldots, n.$$

   Compute the sample column correlation $\widehat{\mathrm{corr}}_{\mathrm{col}}$ as in (7a) and row correlation $\widehat{\mathrm{corr}}_{\mathrm{row}}$ as in (7b).
2. For each $\lambda \in T$ and $\nu \in T'$:
   (a) Obtain the estimated inverse correlation matrices $\widehat{A}^{-1}$, and $\widehat{B}^{-1}$ with glasso($\widehat{\mathrm{corr}}_{\mathrm{col}}, \lambda$) and glasso($\widehat{\mathrm{corr}}_{\mathrm{row}}, \nu$), respectively. Let $\widehat{\Omega}(\lambda) := \widehat{A}_*^{-1}$ and $\widehat{\Pi}(\nu) := \widehat{B}_*^{-1}$, where $\widehat{A}_*$ and $\widehat{B}_*$ are as defined in (42).
   (b) Let $\widehat{E}(\lambda)$ denote the set of edges in the estimated $\widehat{\Omega}(\lambda)$. Now compute $\mathrm{FNR}(\lambda)$, $\mathrm{FPR}(\lambda)$ and $\mathrm{MCC}(\lambda)$ as defined in Table 1. To obtain $\mathrm{FNR}(\nu)$, $\mathrm{FPR}(\nu)$ and $\mathrm{MCC}(\nu)$, we replace $\widehat{E}(\lambda)$ with $\widehat{F}(\nu)$, which denotes the set of edges in $\widehat{\Pi}(\nu)$, E with F, and m with f. Compute the relative errors $\|\widehat{\Omega}(\lambda) - \Omega\|/\|\Omega\|$ and $\|\widehat{\Pi}(\nu) - \Pi\|/\|\Pi\|$, where $\|\cdot\|$ denotes the operator or the Frobenius norm.

After 100 trials, we plot each of the following as λ changes over a range of values in T: $(\overline{\mathrm{FNR}} + \overline{\mathrm{FPR}})(\lambda)$ and $\overline{\mathrm{MCC}}(\lambda)$ for $\widehat{E}(\lambda)$, where $\overline{\mathrm{FNR}}$, $\overline{\mathrm{FPR}}$ and $\overline{\mathrm{MCC}}$ are averaged over the 100 trials, and the average relative errors in the operator and the Frobenius norm. Similarly, we plot these as ν changes over a range of values in T'. Figure 2 shows how these four metrics change as the ℓ_1 regularization parameters λ and ν increase over full paths where covariance A comes from either AR(1) or the Star-Block model, and Π comes from the random graph model. These plots show that the Gemini method is able to select the correct structures as well as achieving low relative errors in the operator and the Frobenius norm when λ and ν are chosen from a suitable range. In addition, as n increases, we see performance gains over almost the entire paths for all metrics as expected. Other model combinations of A, B which are not shown here confirm similar findings.

In Figure 2, we also illustrate choosing the penalty parameters λ and ν by 10-fold cross-validation. To do so, we run the following for 10 trials. In each trial, we partition the rows of each $X^{(t)}, t = 1, \ldots, n$ into 10 folds. For each

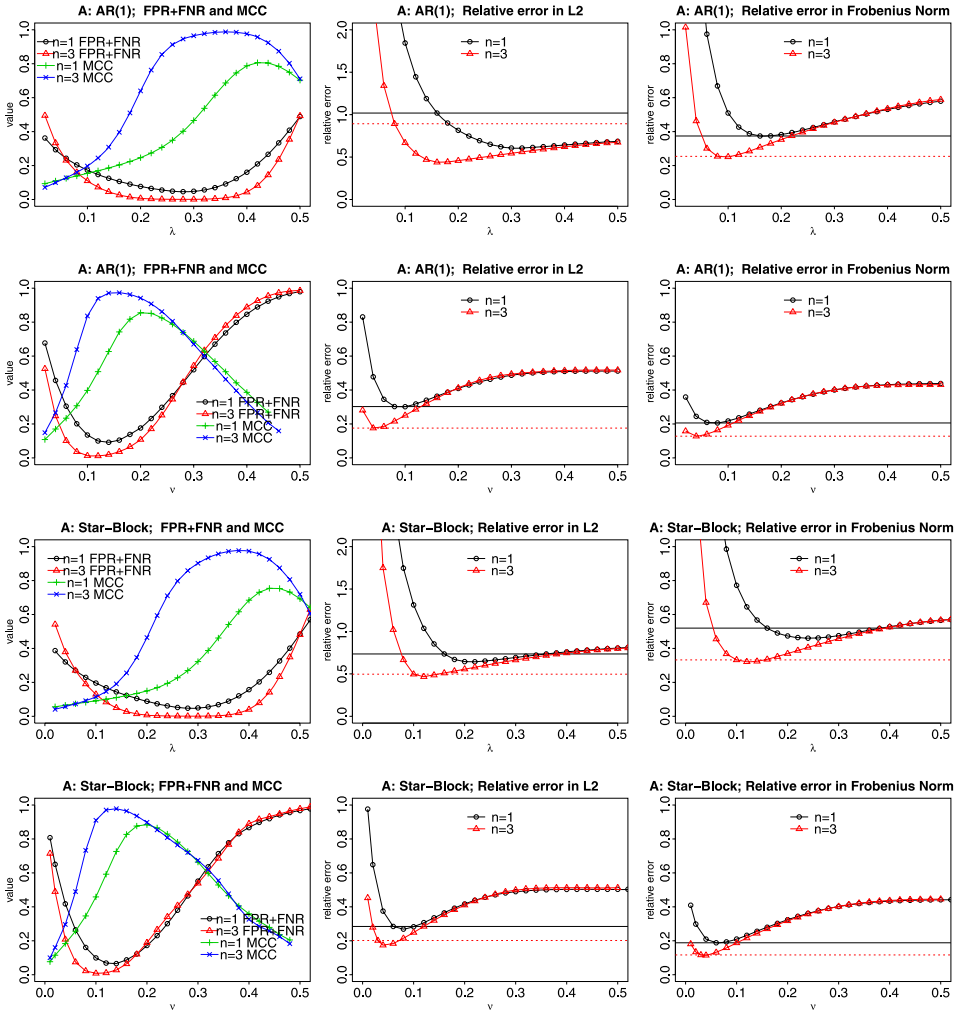

FIG. 2. $m = 400$, $f = 80$. B^{-1} *follows random graph model with* $d = 80$ *and* $w \in [0.1, 0.3]$ *throughout these plots. In the top two panels, covariance A follows the* AR(1) *model with* $\rho = 0.5$; *for the bottom two panels, A follows the Star-Block model. The top and the third panel are for* $\widehat{\Omega}(\lambda)$; *the second and the bottom panel are for* $\widehat{\Pi}(\nu)$. *As* λ *or* ν *increases, FPs decrease and FNs increase. As a result,* FPR + FNR *first decreases and then increases, and on the other hand,* MCC *first increases as the estimated graphs becomes more accurate, and then decreases due to missing edges caused by large penalization. The relative errors also first decrease and then increase before leveling off. This is because decreased FPs first help reduce the estimation errors; however, as penalization increases, the estimated graphs miss more and more edges until only diagonal entries remain in the inverse covariance estimates. Solid and dashed horizontal lines in the second and third columns show the performances of Gemini for cross-validated tuning parameters: in the top two panels,* $\lambda_{\mathrm{CV}} = 0.16$ *and* $\nu_{\mathrm{CV}} = 0.08$ *for* $n = 1$, *and* $\lambda'_{\mathrm{CV}} = 0.08$ *and* $\nu'_{\mathrm{CV}} = 0.04$ *for* $n = 3$. *For the bottom two panels,* $\lambda_{\mathrm{CV}} = 0.16$ *and* $\nu_{\mathrm{CV}} = 0.10$ *for* $n = 1$, *and* $\lambda'_{\mathrm{CV}} = 0.06$ *and* $\nu'_{\mathrm{CV}} = 0.03$ *for* $n = 3$. *These tuning parameters tend to stay near the* λ *or* ν *that minimizes the relative error in the Frobenius norm.*

fold, the validation set consists of the subset of rows of $X^{(1)}, \ldots, X^{(n)}$ sharing the same indices and its complement set serves as the training data. Denote by $\widehat{\mathrm{corr}}_T$ and $\widehat{\mathrm{corr}}_V$ the column-wise sample correlations based upon the training and the validation data, which are computed in the same manner as in (7a). We define $\mathrm{score}_A(\lambda) = \mathrm{tr}(\widehat{\Theta}_\lambda \widehat{\mathrm{corr}}_V) - \log|\widehat{\Theta}_\lambda|$, where $\widehat{\Theta}_\lambda = \mathrm{glasso}(\widehat{\mathrm{corr}}_T, \lambda)$. The final score for a particular λ is the average over 10 trials (with 10 folds in each trial) and the one with the lowest score is chosen to be λ_{CV}. Similarly, we use column partitions to obtain ν_{CV}. We leave the theoretical analysis on cross-validation as future work.

7.2. *ROC comparisons*. In this section, we compare the performances of the two methods, namely, the baseline Gemini and its three-step FF variant over the full paths by examining their ROC curves. Each curve is an average over 50 trials. We fix $f = 80$, $m = 400$, $n = 1$.

To simplify our notation, we summarize the penalty parameters which we use for indexing the ROC curves as follows:

$$\lambda = \lambda_{B_0}, \qquad \nu = \lambda_{A_0}, \qquad \phi = \lambda_{B_1}, \qquad \upsilon = \lambda_{A_1}.$$

To illustrate the overall performances of the baseline Gemini method for estimating the graphs of Ω and Π, we use pairs of metrics $(\overline{\mathrm{FPR}}(\lambda), 1 - \overline{\mathrm{FNR}}(\lambda))$ and $(\overline{\mathrm{FPR}}(\nu), 1 - \overline{\mathrm{FNR}}(\nu))$, respectively, which we obtain as the average over 50 trials of steps 1 and 2 as described in Section 7.1. To plot the ROC curves for the FF method, we start with estimating Π with the Gemini estimator. Due to computational complexity, we specify the input parameters of the subsequent steps sequentially. These choices are not feasible in practical settings. We run through this idealized example for the sake of comparing with the baseline Gemini estimators. Repeat the following 50 times: Let $T := \{0.02, 0.04, \ldots, 0.72\}$.

1–2. Run steps 1, 2 as in Section 7.1, while only computing the metrics for $\widehat{\Pi}(\nu)$, where $\nu \in T$.

3. To execute the second step of the FF algorithm, we use the following three outputs from step 2 of the current procedure to act as B_1 to compute $\widetilde{A}(B_1)$. We choose the output B_1 such that its corresponding ν is chosen to be $\nu_1 = \arg\min_{\nu \in T}(\mathrm{FNR} + \mathrm{FPR})(\nu)$, $\nu_2 = \arg\min_{\nu \in T} \|\widehat{\Pi}(\nu) - \Pi\|_2 / \|\Pi\|_2$, and $\nu_3 = \arg\min_{\nu \in T} \|\widehat{\Pi}(\nu) - \Pi\|_F / \|\Pi\|_F$. Denote these by B_1^1, B_1^2 and B_1^3. We now run the second step of the FF method for each B_1^i, where $i = 1, 2, 3$, with penalty parameter $\phi \in T$ changing over the full path while obtaining the inverse estimators $\widehat{\Omega}^i(\phi)$ for Ω and computing $\mathrm{FNR}^i(\phi)$, and $\mathrm{FPR}^i(\phi)$ for each estimated edge set. These contribute to 3 ROC curves for estimating the edges in E.

4. To execute the last step of the FF method, we use the following three outputs from step 3 as A_1 to compute $\widetilde{B}(A_1)$. We choose the output A_1 such that its corresponding (i, ϕ) is chosen to be optimal with respect to one of the following metrics: $(i_1, \phi_1) = \arg\min_{\phi \in T, i=1,2,3}(\mathrm{FNR}^i + \mathrm{FPR}^i)(\phi)$, $(i_2, \phi_2) = \arg\min_{\phi \in T, i=1,2,3} \|\widehat{\Omega}^i(\phi) - \Omega\|_2 / \|\Omega\|_2$, and $(i_3, \phi_3) =$

$\arg\min_{\phi\in T, i=1,2,3} \|\widehat{\Omega}^i(\phi) - \Omega\|_F / \|\Omega\|_F$. The choices then become (υ_{i_j}, ϕ_j), $j = 1, 2, 3$, which we simply denote by ϕ_1, ϕ_2, ϕ_3. Thus, there are again three choices for A_1. We now run the third step of the FF method for each $\widetilde{B}(A_1)$ with $\upsilon \in T$ changing over the full path, while computing $\mathrm{FNR}^j(\upsilon)$ and $\mathrm{FPR}^j(\upsilon)$, where $j = 1, \ldots, 3$, for each estimated edge set. These contribute to 3 ROC curves for estimating the edges of F.

The ROC curves are plotted in Figure 3 using pairs of metrics $(\overline{\mathrm{FPR}}^i(\phi), 1 - \overline{\mathrm{FNR}}^i(\phi))$ and $(\overline{\mathrm{FPR}}^j(\upsilon), 1 - \overline{\mathrm{FNR}}^j(\upsilon))$, $i, j = 1, 2, 3$, which are averaged over 50 trials. Throughout the plots on the left column of Figure 3, we see clear performance gains of the FF method over the baseline Gemini on estimating $\Omega = A^{-1}$, when the initial penalty ν is chosen properly. For $\Pi = B^{-1}$ in the middle column, we do not always see improvements when w is drawn from $[0.6, 0.8]$. We do see some improvements in case w is drawn from $[0.1, 0.3]$ and when the total correlation ρ_B^2 is small. Overall, the performance gains for Π are not as substantial as those for Ω. These observations are consistent with our theory and discussion in Section 6.1.

7.3. *Summary on the ROC curves.* We use the following metrics to compare matrix B and A across different models or parameters:

1. Total correlation: $\rho_A^2 = \sum_{i<j} \rho_{ij}^2(A)/\binom{m}{2}$ and $\rho_B^2 = \sum_{i<j} \rho_{ij}^2(B)/\binom{f}{2}$.
2. $\|B\|_F / \mathrm{tr}(B)$ and $\|A\|_F / \mathrm{tr}(A)$: these affect the entry-wise error bound in sample correlation estimates for $\rho_{ij}(A)$ and $\rho_{ij}(B)$, for all $i \neq j$, for the baseline Gemini estimators.
3. The pairs of ℓ_1-metrics $(|\rho(B)^{-1}|_{1,\mathrm{off}}, |\rho(B)^{-1}|_1)$ and $(|\rho(A)^{-1}|_{1,\mathrm{off}}, |\rho(A)^{-1}|_1)$.

The total correlation metric comes from [6]. We use it to characterize the average squared magnitudes for correlation coefficients of $\rho(A)$ or $\rho(B)$. They are clearly relevant for the FF method as the entry-wise error bound for estimating $\rho_{ij}(A)$ and $\rho_{ij}(B)$, for all $i \neq j$, depends on the magnitude of the entry itself (cf. Theorems 6.2 and 6.4).

We summarize our findings across the ROC curves in the right column in Figure 3. First, we focus on the case when A is fixed and B is changing. When Π follows the random graph model, we observe that for both the baseline Gemini estimators and their FF variants, the performances in terms of estimating edges for Ω are better when the weights for Π are chosen from $[0.1, 0.3]$ for both $d = 90$ and $d = 180$. Here, the sparsity for Π is not the decisive factor. This is consistent with our theory, in view of Table 2, that $\|B\|_F / \mathrm{tr}(B)$ affects the entry-wise error bound for the baseline Gemini correlation estimate $\widehat{\Gamma}(A)$ as shown in Theorem 4.4, and the pair of metrics $(|\rho(B)^{-1}|_{1,\mathrm{off}}, |\rho(B)^{-1}|_1)$ affect that for the FF correspondent in (30) as shown in Theorem 6.2. The performances in terms of edge recovery for

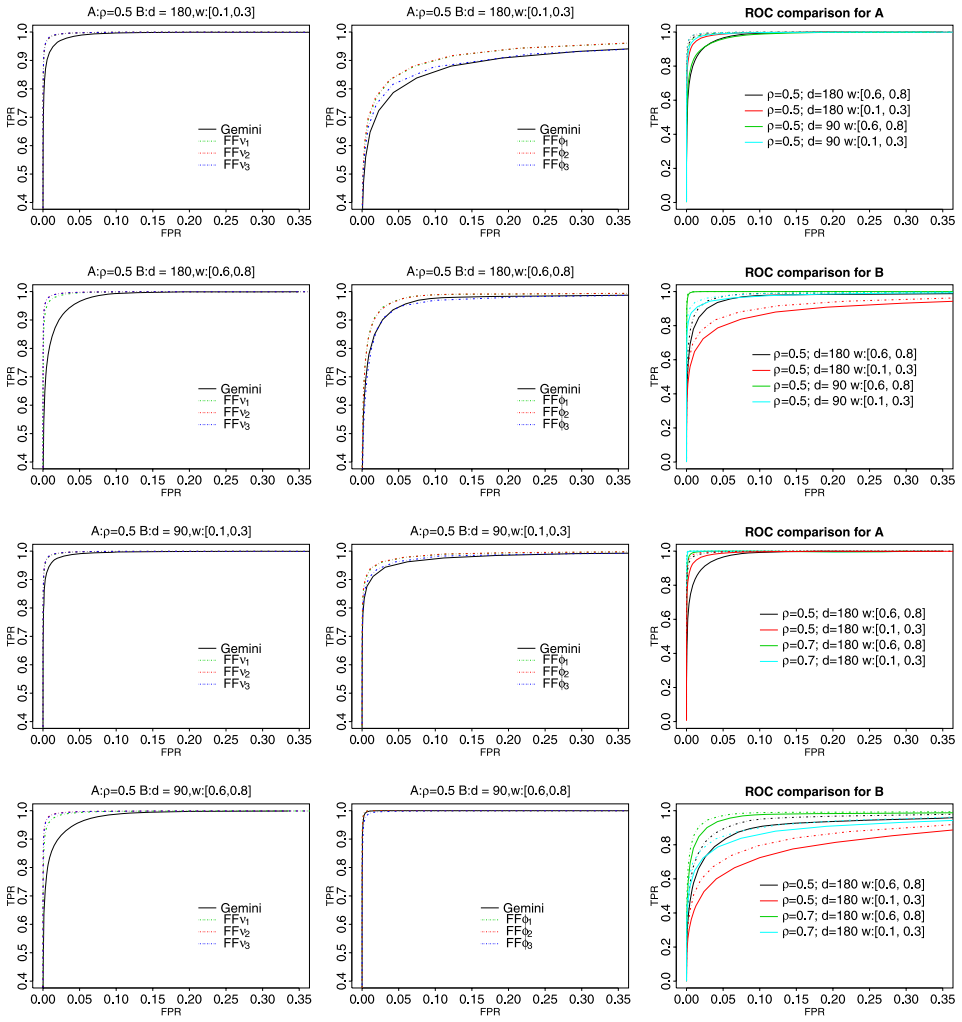

FIG. 3. $m = 400$, $f = 80$, $n = 1$. *Solid lines are for Gemini. Plots in the left column are for A and the middle column are for B. The three dotted lines in each plot on the left column correspond to the three optimization criteria* ν_1, ν_2, ν_3 *as specified in step* 3. *For the middle column, they correspond to* $(i_1, \phi_1), (i_2, \phi_2), (i_3, \phi_3)$, *as specified in step* 4. *In the right column: in top two plots, we choose A from* AR(1) *model with* $\rho = 0.5$ *while changing the settings of* B^{-1} *as in Table 2; in bottom two plots, we choose A from* AR(1) *model with* $\rho = 0.5$ *or* 0.7 *while changing the settings of* B^{-1} *with* $d = 180$. *Dotted lines in plots for A on the right column are chosen according to the optimization criterion* ν_1, *and in plots for B, they are chosen according to the criterion* ϕ_1.

Π take a different order. The sparse random graphs with $d = 90$ see better performances than those with $d = 180$ for both the Gemini and the FF methods. For graphs with the same sparsity, the one with the larger weight performs better. This is consistent with our theory in Section 14.1.

TABLE 2
Metrics for comparing the ROC curves

Metric	$d = 90$ $w : [0.1, 0.3]$	$d = 180$ $w : [0.1, 0.3]$	$d = 90$ $w : [0.6, 0.8]$	$d = 180$ $w : [0.6, 0.8]$
ρ_B^2	0.053	0.06	0.094	0.12
$\|B\|_F / \operatorname{tr}(B)$	0.128	0.13	0.155	0.16
ℓ_1-metrics	(55, 152)	(71, 166)	(99, 225)	(102, 216)

Next, we choose two covariance matrices for both A and B: for B, we choose the two cases with different edge weights with $d = 180$; and for A, we set the parameter ρ to 0.5 or 0.7. The metrics for the two choices of A are: for $\rho = 0.5$, we have $\rho_A^2 = 0.04$, $\|A\|_F / \operatorname{tr}(A) = 0.065$, and ℓ_1-metrics $= (532, 1198)$. The corresponding numbers for $\rho = 0.7$ are: 0.07, 0.085, and $(1095, 2262)$, respectively.

First, we note that the two cases of B show the same trend when A is fixed. In the right bottom two plots in Figure 3, for the graphs of Ω, we find it easier to estimate when their covariance matrices come with parameter $\rho = 0.7$, which results in larger ℓ_1 metrics, and hence larger weights on the inverse chain graph; for the graphs of Π, we observe relatively larger performance gains when $\rho = 0.5$ for A, with the most significant occurring when $w \in [0.1, 0.3]$ for Π, where both $\rho(A)^{-1}$ and $\rho(B)^{-1}$ have smaller ℓ_1 metrics and the total correlation $\rho_B^2 = 0.06$ is also small. The least improvement we see occurs in case all three metrics are large: $\rho = 0.7$, $w \in [0.6, 0.8]$, and $\rho_B^2 = 0.12$. These findings are consistent with results in Theorem 6.2 and 6.4, where we explicitly show the influence of the pairs of ℓ_1-metrics on the error bounds for the FF sample correlation estimates.

7.4. *Application to EEG data.* In this section, we present results of applying Gemini on real data. We used the EEG (electroencephalography) data available from the UCI Machine learning repository [20], which was collected as part of the COGA (Collaborative Studies on Genetics of Alcoholism) project [27]. The data set we used contains measurements from 64 electrodes (channels) placed on two subjects' (one alcoholic and one control) scalps, which were sampled at 256 Hz (3.9-msec epoch) for 1 second. The data consists of 10 runs under three different stimulus paradigm. For each paradigm, we construct an $f \times m$ matrix, X, for each subject's each run, where $f = 64$ and $m = 256$. Each row in X represents a channel and each column represents a measurement epoch. We normalize each row vector such that its mean is 0 and variance is 1. The 10 runs are treated as 10 replicates, and fed to Gemini to estimate both the dependence structures of channels and measurements. We show the resulting graphs for control subject `c02c0000337` under one stimulus paradigm in Figure 4. The estimated graph among channels largely reflects the spatial organization of the brain, and the estimated graph among measurement epochs suggests relatively short-order serial dependence.

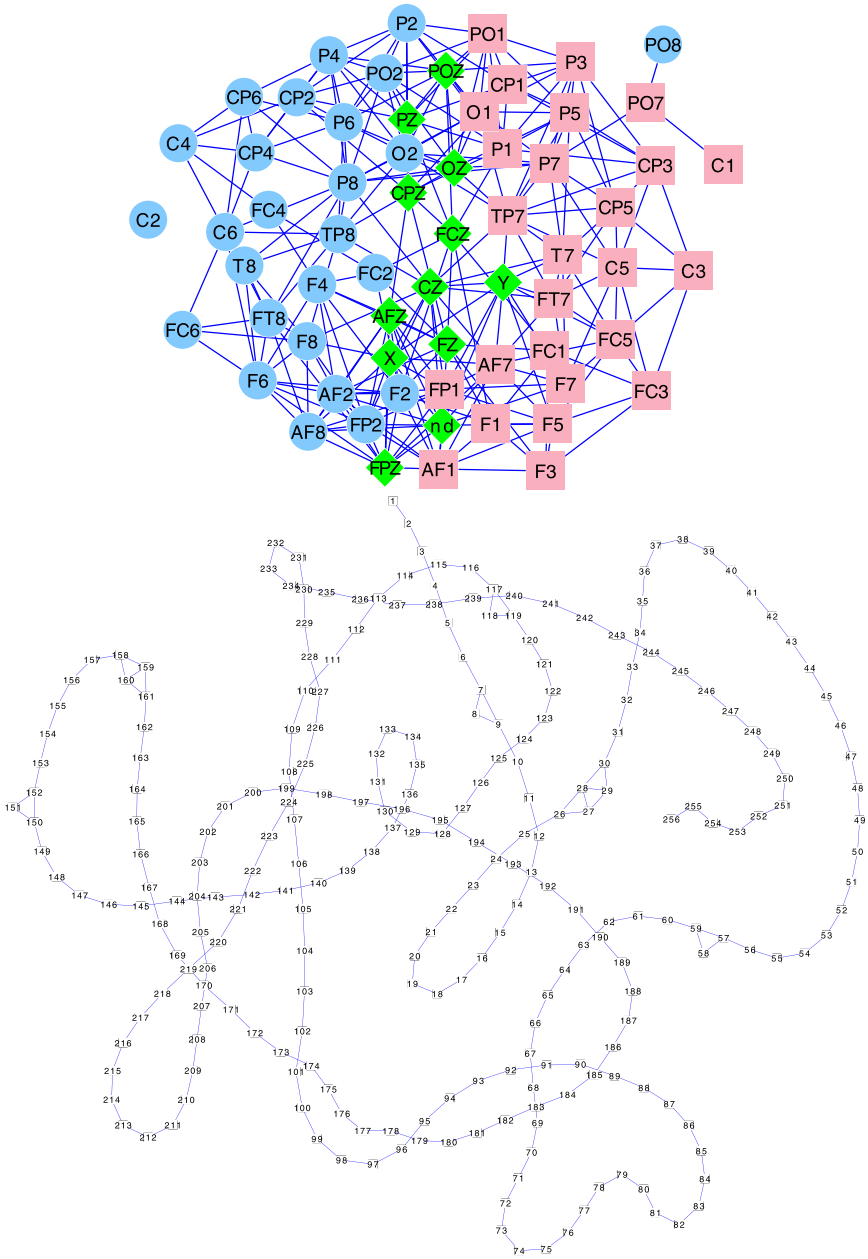

FIG. 4. Top: *Estimated graph of channels with penalty* $\lambda = 0.40$. *Nodes are labeled with EEG electrode identifiers. Circles, squares and diamonds represent electrodes placed on the left, right, and middle of the head respectively. The graph structure indicates that nodes interact mostly with nodes that are physically close to them.* Bottom: *Estimated graph among measurements with penalty* $\nu = 0.78$. *Nodes are labeled with epochs from* 1 *to* 256. *The graph is primarily a long chain showing sequential dependences among epochs with a few extra edges between some neighbors.*

8. Conclusion. In this paper, we presented two methods for estimating graphs in a matrix variate normal model. The baseline Gemini method is rather simple and provides the same rates of convergence as the Noniterative Penalized Flip-Flop method in the operator and the Frobenius norm. In Gemini, a unique pair of optimal solutions for the correlation matrices and their inverses are obtained via the graphical Lasso algorithm. Under sparsity constraints and upon multiplication by proper weight matrices, the penalized estimators are strikingly effective in approximating the row and column covariance matrices. Under sparsity conditions as detailed in (A1) and (A2), the NiPFF method shows some improvement over the baseline algorithm in estimating A_0^{-1}, which is assumed to be the one with the larger dimension, so long as $\rho(B_0)^{-1}$ satisfies a certain additional sparsity condition, namely, its vector ℓ_1 metrics are bounded in the order of its dimensionality. However, we show in both theoretical analysis and simulation results that the performance gains for estimating B_0^{-1} using the NiPFF method at the third step are rather limited; hence, we do not advocate iterating beyond the first three steps. Although our primary interests are in estimating correlations and partial correlations among and between both rows and columns when X follows a matrix variate normal distribution, our methods clearly can be extended to the general cases when the data matrix X follows other type of matrix-variate distributions.

Acknowledgements. The author is grateful for the helpful discussions with Xuming He, John Lafferty, Mark Rudelson, Kerby Shedden and Stanislaw Szarek. The author thanks the Co-Editor Runze Li, an Associate Editor and the anonymous referees for their valuable comments and suggestions.

SUPPLEMENTARY MATERIAL

Supplementary material for "Gemini: Graph estimation with matrix variate normal instances" (DOI: 10.1214/13-AOS1187SUPP; .pdf). The technical proofs are given in the supplementary material [29].

REFERENCES

[1] ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.* **4** 764–790. MR2758420

[2] BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[3] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[4] DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

[5] DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.* **64** 105–123.

[6] EFRON, B. (2009). Are a set of microarrays independent of each other? *Ann. Appl. Stat.* **3** 922–942. MR2750220

[7] FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

[8] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[9] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[10] GUPTA, A. K. and VARGA, T. (1992). Characterization of matrix variate normal distributions. *J. Multivariate Anal.* **41** 80–88. MR1156682

[11] KALAITZIS, A., LAFFERTY, J., LAWRENCE, N. and ZHOU, S. (2013). The bigraphical lasso. In *Proceedings of the* 30*th International Conference on Machine Learning* (*ICML*-13). *JMLR W&CP* **28** 1229–1237. Atlanta, GA.

[12] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

[13] LENG, C. and TANG, C. Y. (2012). Sparse matrix graphical models. *J. Amer. Statist. Assoc.* **107** 1187–1200. MR3010905

[14] LU, N. and ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457. MR2187860

[15] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[16] PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

[17] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

[18] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

[19] TSILIGKARIDIS, T., HERO, A. O. III and ZHOU, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Trans. Signal Process.* **61** 1743–1755. MR3038387

[20] UCI (1999). UCI machine learning repository. Available at http://archive.ics.uci.edu/ml/datasets/EEG+Database.

[21] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[22] WEICHSEL, P. M. (1962). The Kronecker product of graphs. *Proc. Amer. Math. Soc.* **13** 47–52. MR0133816

[23] WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans. Signal Process.* **56** 478–491. MR2445531

[24] YIN, J. and LI, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Multivariate Anal.* **107** 119–140. MR2890437

[25] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[26] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

[27] ZHANG, X. L., BEGLEITER, H., PORJESZ, B., WANG, W. and LITKE, A. (1995). Event related potentials during object recognition tasks. *Brain Res. Bull.* **38** 531–538.

[28] ZHANG, Y. and SCHNEIDER, J. (2010). Learning multiple tasks with a sparse matrix-normal penalty. In *Advances in Neural Information Processing Systems* **23** (*NIPS* 2010) (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.).

[29] ZHOU, S. (2013). Supplement to "Gemini: Graph estimation with matrix variate normal instances." DOI:10.1214/13-AOS1187SUPP.

[30] ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2010). Time varying undirected graphs. *Machine Learning* **80** 298–319.

[31] ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.* **12** 2975–3026. MR2854354

[32] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: shuhengz@umich.edu