




Sparse Multi-Dimensional Graphical Models: A Unified Bayesian Framework

Yang Ni, Francesco C. Stingo & Veerabhadran Baladandayuthapani


To cite this article: Yang Ni, Francesco C. Stingo & Veerabhadran Baladandayuthapani (2017): Sparse Multi-Dimensional Graphical Models: A Unified Bayesian Framework, Journal of the American Statistical Association, DOI: [10.1080/01621459.2016.1167694](https://doi.org/10.1080/01621459.2016.1167694)


To link to this article: <http://dx.doi.org/10.1080/01621459.2016.1167694>

 View supplementary material 

 Accepted author version posted online: 06 Apr 2016.
Published online: 12 Apr 2017.

 Submit your article to this journal 

 Article views: 310

 View related articles 

 View Crossmark data 

Sparse Multi-Dimensional Graphical Models: A Unified Bayesian Framework

Yang Ni^{a,b}, Francesco C. Stingo^{a,c}, and Veerabhadran Baladandayuthapani^a

^aDepartment of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX; ^bDepartment of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX; ^cDipartimento di Statistica, Informatica, Applicazioni “G.Parenti,” University of Florence, Florence, Italy

ABSTRACT

Multi-dimensional data constituted by measurements along multiple axes have emerged across many scientific areas such as genomics and cancer surveillance. A common objective is to investigate the conditional dependencies among the variables along each axes taking into account multi-dimensional structure of the data. Traditional multivariate approaches are unsuitable for such highly structured data due to inefficiency, loss of power, and lack of interpretability. In this article, we propose a novel class of multi-dimensional graphical models based on matrix decompositions of the precision matrices along each dimension. Our approach is a unified framework applicable to both directed and undirected decomposable graphs as well as arbitrary combinations of these. Exploiting the marginalization of the likelihood, we develop efficient posterior sampling schemes based on partially collapsed Gibbs samplers. Empirically, through simulation studies, we show the superior performance of our approach in comparison with those of benchmark and state-of-the-art methods. We illustrate our approaches using two datasets: ovarian cancer proteomics and U.S. cancer mortality. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2015
Revised December 2015

KEYWORDS

Decomposable and nondecomposable graphs; Directed acyclic graphs; LDL decomposition; Partially collapsed Gibbs sampler

1. Introduction

With rapid technological advancements, increasing numbers of newly collected datasets have intrinsic multi-dimensional structure, that is, the observed statistical atom/sample is a d -dimensional array with $d \in \mathbb{N}^+$. The simplest example is the two-dimensional case, for example, genomic data measured under different experimental conditions or with different molecular platforms, where one dimension (rows) is genes/proteins and the other dimension (columns) could be experimental conditions or platforms. Higher dimensional data such as cancer mortality counts across multiple cancers, geographical locations, and years are also widely available. Our objective is to investigate the conditional dependencies among the variables along each dimension. However, this poses a challenging statistical problem, both methodologically and computationally, of defining a statistical framework that properly models and exploits the multi-dimensional structure of such types of data.

Considering the two-dimensional case, intuitively, one can stack any $q \times p$ random matrix into a vector (where q, p denote the cardinality of each dimension) and model the resulting vector using standard approaches, for example, via multivariate Gaussian graphical models (GGMs). However, this naive approach ignores the multi-dimensional structure of the data, which would result in inefficient and inaccurate estimation (as we shall see in the simulation studies). Moreover, the computational burden would be extremely heavy when either or both q, p are moderately large since the dimension of the covariance matrix of the vectorized data is $qp \times qp$. More importantly, by vectorizing the data, we also lose the interpretability associated with each of the dimensions since, in practice, the

variables are noninterchangeable across dimensions. Analogously, these concerns increase in severity for higher dimensional array-variate data. Hence, it is highly desirable to develop models that take advantage of structural information inherent in the multi-dimensional data, reduce the dimensionality of the covariance matrix and generate context-specific interpretable results.

Multivariate GGMs (Lauritzen 1996; Whittaker 2009) are commonly used to probabilistically model conditional independence in multivariate Gaussian data and have been successfully applied in many fields, including genomics, health sciences, macroeconomics, and many others. The estimation methods in GGMs have been extensively studied in the literature for both directed (Friedman et al. 2000; Spirtes, Glymour, and Scheines 2000; Geiger and Heckerman 2002; Shojai and Michailidis 2010; Stingo et al. 2010; Yajima et al. 2015) and undirected graphs (Dobra et al. 2004; Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Banerjee, El Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008; Carvalho and Scott 2009; Kundu, Baladandayuthapani, and Mallick 2013; Stingo and Marchetti 2015).

In the context of large multi-dimensional data, several methods for matrix-variate Gaussian graphical models (mGGMs) have been proposed. Carvalho et al. (2007) first introduced Bayesian dynamic matrix-variate graphical models for multivariate time series analysis with row covariance matrix fixed. Wang and West (2009) proposed a fully Bayesian approach based on the marginal likelihood for analyzing mGGMs with a focus on decomposable graphs. Dobra, Lenkoski, and Rodriguez (2011) developed a Markov chain Monte Carlo (MCMC)

algorithm based on the full likelihood for the inference and model determination of unrestricted mGGMs. Penalized likelihood approaches have also been developed: Allen et al. (2010) proposed an ℓ_1 and ℓ_2 penalized likelihood approach when the sample size is one, which was extended by Yin and Li (2012) to multiple iid random matrices. Similarly, Leng and Tang (2012) defined an ℓ_1 and smoothly clipped absolute deviation (SCAD) penalty for a matrix-variate Gaussian likelihood. Zhou (2014) applied an ℓ_1 penalty on only the off-diagonal elements of the inverse correlation matrix. All the aforementioned methods are restricted solely to undirected matrix-variate graphs and do not generalize to the multiple (> 2) dimensions.

The contributions of this article are four-fold. (1) We propose a novel matrix-variate directed acyclic graph (mDAG) framework to model matrix-variate Gaussian data when there is a prior ordering in both columns and rows. (2) We extend our approach to model determination of undirected mGGMs, where we focus only on decomposable graphs. (3) Our approach is a unified framework for both directed and undirected graphical models, that is, not every dimension has to be modeled by the same type of graph. For example, in a matrix-variate graph, the row graph may be represented by a directed graph when the row prior ordering is known (e.g., time or known genomic pathways); whereas the column graph may be represented by an undirected graph when such prior ordering is missing. This *hybrid* mGGM strategy allows for more flexibility of the methods. (4) Our model formulation allows for natural generalizations to multiple (> 2) array-variate data, called array-variate GGMs (aGGMs). As we show in our application to cancer mortality data, we can apply our framework to more than two dimensions and go beyond the type of inference typically allowed by state-of-the-art approaches.

More importantly, our modeling approach can be paired with efficient MCMC algorithms based on marginal likelihoods. When dimensions are moderately large, a partially collapsed Gibbs sampler (PCGS, van Dyk and Park 2008) is adopted for efficient sampling. Our modeling and computational strategies for multi-dimensional graphical models resulted in superior performances, in terms of both learning the structure and estimating the precision matrix, when compared to alternative naive DAG approaches as well as to the state-of-the-art Bayesian (Dobra, Lenkoski, and Rodriguez 2011) and non-Bayesian (Leng and Tang 2012) mGGM approaches.

Our Bayesian approaches allow us to naturally account for uncertainty in the graph structure and to produce regularized and sparse estimators. The graph uncertainty is especially important in the context of high-dimensional complex data, since with a limited sample size, several graphs may explain the data equally well and hence point estimators are often not adequate. We address this situation of accounting for model uncertainties under a Bayesian paradigm, where we can derive the posterior probability associated with the point estimators.

The rest of this article is organized as follows. We present the mDAG model in Section 2. We extend it for modeling undirected and hybrid matrix-variate graphs in Section 3 and for modeling array-variate graphs in Section 4. We summarize the posterior estimation and inferential algorithms in Section 5. We present detailed simulation studies in Section 6 and real-data applications in Section 7. Section 8 provides our closing discussion.

2. Matrix-Variate Gaussian Graphical Models

We start with the simplest two-dimensional case, that is, matrix-variate graphs. Let \mathbf{Z} be a random matrix with q rows and p columns and let $\mathbf{Y} = (Y_1, \dots, Y_m)^T = \text{vec}(\mathbf{Z})$ be the vectorization of \mathbf{Z} : $Y_l = Z_{ij}$ for $i = 1, \dots, q$, $j = 1, \dots, p$, $m = qp$ and $l = (j-1)q + i$. A random matrix \mathbf{Z} follows a matrix-variate Gaussian distribution, with mean $\mathbf{M}(q \times p)$, row covariance $\mathbf{U}(q \times q)$ and column covariance $\mathbf{V}(p \times p)$, denoted by $\mathbf{Z} \sim MN_{q \times p}(\mathbf{M}, \mathbf{U}, \mathbf{V})$, if \mathbf{Y} follows a multivariate Gaussian distribution $\mathbf{Y} \sim N_m(\text{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$, in which \otimes is the Kronecker product. The probability density function is given as (Gupta and Nagar 2000):

$$p(\mathbf{Z}|\mathbf{M}, \mathbf{U}, \mathbf{V}) = (2\pi)^{-m/2} |\mathbf{U}|^{-q/2} |\mathbf{V}|^{-p/2} \times \exp \left[-\frac{1}{2} \text{tr}\{(\mathbf{Z} - \mathbf{M})^T \mathbf{U}^{-1} (\mathbf{Z} - \mathbf{M}) \mathbf{V}^{-1}\} \right].$$

Since \mathbf{U} and \mathbf{V} are our main focus, we will assume $\mathbf{M} = \mathbf{0}$ throughout the article. Let $\mathbf{\Omega} = \mathbf{U}^{-1}$ and $\mathbf{\Lambda} = \mathbf{V}^{-1}$ be the precision matrices that encode the conditional independence,

- Rows: $\mathbf{Z}^{i_1} \perp\!\!\!\perp \mathbf{Z}^{i_2} | \mathbf{Z}^{-i_1 i_2}$ if and only if $\mathbf{\Omega}_{i_1 i_2} = \mathbf{\Omega}_{i_2 i_1} = 0$, where $-i_1 i_2 = \{1, \dots, q\} \setminus \{i_1, i_2\}$.
- Columns: $\mathbf{Z}_{j_1} \perp\!\!\!\perp \mathbf{Z}_{j_2} | \mathbf{Z}_{-j_1 j_2}$ if and only if $\mathbf{\Lambda}_{j_1 j_2} = \mathbf{\Lambda}_{j_2 j_1} = 0$, where $-j_1 j_2 = \{1, \dots, p\} \setminus \{j_1, j_2\}$.

Graphical models are useful tools to obtain regularized estimators of covariance/precision matrices, with directed and undirected graphical models being the most common. For the matrix-variate case, $\mathbf{\Omega}$ and $\mathbf{\Lambda}$ are each associated with a graph, denoted respectively by $\mathcal{G}_{\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{\Lambda}}$. Notice that $\mathcal{G}_{\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{\Lambda}}$ do not have to be of the same type. Therefore, there are potentially three different types of matrix-variate graphical models:

- directed: both $\mathcal{G}_{\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{\Lambda}}$ are directed (Section 2.1);
- undirected: both $\mathcal{G}_{\mathbf{\Omega}}$ and $\mathcal{G}_{\mathbf{\Lambda}}$ are undirected (Section 3.1); and
- hybrid: $\mathcal{G}_{\mathbf{\Omega}}$ ($\mathcal{G}_{\mathbf{\Lambda}}$) is directed and $\mathcal{G}_{\mathbf{\Lambda}}$ ($\mathcal{G}_{\mathbf{\Omega}}$) is undirected (Section 3.2).

All existing undirected mGGM (or mGGM for short) approaches deal with only type (ii). The previously proposed Bayesian methods (Wang and West 2009; Dobra, Lenkoski, and Rodriguez 2011) assume conjugate hyper-inverse Wishart/G-Wishart priors on the row and column covariance/precision matrices. In this article, we take a different approach based on Cholesky-type decomposition of the precision matrix. This construction allows us to explicitly induce sparsity through priors and develop computationally efficient algorithms for posterior inference, providing a unified framework that covers all three cases: directed, undirected, and hybrid cases.

2.1. Matrix-Variate Directed Acyclic Graphs

Suppose $\mathbf{Y} = \text{vec}(\mathbf{Z}) \sim N(0, \mathbf{\Phi}^{-1})$, where $\mathbf{\Phi} = \mathbf{\Lambda} \otimes \mathbf{\Omega}$ is the precision matrix. We first consider a multivariate DAG, $\mathcal{G} = (V, E)$ which consists of a set of vertices $V = \{1, \dots, m\}$ that represent random variables $\mathbf{Y} = \mathbf{Y}_V$ and a set of directed edges, $E \subset V \times V$, that represent conditional independence among the variables. The joint distribution of a DAG can be factorized

into local/conditional distributions:

$$p(Y) = \prod_{i=1}^m p(Y_i | Y_{pa(i)}),$$

where $pa(i)$ is the parent¹ set of vertex i . Given a prior ordering of Y (assumed to be $\{1, \dots, m\}$ without loss of generality), a Gaussian DAG can be viewed as a system of recursive linear regressions:

$$\begin{aligned} Y_1 + b_{12}Y_2 + b_{13}Y_3 + \dots + b_{1,m-1}Y_{m-1} + b_{1m}Y_m &= \epsilon_1 \sim N(0, t_1) \\ Y_2 + b_{23}Y_3 + \dots + b_{2,m-1}Y_{m-1} + b_{2m}Y_m &= \epsilon_2 \sim N(0, t_2) \\ &\vdots \\ Y_m &= \epsilon_m \sim N(0, t_m). \end{aligned}$$

Let $E = (\epsilon_1, \dots, \epsilon_m)^T$, $T = \text{diag}(t_1, \dots, t_m)$ and let $B = (b_{ij})$ be a unit upper triangular matrix. Then, the system of recursive regressions can be written in matrix form: $BY = E$ with $E \sim N(0, T)$. It immediately follows that $\Phi = B^T T^{-1} B$, which is simply the LDL decomposition (a variant of Cholesky decomposition) of the precision matrix Φ . Suppose Λ and Ω have LDL decompositions $\Lambda = B_\Lambda^T T_\Lambda^{-1} B_\Lambda$ and $\Omega = B_\Omega^T T_\Omega^{-1} B_\Omega$, where B_Λ and B_Ω are $p \times p$ and $q \times q$ unit upper triangular matrices and $T_\Lambda = \text{diag}(T_\Lambda^{(1)}, \dots, T_\Lambda^{(p)})$ and $T_\Omega = \text{diag}(T_\Omega^{(1)}, \dots, T_\Omega^{(q)})$ are $p \times p$ and $q \times q$ diagonal matrices. Then, we have

$$\begin{aligned} B^T T^{-1} B &= \Phi = \Lambda \otimes \Omega = (B_\Lambda^T T_\Lambda^{-1} B_\Lambda) \otimes (B_\Omega^T T_\Omega^{-1} B_\Omega) \\ &= (B_\Lambda^T \otimes B_\Omega^T) (T_\Lambda^{-1} \otimes T_\Omega^{-1}) (B_\Lambda \otimes B_\Omega). \end{aligned}$$

The last equality is due to the mixed-product property of the Kronecker product. Because of the uniqueness of the LDL decomposition of the positive-definite matrix, we obtain $B = B_\Lambda \otimes B_\Omega$ and $T = T_\Lambda \otimes T_\Omega$. The system of recursive regressions above can be written as

$$(B_\Lambda \otimes B_\Omega)Y = E \text{ with } E \sim N(0, T_\Lambda \otimes T_\Omega). \quad (1)$$

Under this reparameterization, we put priors directly on regression parameters (B_Ω, T_Ω) and (B_Λ, T_Λ) , which induce priors on Λ and Ω (details given in Section 2.2). In the context of multivariate GGMs, approaches based on the Cholesky-type decomposition of the precision matrix have been previously proposed by Wermuth (1980) and Roverato (2000). Wermuth (1980) proved the equivalence between a decomposable GGM and a linear recursive equation whose regression coefficients have the same reducible zero pattern as the precision matrix of a GGM, which allows us to extend mDAG to the undirected case (Section 3.1). Based on the Cholesky decomposition, Roverato (2000) showed a hyper inverse-Wishart matrix can be decomposed into independent normal and square roots of chi-squared variables, which provides a guidance for our prior and hyperparameter specifications (Section 2.2). Our use of the Cholesky-type decomposition of precision matrices leads to the definition of mDAG which is formally stated in the following Definition 1.

Definition 1. The joint distribution of a random matrix Z factorizes according to a matrix-variate Gaussian DAG if and only if

$Y = \text{vec}(Z)$ satisfies Equation (1) with the following probability density function:

$$\begin{aligned} p(Z | B_\Omega, B_\Lambda, T_\Omega, T_\Lambda) &= (2\pi)^{-m/2} |B_\Omega^T T_\Omega^{-1} B_\Omega|^{q/2} |B_\Lambda^T T_\Lambda^{-1} B_\Lambda|^{p/2} \\ &\times \exp \left\{ -\frac{1}{2} \text{tr} (Z^T B_\Omega^T T_\Omega^{-1} B_\Omega Z B_\Lambda^T T_\Lambda^{-1} B_\Lambda) \right\}. \end{aligned} \quad (2)$$

Note that model (1) is not identifiable because $T_\Lambda \otimes T_\Omega = (cT_\Lambda) \otimes (c^{-1}T_\Omega)$ for any constant $c > 0$. We resolve the identifiability issue by fixing $T_\Omega^{(q)} = 1$ (equivalently, $U_{qq} = 1$). The factorization in (1) implies that

- Rows: $Z^i \perp\!\!\!\perp Z^{nd_\Omega(i) \setminus pa_\Omega(i)} | Z^{pa_\Omega(i)}$,
- Columns: $Z_j \perp\!\!\!\perp Z^{nd_\Lambda(j) \setminus pa_\Lambda(j)} | Z^{pa_\Lambda(j)}$,

where $pa_\Omega(i)$ and $nd_\Omega(i)$ are the parents and nondescendants of node i with respect to row graph \mathcal{G}_Ω , which can be read off the graph B_Ω , and similarly, $pa_\Lambda(j)$ and $nd_\Lambda(j)$ are the parents and nondescendants of node j with respect to column graph \mathcal{G}_Λ . This property is also known as the directed (local) Markov property.

2.2. Parameter Priors and Model Selection

We complete our model by prior specification. The likelihood (2) involves two sets of parameters (B_Ω, T_Ω) and (B_Λ, T_Λ) . We first discuss the priors for (B_Ω, T_Ω) ; the priors for (B_Λ, T_Λ) are defined analogously at the end of this section.

The normal-inverse-gamma family is a conjugate prior family for regression coefficients and residual variances in linear regressions and multivariate Gaussian DAGs. In our mDAG context, a normal-inverse-gamma prior is conditional conjugate. It allows for partial analytical integration of (B_Ω, T_Ω) . To induce sparsity, we introduce latent variables $\Gamma_\Omega = (\Gamma_\Omega^{(k,i)})$ and assume the following discrete mixture priors,

$$\begin{aligned} B_\Omega^{(k,i)} | T_\Omega^{(k)}, \Gamma_\Omega^{(k,i)} &\sim \Gamma_\Omega^{(k,i)} N(0, \tau_\Omega T_\Omega^{(k)}) + (1 - \Gamma_\Omega^{(k,i)}) \delta_0; \\ T_\Omega^{(k)} &\sim IG(\alpha_\Omega^{(k)}, \beta_\Omega^{(k)}), \end{aligned} \quad (3)$$

for $k = 1, \dots, q$ and $i = k+1, \dots, q$, where δ_0 is a point mass at zero and $\Gamma_\Omega^{(k,i)}$ follows a Bernoulli prior with success probability ρ_Ω , which has a conjugate beta hyperprior,

$$\Gamma_\Omega^{(k,i)} | \rho_\Omega \sim \text{Bernoulli}(\rho_\Omega); \quad \rho_\Omega \sim \text{Beta}(a_{\rho_\Omega}, b_{\rho_\Omega}). \quad (4)$$

We set the hyperparameters of the inverse-gamma priors at $\alpha_\Omega^{(k)} = \frac{\delta_\Omega + n_\Omega^{(k)}}{2}$, $\beta_\Omega^{(k)} = \frac{1}{2\tau_\Omega}$, with $n_\Omega^{(k)} = \sum_{i=k+1}^q \Gamma_\Omega^{(k,i)}$. By the standard normal distribution theory and following the notation for the inverse Wishart distribution of Dawid (1981), the distribution of (B_Ω, T_Ω) is determined by that of Ω , and this prior and hyperparameter setting is equivalent to the prior induced by $U = \Omega^{-1} \sim IW(\delta_\Omega, \tau_\Omega^{-1} I_q)$ with $\delta_\Omega > 0$ degrees of freedom (Dawid and Lauritzen 1993; Roverato 2000; Dobra et al. 2004). Hence, the prior (3) is consistent with the encompassing inverse Wishart priors on the full covariance matrix U . However, the binary variable Γ_Ω explicitly induces sparsity on B_Ω and represents the structure of the row DAG \mathcal{G}_Ω . The hierarchical priors (4) on Γ_Ω provide an automatic multiplicity control since the posterior distributions of ρ_Ω will shrink toward zero as the total

¹For brevity, we provide the definitions of basic graph theory concepts used throughout this article in Supplementary Material A, including parents, nondescendants, decomposability, adjacency matrix, complete graph, ordering, perfect ordering, Markov equivalence, v-structure, and perfect DAG.

number of variables increases (Scott and Berger 2010). The priors for \mathbf{B}_Λ , \mathbf{T}_Λ , $\mathbf{\Gamma}_\Lambda$, ρ_Λ are defined in a similar fashion,

$$\begin{aligned} B_\Lambda^{(j,i)} \mid T_\Lambda^{(j)}, \Gamma_\Lambda^{(j,i)} &\sim \Gamma_\Lambda^{(j,i)} N(0, \tau_\Lambda T_\Lambda^{(j)}) + (1 - \Gamma_\Lambda^{(j,i)}) \delta_0; \\ T_\Lambda^{(j)} &\sim \text{IG}(\alpha_\Lambda^{(j)}, \beta_\Lambda^{(j)}) \\ \Gamma_\Lambda^{(j,i)} \mid \rho_\Lambda &\sim \text{Bernoulli}(\rho_\Lambda); \quad \rho_\Lambda \sim \text{Beta}(a_{\rho_\Lambda}, b_{\rho_\Lambda}) \end{aligned}$$

for $j = 1, \dots, p$ and $i = j + 1, \dots, p$, with $\alpha_\Lambda^{(j)} = \frac{\delta_\Lambda + n_\Lambda^{(j)}}{2}$, $\beta_\Lambda^{(j)} = \frac{1}{2\tau_\Lambda}$ and $n_\Lambda^{(j)} = \sum_{i=j+1}^p \Gamma_\Lambda^{(j,i)}$.

3. Undirected and Hybrid Matrix-Variate Graphical Models

In this section, we extend our inference procedure for mDAGs to model determination of undirected mGGMs and hybrid mGGMs. We show that both undirected and hybrid mGGMs have likelihoods and posteriors equivalent to those of mDAGs; therefore, the same posterior inference (discussed in Section 5) for mDAGs can be applied to these two cases as well. And we also discuss why vectorization is not feasible from a computational point of view.

3.1. Undirected MGGMs

When prior orderings are not available, undirected graphs are usually preferred to directed graphs. Here, we restrict our attention to decomposable graphs as, in general, there is no one-to-one correspondence between mDAGs and mGGMs.

Following the notation of Wermuth (1980), a set R of pairs of indices is said to be *reducible* if $\forall (i, j) \in R$ with $i < j$, either $(h, i) \in R$ or $(h, j) \in R, \forall h = 1, \dots, i - 1$. The *null set* R with respect to a matrix \mathbf{M} is defined as $R = \{(i, j) \mid M_{ij} = 0\}$. Then for any precision matrix Φ of a multivariate Gaussian distribution, its underlying graph \mathcal{G}_Φ is decomposable if and only if there exists an ordering of Φ such that \mathbf{B}_Φ has the same reducible null set as Φ , where $\mathbf{B}_\Phi^T \mathbf{T}_\Phi^{-1} \mathbf{B}_\Phi = \Phi$ is the LDL decomposition (Proposition 5 of Wermuth 1980). Such ordering is generally not unique and can be obtained by, for example, reversing the perfect ordering (Lauritzen 1996). In essence, Proposition 5 of Wermuth (1980) describes the Markov equivalence relationship between a decomposable graph and a perfect DAG (with no v-structures) through LDL decomposition, that is, Φ and \mathbf{B}_Φ share the same zero patterns given the reverse perfect ordering. In our particular case, assuming the rows and columns of the data matrix \mathbf{Z} comply with the reverse perfect ordering of row graph \mathcal{G}_Ω and column graph \mathcal{G}_Λ , the likelihood for mGGMs is given by

$$p(\mathbf{Z} \mid \Omega, \Lambda) = p(\mathbf{Z} \mid \mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda),$$

where $\Omega = \mathbf{B}_\Omega^T \mathbf{T}_\Omega^{-1} \mathbf{B}_\Omega$, $\Lambda = \mathbf{B}_\Lambda^T \mathbf{T}_\Lambda^{-1} \mathbf{B}_\Lambda$ and $p(\mathbf{Z} \mid \mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ is given in equation (2). Since the normal-inverse-gamma priors of the regression coefficients and error variances specified in Section 2.2 are consistent with the (hyper-) inverse Wishart prior of the covariance matrices, the posteriors of the undirected mGGMs and mDAGs are also equivalent: $p(\Omega, \Lambda \mid \mathbf{Z}) = p(\mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda \mid \mathbf{Z})$.

This alternative parameterization results in variation-independent parameters; the parameter space can be expressed as a Cartesian product of one-dimensional spaces, which is particularly desirable for mGGMs where the identifiability issue usually needs to be addressed carefully. Wang and West (2009) and Dobra, Lenkoski, and Rodriguez (2011) fixed $V_{11} = 1$ or $\Lambda_{11} = 1$ and defined the prior on \mathbf{V} or Λ through a parameter expansion technique. This complication arises because the entries in the precision matrix are not variation independent (constrained by positive definiteness). In contrast, with our variation-independent parameterization, we can fix $T_\Omega^{(q)} = 1$ without needing to apply any additional adjustments to the prior of the other parameters.

3.2. Hybrid MGGMs

Since mDAG is a unified framework, that is, both directed and undirected graphs are modeled through directed graphs, it can be naturally extended to hybrid matrix-variate graphs. This is especially useful in cases where the row (column) ordering is available (e.g., time, known graph/pathway information), but the column (row) ordering is missing (e.g., geographical location, experimental conditions). Suppose, without loss of generality, we model \mathcal{G}_Ω as a directed graph with parameters $(\mathbf{B}_\Omega, \mathbf{T}_\Omega)$ and \mathcal{G}_Λ as an undirected graph with parameter Λ . Using an argument similar to that given in Section 3.1, the likelihood can be defined as

$$p(\mathbf{Z} \mid \mathbf{B}_\Omega, \mathbf{T}_\Omega, \Lambda) = p(\mathbf{Z} \mid \mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda),$$

where $\Lambda = \mathbf{B}_\Lambda^T \mathbf{T}_\Lambda^{-1} \mathbf{B}_\Lambda$ and $p(\mathbf{Z} \mid \mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ is given in Equation (2), and the posterior distribution is defined as $p(\mathbf{B}_\Omega, \mathbf{T}_\Omega, \Lambda \mid \mathbf{Z}) = p(\mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{B}_\Lambda, \mathbf{T}_\Lambda \mid \mathbf{Z})$.

3.3. Computational Concern with Vectorization of Multi-Dimensional Data

In principle, any matrix-variate data can be modeled by using an established multivariate GGM on the vectorized data. However, working directly on the vectorized data would greatly increase the size of the graph and lead to an extremely heavy computational cost. One possible remedy is to impose decomposability, which has previously succeeded in reducing the computational cost in multivariate GGMs (Jones et al. 2005). For example, a hyper inverse-Wishart prior on the precision matrix of a decomposable graph results in tractable marginal likelihoods (Dawid and Lauritzen 1993; Lauritzen 1996). Nevertheless, such an assumption turns out to be too stringent for our case, which is explained by the following argument.

Let \mathbf{A}_Λ and \mathbf{A}_Ω be the adjacency matrices of graphs \mathcal{G}_Λ and \mathcal{G}_Ω , respectively. The Kronecker product of the graphs, $\mathcal{G}_\Phi = \mathcal{G}_\Lambda \otimes \mathcal{G}_\Omega$, is defined as the graph with adjacency matrix $\mathbf{A}_\Phi = \mathbf{A}_\Lambda \otimes \mathbf{A}_\Omega$ (Weichsel 1962). Notice that even if both \mathcal{G}_Λ and \mathcal{G}_Ω are decomposable, their Kronecker product graph, $\mathcal{G}_\Phi = \mathcal{G}_\Lambda \otimes \mathcal{G}_\Omega$, is not necessarily decomposable. One simple example would be $\mathbf{A}_\Lambda = \mathbf{A}_\Omega = [1, 1, 0; 1, 1, 1; 0, 1, 1]$. In Proposition 1, we provide a sufficient and necessary condition for \mathcal{G}_Φ to be decomposable.

Proposition 1. Suppose \mathcal{G}_Λ and \mathcal{G}_Ω are decomposable graphs. The Kronecker product, $\mathcal{G}_\Phi = \mathcal{G}_\Lambda \otimes \mathcal{G}_\Omega$, is decomposable if and only if either \mathcal{G}_Λ or \mathcal{G}_Ω is complete or disconnected with complete components.

The proof is provided in Supplementary Material B. The implication of [Proposition 1](#) is that by assuming that the graph \mathcal{G}_Φ of the vectorized data is decomposable, we implicitly require \mathcal{G}_Λ or \mathcal{G}_Ω to be either complete or disconnected with complete components, which is a seemingly unrealistic assumption for most practical applications where sparseness is expected. Our matrix-variate approach, on the other hand, only assumes decomposability for \mathcal{G}_Λ and \mathcal{G}_Ω , which is much less stringent. A similar result holds for perfect DAGs which is stated as a corollary in Supplementary Material B.

4. Array-Variate Graphical Models

Our matrix-variate graphical model framework can be easily extended to d -dimensional aGGMs for both directed and undirected cases. A $q_1 \times q_2 \times \dots \times q_d$ random array \mathbf{Z} is said to follow a centered array-variate Gaussian distribution $\mathbf{Z} \sim AN(\mathbf{0}, \mathbf{\Omega}_1^{-1}, \mathbf{\Omega}_2^{-1}, \dots, \mathbf{\Omega}_d^{-1})$, where $\mathbf{\Omega}_i(q_i \times q_i)$ is the precision matrix of the i th dimension if \mathbf{Y} follows a multivariate normal distribution $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{\Omega}_d^{-1} \otimes \dots \otimes \mathbf{\Omega}_1^{-1})$, where $\mathbf{Y} = \text{vec}(\mathbf{Z})$ is the vector obtained by stacking the elements of \mathbf{Z} in the order of its dimensions, that is, $\mathbf{Y}_l = \mathbf{Z}_{i_1, \dots, i_d}$ for $l = \sum_{h=2}^d [(i_h - 1) \prod_{j=1}^{h-1} q_j] + i_1$. We refer to Akdemir and Gupta (2011) for more details about the density and properties of array-variate Gaussian distributions. The array-variate DAG (aDAG), given a reverse perfect ordering, is defined as

$$(\mathbf{B}_{\Omega_d} \otimes \dots \otimes \mathbf{B}_{\Omega_1})\mathbf{Y} = \mathbf{E} \text{ with } \mathbf{E} \sim N(\mathbf{0}, \mathbf{T}_{\Omega_d} \otimes \dots \otimes \mathbf{T}_{\Omega_1}), \quad (5)$$

where $\mathbf{\Omega}_i = \mathbf{B}_{\Omega_i}^T \mathbf{T}_{\Omega_i}^{-1} \mathbf{B}_{\Omega_i}$ is the LDL decomposition of the i th precision matrix. Analogously, the identifiability issue $\mathbf{T}_{\Omega_d} \otimes \dots \otimes \mathbf{T}_{\Omega_1} = (c_1 \dots c_{d-1} \mathbf{T}_{\Omega_d}) \otimes (\mathbf{T}_{\Omega_{d-1}}) \otimes \dots \otimes (\mathbf{T}_{\Omega_1})$ can be solved by fixing $\mathbf{T}_{\Omega_i}^{(q_i)} = 1$ for $i = 1, \dots, d-1$. If we let $\mathbf{B}_\Omega = \mathbf{B}_{\Omega_{d-1}} \otimes \dots \otimes \mathbf{B}_{\Omega_1}$, $\mathbf{B}_\Lambda = \mathbf{B}_{\Omega_d}$, $\mathbf{T}_\Omega = \mathbf{T}_{\Omega_{d-1}} \otimes \dots \otimes \mathbf{T}_{\Omega_1}$, $\mathbf{T}_\Lambda = \mathbf{T}_{\Omega_d}$, then model (5) takes exactly the same form as model (1). And likewise, the full likelihood and the marginal likelihood (if we assume the same priors as in [Section 2.2](#)) are precisely given by (2) and (6), respectively.

Similar to the matrix-variate case, we could fit any $(d+h)$ -dimensional data using a lower d -dimensional graphical model via vectorizing, for example, the first h dimensions. However, this is not computationally feasible for the high-dimensional applications in which we are interested. One may want to solve this issue by imposing a decomposable graph on the vectorized data. The following corollary shows why it is impractical.

Corollary 1. Suppose $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_h$ are decomposable graphs. The Kronecker product, $\mathcal{G}_{1:h} = \mathcal{G}_1 \otimes \mathcal{G}_2 \otimes \dots \otimes \mathcal{G}_h$, is decomposable if and only if at least $h-1$ graphs are either complete or disconnected with complete components.

This sufficient and necessary condition for the decomposability of the general multi-dimensional graph can be obtained by applying [Proposition 1](#), recursively. The corollary implies that

we must restrict $h-1$ graphs of $\{\mathcal{G}_1, \dots, \mathcal{G}_h\}$ to be either complete or disconnected with complete components in order for $\mathcal{G}_{1:h} = \mathcal{G}_1 \otimes \mathcal{G}_2 \otimes \dots \otimes \mathcal{G}_h$ to be decomposable, which is a very restrictive modeling assumption.

5. Posterior Inference

In this section, we propose two efficient computational strategies based on collapsed MCMC algorithms (Liu 1994). We discuss these algorithms for directed mGGMs and provide extensions to undirected mGGMs, hybrid mGGMs, and aGGMs in Supplementary Material C.

5.1. Full Conditionals and the Marginal Likelihood

At each step of an MCMC algorithm, a new value is drawn from a full conditional distribution that can be either in a closed form or replaced by a computationally more expensive Metropolis-Hastings step. The full conditional distribution for $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ has a closed form, which is provided in the Appendix, and the full conditional for $(\mathbf{B}_\Omega, \mathbf{T}_\Omega)$ can be similarly obtained. When the focus is on model selection, model parameters are usually integrated out to increase the computational efficiency. The row and column covariance/precision matrices of the matrix-variate Gaussian distribution cannot be integrated out simultaneously because the prior of the row precision matrix is no longer conjugate to the marginal likelihood after integrating out the column precision matrix, and vice versa. Hence, we can only marginalize over either $(\mathbf{B}_\Omega, \mathbf{T}_\Omega)$ or $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$, but not both. Suppose we observe n samples of \mathbf{Z} and arrange the data into a $q \times p \times n$ array, denoted by $\underline{\mathbf{Z}}$. The marginal likelihood after integrating out $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ is given by

$$\begin{aligned} \ell_\Omega(\mathbf{B}_\Omega, \mathbf{T}_\Omega, \mathbf{\Gamma}_\Omega, \mathbf{\Gamma}_\Lambda | \underline{\mathbf{Z}}) \\ = (2\pi)^{-\frac{nm}{2}} \left(\prod_{k=1}^q \mathbf{T}_\Omega^{(k)} \right)^{-\frac{np}{2}} \prod_{j=1}^p |\mathbf{\Sigma}_\Lambda^{(j)}|^{\frac{1}{2}} |\mathbf{\Pi}_\Lambda^{(j)}|^{-\frac{1}{2}} \\ \times \frac{\Gamma(\alpha_\Lambda^{(j)} + \frac{nq}{2}) (\beta_\Lambda^{(j)})^{\alpha_\Lambda^{(j)}}}{\Gamma(\alpha_\Lambda^{(j)}) (\beta_\Lambda^{(j)} + d_\Lambda^{(j)})^{\alpha_\Lambda^{(j)} + \frac{nq}{2}}} \end{aligned} \quad (6)$$

with $\mathbf{\Sigma}_\Lambda^{(j)}$ and $d_\Lambda^{(j)}$ as defined in the Appendix. The marginal likelihood $\ell_\Lambda(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda, \mathbf{\Gamma}_\Lambda, \mathbf{\Gamma}_\Omega | \underline{\mathbf{Z}})$ can be similarly obtained.

5.2. Metropolis-Within-Gibbs Sampler: Unbalanced Dimensions

When q (dimension of \mathcal{G}_Ω) and p (dimension of \mathcal{G}_Λ) are unbalanced, for example, $q \ll p$ and p is moderately large, we can exploit the marginalization of the likelihood over the continuous parameters, $\mathbf{B}_\Lambda, \mathbf{T}_\Lambda$, which yields a more efficient MCMC algorithm than full-likelihood-based MCMC. The sampling scheme goes as follows.

Algorithm Metropolis-within-Gibbs Sampler (MGS)

- (I) Update $\mathbf{\Gamma}_\Lambda$ using the Metropolis-Hastings (M-H) algorithm. For $j = 1, \dots, p$, a new $\mathbf{\Gamma}_\Lambda^{(j, j+1:p)}$ is proposed in two possible ways with equal probability: (1) randomly

swapping a zero and one; (2) randomly choosing one element and switching it on (off) if it was off (on).

(II) Update ρ_Λ from the conditional distribution

$$\rho_\Lambda | \Gamma_\Lambda \sim \text{Beta} \left(a_{\rho_\Lambda} + \sum_{j=1}^{p-1} n_\Lambda^{(j)}, b_{\rho_\Lambda} \frac{p(p-1)}{2} - \sum_{j=1}^{p-1} n_\Lambda^{(j)} \right).$$

(III) Update $(\mathbf{B}_\Omega, \Gamma_\Omega)$ jointly by using the M-H algorithm.

The new Γ_Ω is proposed in the same way as Γ_Λ .

(IV) Update $(\mathbf{B}_\Omega, \mathbf{T}_\Omega)$ by using the M-H algorithm.

(V) Update ρ_Ω from the conditional distribution

$$\rho_\Omega | \Gamma_\Omega \sim \text{Beta} \left(a_{\rho_\Omega} + \sum_{k=1}^{q-1} n_\Omega^{(k)}, b_{\rho_\Omega} + \frac{q(q-1)}{2} - \sum_{k=1}^{q-1} n_\Omega^{(k)} \right).$$

This Metropolis-within-Gibbs sampling (MGS) scheme works efficiently when the assumption $q \ll p$ is met. However, Step (III) can be inefficient when q is moderately large possibly because the discrete variable Γ_Ω has to be sampled together with the continuous variable \mathbf{B}_Ω , which yields a low acceptance rate. Hence, we develop a more efficient algorithm in the next section to address this issue.

5.3. Partially Collapsed Gibbs Sampler: Balanced Dimensions

When q and p are both moderately large, a better computational strategy can be achieved if $\Gamma_\Lambda, \Gamma_\Omega$ are sampled separately from $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda), (\mathbf{B}_\Omega, \mathbf{T}_\Omega)$. To implement such a strategy we need to overcome two main difficulties: (a) we are not able to analytically integrate out both $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ and $(\mathbf{B}_\Omega, \mathbf{T}_\Omega)$, which prevents us from sampling from $p(\Gamma_\Lambda, \Gamma_\Omega | \mathbf{Z})$; and (b) generally, the set of conditional distributions of a Gibbs sampler has to be functionally compatible in the sense that it has to come from the same (joint) distribution. This impedes us, for example, sampling Γ_Λ from $p(\Gamma_\Lambda | \mathbf{Z}, \mathbf{B}_\Omega, \Gamma_\Omega, \mathbf{T}_\Omega)$ and sampling Γ_Ω from $p(\Gamma_\Omega | \mathbf{Z}, \mathbf{B}_\Lambda, \Gamma_\Lambda, \mathbf{T}_\Lambda)$. van Dyk and Park (2008) proposed a partially collapsed Gibbs sampler (PCGS), which is useful in this case to better exploit the marginalization. The PCGS relies on a set of functionally incompatible conditional distributions (i.e., some conditional distributions might come from the marginal distribution so that the joint distribution of this set of conditional distributions is undefined), which would generally lead to unknown convergence properties for the ordinary Gibbs sampler. However, carefully using the three operations of marginalization, permutation, and trimming, the resulting PCGS is guaranteed to have a known stationary distribution and usually converges faster than an MCMC algorithm on the entire set of parameters (van Dyk and Jiao 2015). Our sampler consists of two symmetric parts defined by the following operator.

Operator $\mathcal{S}(\Lambda | \Omega, \mathbf{Z})$

(I) Update Γ_Λ . This is the same as Step (I) of the algorithm MGS (with marginal likelihood ℓ_Ω).

(II) Update ρ_Λ . This is the same as Step (II) of the algorithm MGS.

(III) Update $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ from its full conditional.

The following two steps define the sampling algorithm.

Algorithm Metropolis-within-Partially-Collapsed-Gibbs Sampler (MPCGS)

(I) $\mathcal{S}(\Lambda | \Omega, \mathbf{Z})$

(II) $\mathcal{S}(\Omega | \Lambda, \mathbf{Z}^T)$

The procedure to obtain the MPCGS is given in Supplementary Material D.

5.4. Model Selection

We select the model with the highest (joint) posterior probability for undirected graphs. However, when the dimension of the (directed) graph is large, the posterior probability of any model is extremely small and consequently many models would have similar posterior probabilities. Hence, we resort to the median probability model (Barbieri and Berger 2004) in this case: an edge is selected if its marginal posterior probability is greater than the threshold 0.5.

6. Simulation Studies for Matrix-Variate Graphs

We empirically evaluate the performance of our proposed approaches in three subsections: directed mGGMs (Section 6.1), decomposable mGGMs (Section 6.2), and nondecomposable mGGMs (Section 6.3). We provide an additional simulation study of directed and undirected aGGMs in Supplementary Material E.

Suppressing the subscripts, the hyperparameters are objectively specified as $\delta = 3, \tau = 5, (a, b) = (0.5, 0.5)$ for all graphs (sensitivity analyses to these hyperparameter settings are provided at the end of Section 6.3). To summarize the simulation results, we calculate the true positive rate ($\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$), false discovery rate ($\text{FDR} = \text{FP}/(\text{TP} + \text{FP})$) and Matthews correlation coefficient (MCC):

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

where TP, TN, FP, and FN stand for the true positives, true negatives, false positives, and false negatives, respectively. MCC measures the quality of the binary classification and takes values between -1 (total disagreement) and $+1$ (perfect classification). A value of 0 suggests that using the classifier is no better than tossing a coin. MCC is a balanced measure and is relatively robust with respect to the different class sizes, which is useful in high-dimensional sparse graphs because the number of negatives (missing edges) is much larger than that of positives (present edges). To evaluate the performance with varying thresholds (marginal probability cut-offs for Bayesian approaches and penalty parameters for non-Bayesian approaches), we calculate the area under the precision-recall (TPR vs. 1-FDR) curve (AUPRC) with FDR controlled at less than 20%. To evaluate the performance of estimating the precision matrices, we let Δ be the generic notation of the

difference between the estimated precision matrix and the true precision matrix and then compute the Frobenius norm of this difference, denoted by $\|\Delta\|_F$. We apply a nonparametric Wilcoxon–Mann–Whitney two-sample rank-sum test (Wilcoxon 1945) on each operating characteristic with significance level $\alpha = 0.05$ to assess the significance of the difference between mDAG and competing methods.

6.1. Case I: Directed MGGMs

In this simulation study, we evaluate the performance of mDAG applied to directed mGGMs. Since we are not aware of any alternative approach in the literature, we take two naive methods as benchmark alternatives for comparison. The first approach estimates the row and column DAGs independently. The row DAG is constructed by treating the p columns as independent samples; likewise, the column DAG is constructed by treating the q rows as independent samples. Then, a multivariate DAG method is applied to learn the row DAG: given an ordering of the rows, the row DAG can be written as a system of linear regression, as we did in Section 2.1. We use LASSO (Tibshirani 1996), a common shrinkage and variable selection method, on each regression to learn the structure. The column DAG is learned in the same manner. We label this approach as Ind+LASSO. The second method simply applies the multivariate DAG approach to vectorized data $\mathbf{Y} = \text{vec}(\mathbf{Z})$, which we call Vec+LASSO. This approach ignores the inherent structure information of the matrix-variate data. The shrinkage parameter of LASSO is chosen by 10-fold cross-validation for both methods.

In this study, we consider four scenarios with different dimensions and sparsities:

- $D_1: q = 30, p = 30, d_\Omega = 1/20, d_\Lambda = 1/5$
- $D_2: q = 5, p = 50, d_\Omega = 1/2, d_\Lambda = 1/10$

- $D_3: q = 5, p = 500, d_\Omega = 1/2, d_\Lambda = 1/1000$
- $D_4: q = 5, p = 1000, d_\Omega = 1/2, d_\Lambda = 1/5000$,

where d_Ω and d_Λ are the measures of sparsity (i.e., ratio of the number of true edges over the number of all possible edges) of graphs \mathcal{G}_Ω and \mathcal{G}_Λ , respectively. For each edge present in the graph, the corresponding entry of \mathbf{B}_Ω and \mathbf{B}_Λ is set randomly from $\pm U(2.0, 2.2)$ and $\pm U(0.2, 0.4)$. The large values in \mathbf{B}_Ω should undermine the performance of Ind+LASSO on \mathcal{G}_Λ since the columns are far from independent, while the small values in \mathbf{B}_Λ provide weak signals. The vectorized data \mathbf{Y} with sample size $n = 100$ is then generated by solving the system of equations $\mathbf{B}\mathbf{Y} = \mathbf{E}$, where $\mathbf{B} = \mathbf{B}_\Lambda \otimes \mathbf{B}_\Omega$ and \mathbf{E} is the error term drawn from $N(0, \mathbf{I}_m)$. We center the data \mathbf{Y} and run the MCMC for 5,000 iterations, discarding the first 500 iterations as burn-in. We repeat our analysis 50 times with the graph structures fixed.

We tabulate the operating characteristics in Table 1 with the bold numbers indicating mDAG significantly (with respect to Wilcoxon–Mann–Whitney test) outperforms both competing methods and the underlined numbers indicating mDAG significantly outperforms one of the competing methods. We observe that mDAG is superior to both naive approaches for all scenarios. Vec+LASSO suffers from both poorly detecting the true positives and controlling the false discoveries, especially when the dimension is higher. For example, in D_3 ($p = 500$), mDAG has much higher TPR (0.955) of $\mathcal{G}_\Phi = \mathcal{G}_\Lambda \otimes \mathcal{G}_\Omega$ than Vec+LASSO (0.580). And yet, mDAG shows an FDR control (0.096) that is much lower than that of Vec+LASSO (0.944). We suspect these significant differences might be due to Vec+LASSO ignoring the inherent structure of the matrix-variate data. Ind+LASSO, on the other hand, performs better than Vec+LASSO in terms of TPR because it takes advantage of the structural information. However, the FDR is still very high, partly because it

Table 1. Directed matrix-variate graphs.

		mDAG			Ind+LASSO			Vec+LASSO
		\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Φ	\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Φ	\mathcal{G}_Φ
D_1	TPR	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.289(0.004)
	FDR	0.001(0.003)	0.000(0.000)	0.001(0.003)	0.716(0.017)	0.927(0.006)	0.953(0.005)	0.862(0.008)
	MCC	0.999(0.002)	1.000(0.000)	1.000(0.001)	0.322(0.033)	0.153(0.020)	0.185(0.013)	0.185(0.007)
	$\ \Delta\ _F$	0.691(0.226)	2.013(0.856)	15.456(1.553)	8.059(0.010)	25.711(0.100)	298.764(0.204)	260.770(0.363)
	AUPRC	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.846(0.177)	1.000(0.000)	0.024(0.003)	0.127(0.014)
D_2	TPR	0.969(0.014)	1.000(0.000)	<u>0.974(0.012)</u>	0.974(0.015)	1.000(0.000)	0.978(0.013)	0.370(0.010)
	FDR	0.001(0.003)	0.033(0.067)	0.022(0.042)	0.813(0.011)	0.454(0.025)	0.848(0.010)	0.793(0.009)
	MCC	0.982(0.008)	0.963(0.074)	0.975(0.023)	0.301(0.021)	0.270(0.138)	0.326(0.016)	0.228(0.008)
	$\ \Delta\ _F$	1.549(0.184)	1.163(0.409)	22.663(2.050)	9.707(0.017)	12.040(0.088)	176.058(0.206)	156.816(0.631)
	AUPRC	0.995(0.006)	0.950(0.137)	0.961(0.088)	0.487(0.073)	1.000(0.000)	0.052(0.006)	0.164(0.005)
D_3	TPR	0.869(0.024)	1.000(0.000)	<u>0.955(0.008)</u>	0.986(0.009)	1.000(0.000)	0.995(0.003)	0.580(0.003)
	FDR	0.004(0.005)	0.110(0.080)	0.096(0.069)	0.998(0.000)	0.450(0.021)	0.996(0.000)	0.944(0.008)
	MCC	0.930(0.013)	0.879(0.088)	0.928(0.036)	0.028(0.001)	0.297(0.113)	0.051(0.001)	0.178(0.013)
	$\ \Delta\ _F$	2.429(0.098)	0.596(0.194)	39.919(2.406)	19.431(0.198)	12.800(0.039)	400.581(3.881)	364.324(0.518)
	AUPRC	0.923(0.020)	0.842(0.204)	0.830(0.173)	0.186(0.045)	1.000(0.000)	0.001(0.000)	0.533(0.001)
D_4	TPR	0.760(0.035)	1.000(0.000)	<u>0.960(0.006)</u>	0.992(0.008)	1.000(0.000)	0.999(0.001)	0.691(0.002)
	FDR	0.011(0.011)	0.097(0.083)	0.093(0.078)	1.000(0.000)	0.460(0.028)	0.999(0.000)	0.971(0.004)
	MCC	0.867(0.020)	0.894(0.091)	0.932(0.041)	0.006(0.000)	0.237(0.158)	0.023(0.001)	0.139(0.010)
	$\ \Delta\ _F$	3.066(0.111)	0.523(0.168)	50.309(3.150)	40.777(0.492)	12.896(0.028)	836.090(9.653)	499.214(0.487)
	AUPRC	0.816(0.033)	0.917(0.168)	0.884(0.153)	0.082(0.037)	1.000(0.000)	0.000(0.000)	0.669(0.000)

NOTES. Operating characteristics for mDAG, Ind+LASSO and Vec+LASSO. The bold numbers indicate mDAG significantly (with respect to Wilcoxon–Mann–Whitney test at significance level $\alpha = 0.05$) outperforms both competing methods and the underlined numbers indicate mDAG significantly outperforms one of the competing methods. The standard deviation of each entry is given within parentheses.

ignores the dependencies between rows (columns) when estimating the column (row) DAG. Especially in high-dimensional settings, for example, $D_4(p = 1000)$, the FDRs of \mathcal{G}_Λ are 0.011 and 1.000 for mDAG and Ind+LASSO, respectively. The trade-off between TPR and FDR strongly favors mDAG, which can be seen from the substantial difference in the balanced measure MCC (e.g., 0.867 and 0.006 for \mathcal{G}_Λ of mDAG and Ind+LASSO in D_4 , respectively). AUPRC presents a similar pattern except that Ind+LASSO has a slightly higher AUPRC than mDAG for the smaller graph with five nodes. We do not report the area under the ROC curve because it is not a suitable measure in our case when the class distribution is highly skewed (the ratio of true negatives and positives is up to 4999:1). A large increase in false positives results in a very small increase in the false positive rate and therefore it tends to be excessively optimistic (Davis and Goadrich 2006). MDAG also performs the best in terms of precision estimation as its $\|\Delta\|_F$ is uniformly lower than those of Ind+LASSO and Vec+LASSO.

6.2. Case II: Decomposable Undirected MGGMs

Here, we examine the performance of mDAG applied to decomposable mGGMs. Specifically, we consider three decomposable graphs:

- U_1 : $q = 5$, $p = 10$, $d_\Omega = 1/2$, $d_\Lambda = 1/3$
- U_2 : $q = 5$, $p = 50$, $d_\Omega = 1/2$, $d_\Lambda = 1/10$
- U_3 : $q = 30$, $p = 30$, $d_\Omega = 1/20$, $d_\Lambda = 1/5$.

Given the true graph structures (ordered with respect to the reverse perfect ordering), we first generate the upper triangular matrices \mathbf{B}_Ω and \mathbf{B}_Λ by setting each nonzero entry (edges present in the true graphs) at ± 0.4 . By Proposition 1, the products $\tilde{\Omega} = \mathbf{B}_\Omega' \mathbf{B}_\Omega$ and $\tilde{\Lambda} = \mathbf{B}_\Lambda' \mathbf{B}_\Lambda$ have the same zero patterns as \mathbf{B}_Ω and \mathbf{B}_Λ and hence are consistent with the graph structures. Then, we scale the diagonal elements to one: $\Omega_{ij} = \tilde{\Omega}_{ij} / \sqrt{\tilde{\Omega}_{ii} \tilde{\Omega}_{jj}}$ and $\Lambda_{ij} = \tilde{\Lambda}_{ij} / \sqrt{\tilde{\Lambda}_{ii} \tilde{\Lambda}_{jj}}$ so that the negative values of the off-diagonal entries have a similar interpretation as a partial correlation in the multivariate GGM model. In our setting, the nonzero partial correlations are concentrated around ± 0.3 . Finally, we sample the vectorized data \mathbf{Y} with sample size $n = 100$ from $N(\mathbf{0}, \Lambda^{-1} \otimes \Omega^{-1})$.

We compare our approach with the state-of-the-art Bayesian mGGM method DLR (Dobra, Lenkoski, and Rodriguez 2011) and the non-Bayesian method LT (Leng and Tang 2012). DLR imposes G-Wishart priors on the row and column precision matrices that are not restricted to be decomposable. Their approach is based on the full likelihood since the marginal likelihood for the matrix-variate graph is not available in a closed form. We run their MCMC algorithm for 10,000 iterations and discard the first 1,000 iterations as burn-in. LT uses a penalized likelihood approach that penalizes row and column precisions by a LASSO or SCAD penalty (Fan and Li 2001). Since the penalized log-likelihood is only conditionally convex, an iterative algorithm is required. We run LT (the algorithm was kindly provided by the authors) with the SCAD penalty because Leng and Tang (2012) showed that the SCAD penalty consistently outperformed the LASSO penalty in their simulations. In the original article, the penalty parameter was chosen by using a separate test dataset that was the same size as the training dataset

in simulations and via cross-validation in real data analysis. The former would be unfair since both Bayesian approaches do not require an additional test dataset. Hence, we tune the penalty parameter by 10-fold cross-validation.

We present in Table 2 the operating characteristics that correspond to the three scenarios U_1 , U_2 , and U_3 . The bold and the underlined numbers have the same interpretation as Table 1. In general, mDAG is superior to the competing methods. In terms of learning the graph structures, mDAG has very good control of FDR while retaining high TPR at the same time. In U_1 , where both dimensions are low ($q = 5$, $p = 10$), mDAG and DLR perform better than LT, particularly in terms of FDR and MCC, with mDAG still achieving better performance than DLR. For instance, the FDRs of \mathcal{G}_Λ are 0.000, 0.184, and 0.561 for mDAG, DLR, and LT, respectively. The tendency of the penalized likelihood approach to produce dense networks and hence have higher FDR has been reported by others (Fitch, Jones, and Massam 2014; Liang, Song, and Qiu 2015). As the dimension increases, the superiority of mDAG is even more evident. For example, in U_3 ($q = 30$, $p = 30$), the MCCs of \mathcal{G}_Λ are 0.992, 0.463, and 0.470 for mDAG, DLR, and LT, respectively. AUPRC is consistent with other measures and again shows the great performance of mDAG. In terms of estimating the precision matrices, mDAG is the most accurate (measured in $\|\cdot\|_F$) among all the competing methods. For example, in U_2 ($q = 5$, $p = 50$), $\|\Delta\|_F$ of Λ are 1.295, 2.649, and 2.334 for mDAG, DLR, and LT, respectively.

6.3. Case III: Nondecomposable Undirected MGGMs.

Although in theory mDAG is only valid for decomposable graphs, in this simulation study we test how well mDAG fits nondecomposable matrix-variate data compared to unrestricted mGGM methods.

We consider three scenarios, U_1^* , U_2^* , and U_3^* which are the same as U_1 , U_2 , and U_3 except that the graphs we currently consider are nondecomposable. Given graph \mathcal{G}_Ω , the precision matrix Ω is generated as follows. Only Step 1 is needed if we do not need to fix $U_{qq} = 1$.

- Step 1. Let Ω be the identity matrix \mathbf{I}_q and draw Ω_{ij} from $\pm U(0.1, 0.3)$ if edge (i, j) is present in \mathcal{G}_Ω .
- Step 2. Set $U_{qq} = 1$ where $U = \Omega^{-1}$.
- Step 3. Invert U and round the resulting Ω to the fourth digit.
- Step 4. Make \mathcal{G}_Ω consistent with Ω .

Since the matrix generated from Step 1 is not necessarily positive definite, we might need to repeat the above procedure until we get a positive definite precision matrix with desired sparsity. The nonzero off-diagonal elements of the resulting Ω are concentrated around ± 0.2 . The precision matrix Λ can be generated in the same manner. For Scenario U_1^* , the row graph \mathcal{G}_Ω (Figure 1(a) of Supplementary Material) is a circle with five nodes and the column graph \mathcal{G}_Λ (Figure 1(b) of Supplementary Material) is a single noncomplete prime component with 10 nodes. The column graph is very similar to the graph used in Jones et al. (2005) to test how well a decomposable model can recover a nondecomposable graph. The nonzero off-diagonal elements of Ω are set to 0.4, a high value that makes this an ideal scenario for the competing methods to perform well.

Table 2. Decomposable undirected matrix-variate graphs.

		mDAG		DLR		LT	
		\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Λ	\mathcal{G}_Ω
U_1	TPR	0.996(0.028)	1.000(0.000)	0.992(0.022)	1.000(0.000)	1.000(0.000)	1.000(0.000)
	FDR	0.000(0.000)	0.000(0.000)	0.184(0.101)	0.000(0.000)	0.561(0.040)	0.282(0.126)
	MCC	0.997(0.021)	1.000(0.000)	0.840(0.096)	1.000(0.000)	0.391(0.076)	0.630(0.201)
	$\ \Delta\ _F$	0.421(0.156)	0.172(0.071)	0.465(0.138)	0.214(0.108)	0.704(0.122)	0.392(0.112)
	AUPRC	1.000(0.000)	1.000(0.000)	0.991(0.027)	1.000(0.000)	0.845(0.108)	1.000(0.000)
U_2	TPR	0.968(0.036)	1.000(0.000)	0.962(0.049)	1.000(0.000)	0.997(0.004)	1.000(0.000)
	FDR	0.013(0.019)	0.000(0.000)	0.870(0.003)	0.000(0.000)	0.793(0.009)	0.046(0.092)
	MCC	0.975(0.029)	1.000(0.000)	0.168(0.019)	1.000(0.000)	0.343(0.015)	0.947(0.108)
	$\ \Delta\ _F$	1.295(0.269)	0.154(0.043)	2.649(0.252)	0.213(0.129)	2.334(0.112)	0.349(0.048)
	AUPRC	0.966(0.047)	1.000(0.000)	0.131(0.010)	1.000(0.000)	0.407(0.059)	1.000(0.000)
U_3	TPR	0.995(0.015)	1.000(0.000)	0.940(0.045)	0.921(0.098)	0.990(0.007)	1.000(0.000)
	FDR	0.008(0.025)	0.001(0.006)	0.604(0.050)	0.367(0.139)	0.619(0.018)	0.853(0.013)
	MCC	0.992(0.025)	1.000(0.003)	0.463(0.081)	0.746(0.132)	0.470(0.023)	0.318(0.021)
	$\ \Delta\ _F$	0.459(0.189)	0.298(0.108)	0.988(0.424)	0.768(0.327)	0.521(0.088)	0.342(0.072)
	AUPRC	0.988(0.057)	1.000(0.000)	0.806(0.084)	0.834(0.138)	0.484(0.042)	1.000(0.000)

NOTES. Operating characteristics for mDAG, DLR, and LT. The bold numbers indicate mDAG significantly (with respect to Wilcoxon–Mann–Whitney test at significance level $\alpha = 0.05$) outperforms both competing methods and the underlined numbers indicate mDAG significantly outperforms one of the competing methods. The standard deviation of each entry is given within parentheses.

For scenarios U_2^* , U_3^* , the performance of the mDAG is highly competitive in terms of both learning the graph structures and estimating the precision matrices, as shown in Table 3. In U_3^* , where $p = q = 30$, we have very good FDR compared to those of the competing methods (e.g., 0.094, 0.615, and 0.861 for mDAG, DLR, and LT, respectively, in estimating \mathcal{G}_Ω). What is more promising is that mDAG does not compromise much TPR for the extremely low FDR. The TPRs of \mathcal{G}_Ω are 0.862, 0.738, and 0.904 for mDAG, DLR, and LT, respectively. As expected, in scenario U_1^* where the dimension is low and the signal is strong, the competing method DLR can almost perfectly recover the structure of \mathcal{G}_Ω (TPR = 1, FDR = 0.003) while mDAG has a higher FDR (0.286) due to the decomposability constraint. This is also reflected in the estimation of the precision matrix. For example, $\|\Delta\|_F$ of Λ are 1.224, 0.496, and 1.164 for mDAG, DLR, and LT, respectively. Although DLR and LT have higher AUPRC than mDAG in some scenarios, we want to emphasize that choosing tuning parameter or threshold for edge inclusion probabilities is part of the estimation procedure and is just as important by itself. We suspect that the poor performance of the competing methods is due to the low signal in the partial correlations, which possibly makes the iterative optimization algorithm of LT and

MCMC algorithm of DLR ineffective in finding the true graph. Regarding the estimation of the precision matrices, mDAG generally has smaller error than the Bayesian approach, DLR, and is comparable to the frequentist approach, LT.

We perform a sensitivity analysis of the hyperparameter specifications in Supplementary Material F, which shows that our approach is robust to the choice of hyperparameter values.

In Supplementary Material E, we report the additional simulation study we conducted for both directed and undirected (three-dimensional) aDAGs. We found that our aDAG approach outperforms the alternative naive methods in terms of both structure learning and precision matrix estimation.

7. Case Studies

7.1. Integrated Protein Networks for Ovarian Cancer

Protein signaling pathways that characterize protein–protein interactions are receiving increased attention, especially in the cancer community because proteins kinases that are frequently mutated in cancer are potential therapeutic targets for the development of new treatments. Many drugs that target mutated

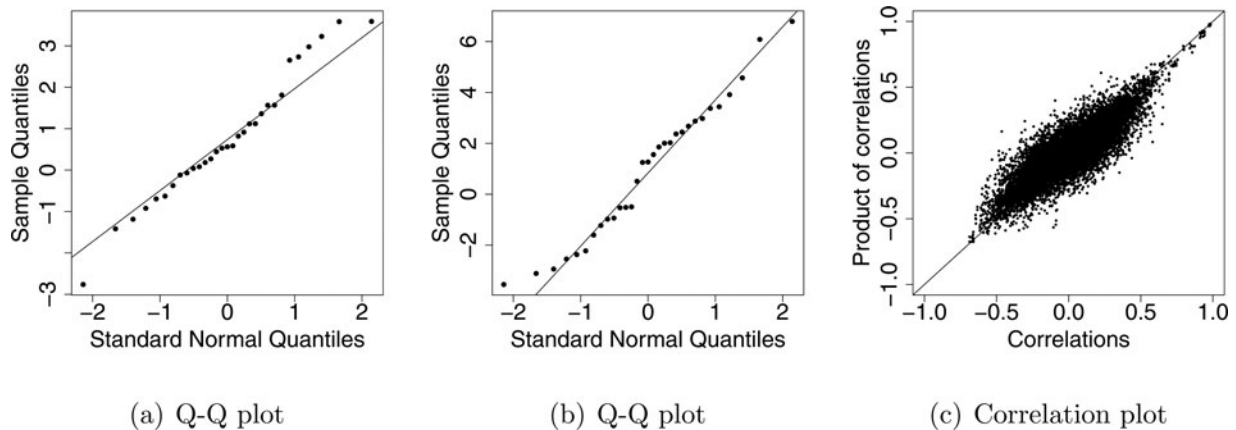


Figure 1. Empirical model checking. (a), (b) Two randomly chosen Q–Q plots of Z_{ij} against the standard normal distribution; (c) correlation plot for covariance separability.

Table 3. Nondecomposable undirected matrix-variate graphs.

		mDAG		DLR		LT	
		\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Λ	\mathcal{G}_Ω	\mathcal{G}_Λ	\mathcal{G}_Ω
U_1^*	TPR	0.684(0.098)	1.000(0.000)	0.938(0.059)	1.000(0.000)	0.998(0.013)	1.000(0.000)
	FDR	0.000(0.000)	0.286(0.000)	0.324(0.108)	0.003(0.024)	0.673(0.036)	0.229(0.106)
	MCC	0.787(0.068)	0.655(0.000)	0.714(0.103)	0.996(0.026)	0.321(0.079)	0.722(0.145)
	$\ \Delta\ _F$	1.224(0.112)	1.438(0.219)	0.496(0.123)	0.212(0.106)	1.164(0.094)	1.185(0.173)
	AUPRC	0.896(0.055)	1.000(0.000)	0.920(0.086)	1.000(0.000)	0.910(0.071)	1.000(0.000)
U_2^*	TPR	0.499(0.024)	1.000(0.000)	0.947(0.043)	1.000(0.000)	0.964(0.009)	1.000(0.000)
	FDR	0.122(0.028)	0.167(0.000)	0.870(0.003)	0.000(0.000)	0.755(0.012)	0.000(0.000)
	MCC	0.636(0.016)	0.816(0.000)	0.163(0.017)	1.000(0.000)	0.388(0.018)	1.000(0.000)
	$\ \Delta\ _F$	2.135(0.086)	0.116(0.039)	2.549(0.241)	0.236(0.136)	1.418(0.094)	0.203(0.025)
	AUPRC	0.433(0.042)	1.000(0.000)	0.127(0.011)	1.000(0.000)	0.295(0.069)	1.000(0.000)
U_3^*	TPR	0.841(0.046)	0.862(0.009)	0.964(0.039)	0.738(0.104)	0.998(0.004)	0.904(0.043)
	FDR	0.336(0.032)	0.094(0.015)	0.618(0.036)	0.615(0.230)	0.594(0.019)	0.861(0.017)
	MCC	0.673(0.040)	0.877(0.006)	0.453(0.065)	0.470(0.223)	0.502(0.024)	0.280(0.031)
	$\ \Delta\ _F$	1.278(0.266)	0.404(0.127)	1.203(0.303)	1.132(0.268)	0.736(0.105)	0.337(0.054)
	AUPRC	0.287(0.040)	0.751(0.074)	0.814(0.068)	0.638(0.231)	0.419(0.049)	0.863(0.008)

NOTES. Operating characteristics for mDAG, DLR, and LT. The bold numbers indicate mDAG significantly (with respect to Wilcoxon–Mann–Whitney test at significance level $\alpha = 0.05$) outperforms both competing methods and the underlined numbers indicate mDAG significantly outperforms one of the competing methods. The standard deviation of each entry is given within parentheses.

protein kinases have shown dramatic clinical benefits (Davies, Hennessy, and Mills 2006).

In this case study, we are interested in ovarian cancer, the fifth leading cause of cancer deaths among U.S. women (Siegel et al. 2014). The overall cure rate is only $\sim 30\%$ (Bast, Hennessy, and Mills 2009). One of the critical signaling pathways for ovarian cancer is the AKT/PI3K pathway, which is activated in $\sim 70\%$ of ovarian cancers (Bast, Hennessy, and Mills 2009). In this study, reverse-phase protein array (RPPA), a relatively new technology (Tibes et al. 2006), was used to assay $p = 50$ protein expressions in the AKT/PI3K pathway from $n = 31$ ovarian cancer cell lines that were grown under $q = 3$ culture conditions: (a) media replete with growth factor; (b) media starved of growth factor; and (c) media starved, then acutely stimulated with growth factor. Details of the experimental procedure can be found in Baladandayuthapani et al. (2014). Our goal is to integrate the protein expression data and capture the common structure of proteins across different culture conditions.

The two assumptions of our model are normality and covariance separability. Here, we provide empirical evidence that both assumptions are appropriate for this dataset. Two randomly chosen quantile–quantile (Q–Q) plots of Z_{ij} against the standard normal distribution are shown in Figures 1(a) and (b). We also perform the Kolmogorov–Smirnov test on all the elements in Z . The average p -value is 0.56, which together with the Q–Q plots suggests a good justification for the normality assumption. Next, we examine whether the covariance structure is separable with respect to the Kronecker product:

$$\text{cor}(Z_{ij}, Z_{i'j'}) = \text{cor}(Z_{ij}, Z_{i'j})\text{cor}(Z_{ij}, Z_{ij'}). \quad (7)$$

In Figure 1(c), we plot the right-hand side against the left-hand side of Equation (7). The strong linear trend (correlation 0.87) justifies the separability of the covariance structure.

We can define a prior ordering for the proteins from manually curated AKT/PI3K signaling pathways (<http://www.genome.jp/kegg/>). We apply the mDAG as a hybrid of the directed graph on proteins and undirected graph on

culture conditions. We run two separate MCMCs, each with 50,000 iterations. The Markov chains exhibit good mixing and are deemed to converge by MCMC diagnostics (details provided in Supplementary Material G). The computation time is ~ 55 minutes on a 3.5 GHz Intel Core i7 processor. We combine the two chains and discard the first 10% of the iterations of each chain as burn-in. The inferred network of the culture conditions is a complete graph, that is, all the conditions are correlated with each other. This is expected since the environment of the culturing is interrelated by the experimental design. In Figure 2, we present the inferred protein network, with solid arrows representing activation while dashed arrows represent inactivation. Node size is proportional to its degree and line width is proportional to its marginal posterior probability. We found 70 regulations (24 inactivations and 46 activations) which are ranked according to their marginal posterior probabilities in Supplementary Material H. Some of these connections are confirmed from previous biological studies. For example, we found that PTEN inactivates Akt.pT308. In fact, the up-regulation of tumor suppressor PTEN inhibits Akt in ovarian cancer cells (Selvendiran et al. 2007), which together with downstream regulations induces cell apoptosis (Li et al. 1998). We also observed that MAPK.pT202.Y204 and p38 inactivate p70S6K.pT389, while PKCa and PDK1 activate p70S6K.pT389. All of them are known to participate in phosphorylation of p70S6K (Mukhopadhyay et al. 1992; Romanelli et al. 1999). In addition, we identified four hub/driver proteins (represented as hexagons in Figure 2): p70S6K.pT389, cJun, HER3 and Stat3, with degrees 10, 9, 6, and 6, respectively. The most highly connected protein, p70S6K.pT389, is known to phosphorylate the S6 ribosomal protein (Chung et al. 1992), which was also detected in our network (p70S6K \rightarrow S6.pS235.S236). It participates in the transduction of signals that regulate motility and the invasion of ovarian cancer cells through the PI3K–Akt–p70S6K signaling pathway, a key factor in metastasis (Bast, Hennessy, and Mills 2009). Hub protein Stat3, which contributes to tumor proliferation, apoptosis and angiogenesis, is activated in more than 70% of ovarian cancers and is negatively associated with

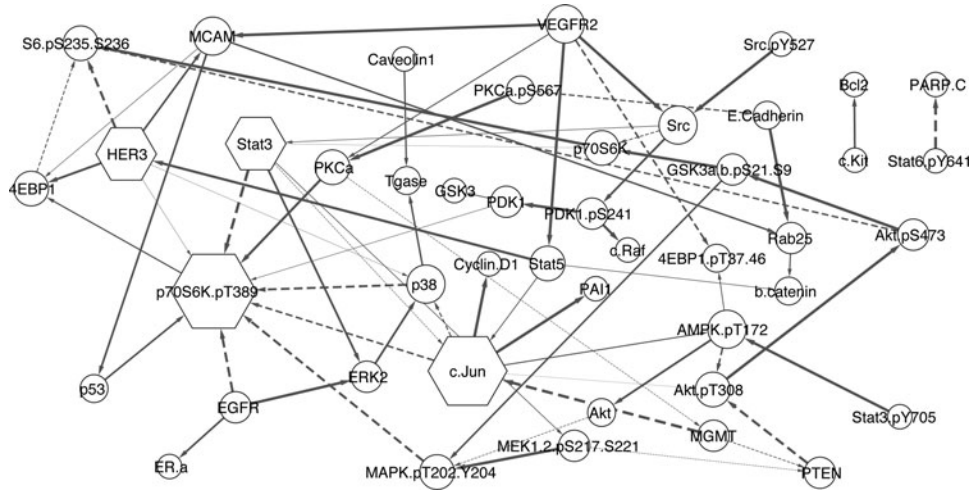


Figure 2. Protein network for ovarian cancer. Solid arrows represent activation while dashed arrows represent inactivation. Node size is proportional to its degree and line width is proportional to its marginal posterior probability. Hexagons are 4 hub proteins.

the patient's survival time (Rosen et al. 2006; Bast, Hennessy, and Mills 2009). Another interesting hub protein is HER3, a member of the epidermal growth factor receptor (EGFR) family, which is also commonly overexpressed in ovarian cancer and negatively associated with patients' survival times (Davies et al. 2014). HER3, together with other EGFR family members (e.g., HER2), activates the AKT/PI3K signaling cascade (Tanner et al. 2006). The less well-studied hub protein, c.Jun, as well as novel regulatory relationships need further biological validations.

In addition, we investigate what type of inference can be obtained if a decomposable multivariate GGM is fitted on the vectorized data. Specifically, we apply two decomposable multivariate GGM methods, a Metropolis-Hastings (MH) algorithm implemented in Jones et al. (2005) and a feature-inclusion stochastic search algorithm (FINCS; Scott and Carvalho 2008) to the vectorized data for comparison, assuming the graph of the culture conditions is complete (as inferred by mDAG). However, both methods detect few edges (details can be found in Supplementary Material H); this result is not unexpected, since the number of nodes is increased from 50 to 150 and the sample size is only 31 due to the vectorization of the data. By contrast, mDAG effectively accounts for the matrix structure of the data and effectively uses culture conditions as samples when estimating the protein network; in this case, the effective sample size is on the order of $q \times n = 3 \times 31 = 93 > p = 50$.

7.2. SEER U.S. Cancer Mortality Counts

Systematic collection of cancer mortality data is critical to the cancer research community. Such information allows researchers and policy makers to monitor mortality trends over time, which is crucial to the identification of effective cancer prevention strategies. Also, investigating the mortality distribution over different regions is useful to explore potential common risk factors for specific cancers. In this article, we consider the U.S. cancer mortality data collected through the National Cancer Institute as the Surveillance, Epidemiology, and End Results (SEER) Program (<http://seer.cancer.gov/data/>). We focus on seven common cancers: stomach, rectum, colon excluding rectum, pancreas, lung, breast, and prostate and exclude other

cancers due to missing values. Overall, cancer mortality rates have been decreasing since 1991 as a consequence of effective cancer prevention strategies, including cancer screening and tobacco control (Siegel et al. 2014). In our analysis, we restrict our attention to a recent 12-year period (2000–2011) to exclude the possible lingering effects of the tobacco epidemic of the 20th century. The northeast region (9 states) is of particular interest as it has the highest cancer incidence rate across the country (U.S. Cancer Statistics Working Group 2014).

Random effect array-variate Poisson (REAP) log-linear model. The Poisson log-linear model is often used to model count observations. However, in practice, count data often exhibit higher variance than the mean, which contradicts the assumption of a Poisson distribution. This phenomenon is also known as overdispersion. A common remedy is to add random effects to the model that explain the extra variability in the data. In addition, the random effects are useful for modeling the structural dependence between responses (Breslow and Clayton 1993). Let $N_{ijk} | \lambda_{ijk} \sim \text{Poisson}(\lambda_{ijk})$ denote the mortality counts for cancer k in state i and year j . We choose the canonical link $\log(\cdot)$, which links the linear predictors to the mean λ_{ijk} :

$$\log(\lambda_{ijk}) = \log(M_{ij}) + \mu_k + Z_{ijk},$$

where the offset M_{ij} is the observed population of state i in year j , μ_k is the cancer-specific fixed effect and Z_{ijk} is the random effect. An array-variate graphical model on \mathbf{Z} will allow us to understand how cancer mortality varies across states and cancer types, and its temporal dynamic. The log-likelihood function is given by

$$\ell(\boldsymbol{\mu}, \mathbf{Z} | \mathbf{N}) \propto \sum_{i,j,k} \{ N_{ijk} (\log(M_{ij}) + \mu_k + Z_{ijk}) - M_{ij} \exp(\mu_k + Z_{ijk}) \},$$

where $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_p\}$, $\mathbf{Z} = (Z_{ijk})$ and $\mathbf{N} = (N_{ijk})$. The priors are specified as follows:

- (1) fixed intercept $\boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$;
- (2) random intercept $\mathbf{Z} \sim AN_{r \times q \times p}(0, \boldsymbol{\Theta}^{-1}, \boldsymbol{\Psi}^{-1}, \boldsymbol{\Lambda}^{-1})$; and
- (3) hyperparameters $\boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Lambda}$ take the same priors as in Section 2.2.

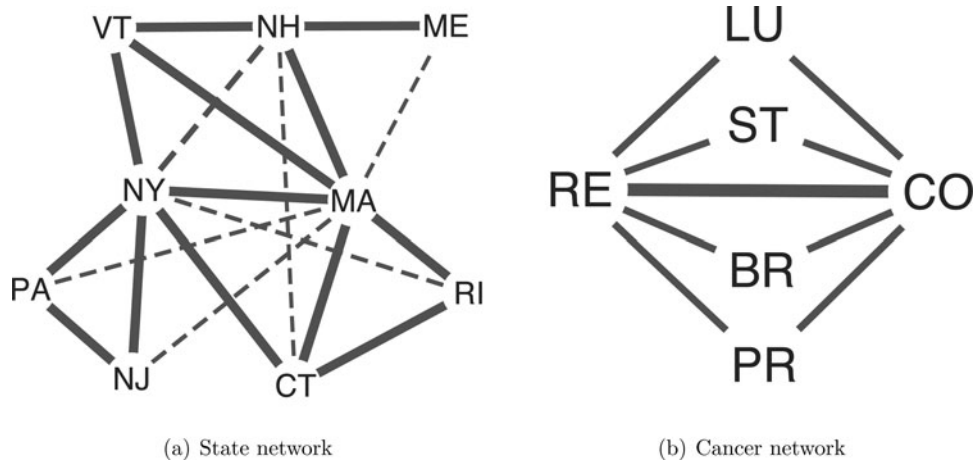


Figure 3. SEER U.S. cancer mortality data. Line width is proportional to posterior probability. (a) Solid lines represent geographically adjacent states; dashed lines nonadjacent states. (b) BR (breast), CO (colon), LU (lung), PR (prostate), RE (rectum), and ST (stomach). Pancreatic cancer is not shown as it is not connected to any other cancer.

In terms of sampling, we need only two additional steps, that is, sampling μ and Z . Then, treating the sampled Z as observation ($n = 1$), Θ , Ψ , Λ are sampled using the algorithm in Section 4. The algorithm to sample μ and Z is given in Supplementary Material I.

Dobra, Lenkoski, and Rodriguez (2011) performed a detailed analysis on similar data where they assigned a multivariate conditional autoregressive (MCAR) prior on the random effects in the Poisson log-linear model. However, there are at least two advantages of our REAP model in analyzing this dataset. First, restricted by the matrix-variate nature of the methodology, the method of Dobra, Lenkoski, and Rodriguez (2011) could not handle three-way data; hence, their investigation focused on the data from just one year (2000). Our REAP model, however, is capable of fully using the data collected over time. Second, and more important, Dobra, Lenkoski, and Rodriguez (2011) had to fix the state network in their analysis because the state network and the cancer network could not be simultaneously estimated within a random effect matrix-variate model. Fixing the state network from the neighborhood structures of states within the U.S. implies that the associations between cancer mortalities within different states are completely determined by their geographical locations. Although the populations in adjacent states might share some risk factors, such an assumption is too rigid and excludes the possibility that the populations in some distant states may share similar risk factors apart from geographical location. For example, the populations of Maine and Pennsylvania, although geographically far apart, have similar adult tobacco use (Centers for Disease Control and Prevention 2013), which is a well-known risk factor for lung cancer and many other cancers. By contrast, our REAP model does not need to fix any network structure and is able to simultaneously learn the state network, temporal network and cancer network. Intuitively, for instance, when estimating the state network with $r = 9$ nodes, we effectively (respecting the dependencies) use the second and third dimensions as samples and the effect sample size is on the order of $q \times p = 12 \times 7 = 84 > r$. Hence, the three networks are jointly estimable, especially when the true networks are presumably sparse. Yet, from an application perspective, spatial and temporal prior information can be easily incorporated into our model. Let $A = (a_{ij})$ denote the adjacency matrix of the neighborhood structures of the states

within the U.S., that is, $a_{ij} = 1$ if states i and j share a common border and $a_{ij} = 0$ otherwise. The spatial information is embedded in the prior: $\Gamma_{\Theta}^{(i,j)} | \xi_{a_{ij}} \sim \text{Bernoulli}(\xi_{a_{ij}})$, where $\xi_0 \sim \text{Beta}(a_{\xi_0}, b_{\xi_0})$ with $a_{\xi_0} \ll b_{\xi_0}$ and $\xi_1 \sim \text{Beta}(a_{\xi_1}, b_{\xi_1})$ with $a_{\xi_1} \gg b_{\xi_1}$. The choice of Beta hyperparameters reflects our prior belief that the populations of adjacent states are more likely to share common cancer risk factors. A similar hierarchical prior structure can be applied to the temporal network, for which A has ones on the diagonal above the main diagonal and zeros everywhere else, which favors the associations between cancer mortalities in consecutive years over those associations in nonconsecutive years.

We ran two separate chains, each with 50,000 iterations. The MCMC diagnostics given in Supplementary Material G show good mixing. The computation time is ~ 30 hours on a 3.5 GHz Intel Core i7 processor. We combined the two chains and discarded the first 10% of iterations of each chain as burn-in. Figure 3 shows the state and cancer networks inferred by the REAP model. The line width is proportional to the posterior edge inclusion probabilities. For the state network (Figure 3(a)), the REAP model recovered all the associations among geographically adjacent states (solid lines) as well as those among some nonadjacent states (dashed lines). Some associations between the populations of nonadjacent states can be explained by common risk factors. For example, although separated by New York geographically, the populations of New Jersey (16.8% with 95% confidence interval (CI) (15.9%, 17.8%)) and Massachusetts (18.2 with 95% CI (17.3%, 19.2%)) have similar prevalence of adult cigarette use, with median prevalence of 21.2% across all states (Centers for Disease Control and Prevention 2013). The populations of New York and Rhode Island, separated by Massachusetts and Connecticut, share nearly equal prevalence of excessive alcohol use (18.3% vs. 18.2%), with overall prevalence of 17.1% (Centers for Disease Control and Prevention 2012). For the cancer network (Figure 3(b)), we observe the strongest association (the thickest edge) between colon and rectal cancers. In fact, colon and rectal cancers are almost indistinguishable genomically (The Cancer Genome Atlas 2012) and share many risk factors such as family history and obesity (Edwards et al. 2010). We plot a heatmap of the posterior edge inclusion probabilities for the temporal network in Figure 4. The colors (probabilities) decay away from the diagonal, which coincides

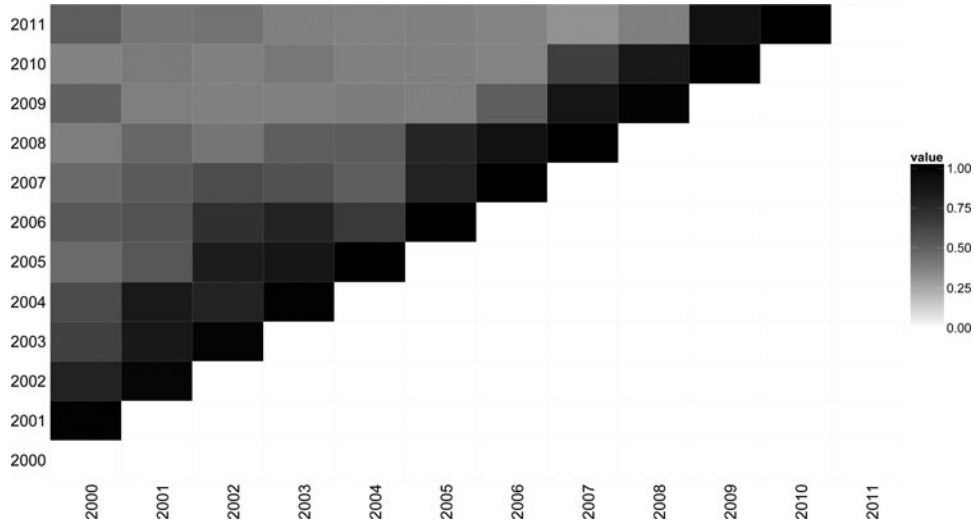


Figure 4. SEER U.S. cancer mortality data. Heatmap of posterior edge inclusion probabilities for temporal network. Darker colors correspond to higher probabilities.

with our intuition that the data from nonconsecutive years are weakly associated while those from consecutive years are closely related.

8. Discussion

We have introduced a novel class of graphical models based on LDL decomposition of precision matrices to infer networks for multi-dimensional data. The LDL decomposition implies a system of linear regressions that connects directed graphs to undirected decomposable graphs. By exploiting such connections, our unified framework can be used to model directed, undirected, and hybrid multi-dimensional graphs. Our modeling approach combined with our computational strategies result in an efficient and flexible statistical framework. Using simulation studies, we have demonstrated the superior performance of our approach in comparison with both benchmark and state-of-the-art methods. We applied our model to two datasets. First, as a matrix-variate example, we integrated RPPA protein expression data for ovarian cancer cell lines grown in three different culture conditions. Some of our findings are consistent with the literature, while others need to be validated by biological experiments. Second, we modeled the three-way SEER U.S. cancer mortality data using the REAP model. We assigned an array-variate Gaussian graphical model prior to the random effects and simultaneously estimated the cancer, temporal and state networks. We used informative priors to incorporate the temporal and spatial information. Again, some of our findings are plausible and supported by the literature.

Our formulation of an undirected graph through LDL decomposition can be thought of as one limitation of our approach as it only applies to decomposable graphs, which we would like to address in our future work. However, the decomposability assumption is not unique to our approach. In fact, it is commonly assumed in much of the recent Bayesian methodological literature on graphical models (Scott and Carvalho 2008; Carvalho and Scott 2009; Wang and West 2009) for computational efficiency because the normalizing constant of the G-Wishart prior is not available in a closed form for nondecomposable graphs. Theoretically, the posterior of decomposable GGMs will converge to minimal triangulations of the true graph (Fitch,

Jones, and Massam 2014). Empirically, our proposed approach, especially in higher dimensional settings, outperforms the competing methods that are developed without a decomposability constraint even when the true graph is nondecomposable, as we showed in Section 6.3.

Appendix: Details of Full Conditionals and Marginal Likelihood

We provide the full conditional distributions and the notations used in the marginal likelihood (6). Consider the j th and k th rows of \mathbf{B}_Λ and \mathbf{B}_Ω . Let $\mathbf{Y}_l = (Y_{l1}, \dots, Y_{ln})^T$ denote the n samples of Y_l and let $\mathbf{Y}_{l+1:m} = [\mathbf{Y}_{l+1}, \dots, \mathbf{Y}_m]$ where $l = (j-1)q + k$. Also let $\mathbf{Y}_l^{(\Lambda\Omega)}, \mathbf{Y}_l^{(\Omega)}, \mathbf{Y}_l^{(\Lambda)}$ be the submatrices of $[\mathbf{Y}_1, \dots, \mathbf{Y}_m]$, with columns corresponding to the regression coefficients $\mathbf{B}_\Lambda^{(j,j+1:p)T} \otimes \mathbf{B}_\Omega^{(k,k+1:q)T}, \mathbf{B}_\Omega^{(k,k+1:q)T}, \mathbf{B}_\Lambda^{(j,j+1:p)T}$, respectively. Then $\Sigma_\Lambda^{(j)}$ and $d_\Lambda^{(j)}$ are given by

$$\Sigma_\Lambda^{(j)} = \left[\sum_{k=1}^q \frac{\mathbf{C}_\Lambda^{(j,k)}}{T_\Omega^{(k)}} + \left(\Pi_\Lambda^{(j)} \right)^{-1} \right]^{-1}$$

$$d_\Lambda^{(j)} = \sum_{k=1}^q \frac{e_\Lambda^{(j,k)}}{2T_\Omega^{(k)}} - \frac{1}{2} \sum_{k=1}^q \frac{\mathbf{c}_\Lambda^{(j,k)T}}{T_\Omega^{(k)}} \Sigma_\Lambda^{(j)} \sum_{k=1}^q \frac{\mathbf{c}_\Lambda^{(j,k)}}{T_\Omega^{(k)}}$$

where

$$\Pi_\Lambda^{(j)} = \text{diag}(\tau_\Lambda^{(j+1)}, \dots, \tau_\Lambda^{(j)})$$

$$\mathbf{C}_\Lambda^{(j,k)} = \sum_{i=1}^n \left(\mathbf{Y}_l^{(\Lambda\Omega)T} \mathbf{e}_i \right)^{mT} \mathbf{B}_\Omega^{(k,k+1:q)T} \mathbf{B}_\Omega^{(k,k+1:q)} \left(\mathbf{Y}_l^{(\Lambda\Omega)T} \mathbf{e}_i \right)^m$$

$$+ \mathbf{Y}_l^{(\Lambda)T} \mathbf{Y}_l^{(\Lambda)} + 2 \sum_{i=1}^{p-j} \left(\mathbf{Y}_l^{(\Lambda\Omega)T} \mathbf{Y}_l^{(\Lambda)} \mathbf{e}_i \right)^{mT} \mathbf{B}_\Omega^{(k,k+1:q)T} \mathbf{e}_i^T$$

$$\mathbf{c}_\Lambda^{(j,k)} = \left(\mathbf{B}_\Omega^{(k,k+1:q)} \left(\mathbf{Y}_l^{(\Lambda\Omega)T} \mathbf{Y}_l \right)^m + \mathbf{Y}_l^T \mathbf{Y}_l^{(\Lambda)} + \mathbf{B}_\Omega^{(k,k+1:q)} \right)$$

$$\times \left(\mathbf{Y}_l^{(\Lambda\Omega)T} \mathbf{Y}_l^{(\Omega)} \mathbf{B}_\Omega^{(k,k+1:q)T} \right)^m + \mathbf{B}_\Omega^{(k,k+1:q)} \mathbf{Y}_l^{(\Omega)T} \mathbf{Y}_l^{(\Lambda)} \Big)^T$$

$$e_\Lambda^{(j,k)} = \mathbf{Y}_l^T \mathbf{Y}_l + \mathbf{B}_\Omega^{(k,k+1:q)} \mathbf{Y}_l^{(\Omega)T} \mathbf{Y}_l^{(\Omega)} \mathbf{B}_\Omega^{(k,k+1:q)T} + 2 \mathbf{B}_\Omega^{(k,k+1:q)} \mathbf{Y}_l^{(\Omega)T} \mathbf{Y}_l$$

with e_i being the unit vector with one in the i th element and zeros elsewhere. And $()^m$ is the inverse operation of vectorization (i.e., $\text{vec}(\mathbf{Z})^m = \mathbf{Z}$), which turns a vector into a matrix whose size is compatible with the matrices next to it. The full conditional distribution for $(\mathbf{B}_\Lambda, \mathbf{T}_\Lambda)$ has a closed form:

$$\mathbf{B}_\Lambda^{(j,j+1:p)T} \left| \mathbf{Z}_{(j-1)q+1:qp}, \mathbf{T}_\Lambda^{(j)}, \boldsymbol{\Pi}_\Lambda^{(j)}, \mathbf{B}_\Omega, \mathbf{T}_\Omega \sim N\left(\boldsymbol{\mu}_\Lambda^{(j)}, \mathbf{T}_\Lambda^{(j)} \boldsymbol{\Sigma}_\Lambda^{(j)}\right),$$

$$\mathbf{T}_\Lambda^{(j)} \left| \mathbf{Z}_{(j-1)q+1:qp}, \mathbf{B}_\Lambda^{(j,j+1:p)}, \mathbf{B}_\Omega, \mathbf{T}_\Omega \sim \text{IG}\left(\alpha_\Lambda^{(j)} + \frac{nq}{2}, \beta_\Lambda^{(j)} + f_\Lambda^{(j)}\right)$$

$$\text{with } \boldsymbol{\mu}_\Lambda^{(j)} = -\boldsymbol{\Sigma}_\Lambda^{(j)} \sum_{k=1}^q \frac{\mathbf{c}_\Lambda^{(j,k)}}{T_\Omega^{(k)}} \quad \text{and} \quad f_\Lambda^{(j)} = \sum_{k=1}^q \frac{1}{2T_\Omega^{(k)}} \\ (\mathbf{B}_\Lambda^{(j,j+1:p)} \mathbf{C}_\Lambda^{(j,k)} \mathbf{B}_\Lambda^{(j,j+1:p)T} + 2\mathbf{c}_\Lambda^{(j,k)T} \mathbf{B}_\Lambda^{(j,j+1:p)T} + e_\Lambda^{(j,k)}).$$

Supplemental Materials

SMGM_supp.pdf contains (A) definitions of the graph terminology used in this article, (B) proof and corollary of Proposition 1, (C) sampling algorithm for undirected and hybrid mGGMs and array-variate graphs, (D) details of partially collapsed Gibbs sampler, (E) additional simulation studies for directed and undirected aGGMs, (F) sensitivity analysis of hyperparameters, (G) MCMC convergence diagnostics, (H) additional tables and figures for RPPA ovarian cancer data analysis, (I) sampling algorithm for random effect array-variate Poisson regression, and (J) data sources.

SMGM_code.zip contains the Matlab files that implement our approaches.

Funding

V. Baladandayuthapani was partially supported by NIH grant R01 CA160736 and NSF grant DMS: 1463233. Both F. C. Stingo and V. Baladandayuthapani were partially supported by the Cancer Center Support Grant (CCSG) (P30 CA016672).

References

- Akdemir, D., and Gupta, A. K. (2011), "Array Variate Random Variables With Multiway Kronecker Delta Covariance Matrix Structure," *Journal of Algebraic Statistics*, 2, 98–113. [5]
- Allen, G. I., Tibshirani, R., et al. (2010), "Transposable Regularized Covariance Models With an Application to Missing Data Imputation," *The Annals of Applied Statistics*, 4, 764–790. [2]
- Baladandayuthapani, V., Talluri, R., Ji, Y., Coombes, K. R., Lu, Y., Hennessy, B. T., Davies, M. A., and Mallick, B. K. (2014), "Bayesian Sparse Graphical Models for Classification With Application to Protein Expression Data," *The Annals of Applied Statistics*, 8, 1443–1468. [10]
- Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *The Journal of Machine Learning Research*, 9, 485–516. [1]
- Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [6]
- Bast, R. C., Hennessy, B., and Mills, G. B. (2009), "The Biology of Ovarian Cancer: New Opportunities for Translation," *Nature Reviews Cancer*, 9, 415–428. [10]
- Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [11]
- Carvalho, C. M., and Scott, J. G. (2009), "Objective Bayesian Model Selection in Gaussian Graphical Models," *Biometrika*, 96, 497–512. [1,13]
- Carvalho, C. M., and West, M. (2007), "Dynamic Matrix-Variate Graphical Models," *Bayesian Analysis*, 2, 69–97. [1]
- Centers for Disease Control and Prevention (2012), "Vital Signs: Binge Drinking Prevalence, Frequency, and Intensity Among Adults—United

- States, 2010," *MMWR. Morbidity and Mortality Weekly Report*, 61, 14. [12]
- Centers for Disease Control and Prevention (2013), *Tobacco Control State Highlights 2012*. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention. National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. [12]
- Chung, J., Kuo, C. J., Crabtree, G. R., and Blenis, J. (1992), "Rapamycin-fkbp Specifically Blocks Growth-Dependent Activation of and Signaling by the 70 kd s6 Protein Kinases," *Cell*, 69, 1227–1236. [10]
- Davies, M., Hennessy, B., and Mills, G. B. (2006), "Point Mutations of Protein Kinases and Individualised Cancer Therapy," *Expert Opinion on Pharmacotherapy*, 7, 2243–2261. [10]
- Davies, S., Holmes, A., Lomo, L., Steinkamp, M. P., Kang, H., Muller, C. Y., and Wilson, B. S. (2014), "High Incidence of erbb3, erbb4, and met Expression in Ovarian Cancer," *International Journal of Gynecologic Pathology*. [11]
- Davis, J., and Goadrich, M. (2006), "The Relationship Between Precision-Recall and roc Curves," in *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 233–240. [8]
- Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274. [3]
- Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, pp. 1272–1317. [3,4]
- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004), "Sparse Graphical Models for Exploring Gene Expression Data," *Journal of Multivariate Analysis*, 90, 196–212. [1,3]
- Dobra, A., Lenkoski, A., and Rodriguez, A. (2011), "Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data," *Journal of the American Statistical Association*, 106, [1,2,4,8,12]
- Edwards, B. K., Ward, E., Kohler, B. A., Ehemann, C., Zaubler, A. G., Anderson, R. N., Jemal, A., Schymura, M. J., Lansdorp-Vogelaar, I., Seeff, L. C., et al. (2010), "Annual Report to the Nation on the Status of Cancer, 1975–2006, Featuring Colorectal Cancer Trends and Impact of Interventions (Risk Factors, Screening, and Treatment) to Reduce Future Rates," *Cancer*, 116, 544–573. [12]
- Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [8]
- Fitch, A. M., Jones, M. B., and Massam, H. (2014), "The Performance of Covariance Selection Methods That Consider Decomposable Models Only," *Bayesian Analysis*, 9, 659–684. [8,13]
- Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1]
- Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, 7, 601–620. [1]
- Geiger, D., and Heckerman, D. (2002), "Parameter Priors for Directed Acyclic Graphical Models and the Characterization of Several Probability Distributions," *The Annals of Statistics*, 30, 1412–1440. [1]
- Gupta, A., and Nagar, D. (2000), "Matrix Variate Distributions," *Monographs and Surveys in Pure and Applied Mathematics*, Boca Raton, FL: Chapman and Hall/CRC Press. [2]
- Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in Stochastic Computation for High-Dimensional Graphical Models," *Statistical Science*, 388–400. [4,8,11]
- Kundu, S., Baladandayuthapani, V., and Mallick, B. K. (2013), "Bayes Regularized Graphical Model Estimation in High Dimensions," *arXiv:1308.3915*. [1]
- Lauritzen, S. L. (1996), *Graphical Models* (vol. 17), Oxford, UK: Oxford University Press. [1,4]
- Leng, C., and Tang, C. Y. (2012), "Sparse Matrix Graphical Models," *Journal of the American Statistical Association*, 107, 1187–1200. [2,8]
- Li, J., Simpson, L., Takahashi, M., Miliareisis, C., Myers, M. P., Tonks, N., and Parsons, R. (1998), "The pten/mmac1 Tumor Suppressor Induces Cell

- Death That is Rescued by the Akt/Protein Kinase b Oncogene," *Cancer Research*, 58, 5667–5672. [10]
- Liang, F., Song, Q., and Qiu, P. (2015), "An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Graphical Models," *Journal of the American Statistical Association*, 110, 1248–1265. [8]
- Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966. [5]
- Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 1436–1462. [1]
- Mukhopadhyay, N., Price, D. J., Kyriakis, J., Pelech, S., Sanghera, J., and Avruch, J. (1992), "An Array of Insulin-Activated, Proline-Directed Serine/Threonine Protein Kinases Phosphorylate the p70 s6 Kinase," *Journal of Biological Chemistry*, 267, 3325–3335. [10]
- Romanelli, A., Martin, K. A., Toker, A., and Blenis, J. (1999), "p70 s6 Kinase is Regulated by Protein Kinase c ζ and Participates in a Phosphoinositide 3-Kinase-Regulated Signalling Complex," *Molecular and Cellular Biology*, 19, 2921–2928. [10]
- Rosen, D. G., Mercado-Uribe, I., Yang, G., Bast, R. C., Amin, H. M., Lai, R., and Liu, J. (2006), "The Role of Constitutively Active Signal Transducer and Activator of Transcription 3 in Ovarian Tumorigenesis and Prognosis," *Cancer*, 107(11), 2730–2740. [11]
- Roverato, A. (2000), "Cholesky Decomposition of a Hyper Inverse Wishart Matrix," *Biometrika*, 87, 99–112. [3]
- Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [4]
- Scott, J. G., and Carvalho, C. M. (2008), "Feature-Inclusion Stochastic Search for Gaussian Graphical Models," *Journal of Computational and Graphical Statistics*, 17, 790–808. [11,13]
- Selvendiran, K., Tong, L., Vishwanath, S., Bratasz, A., Trigg, N. J., Kutala, V. K., Hideg, K., and Kuppasamy, P. (2007), "Ef24 Induces g2/m Arrest and Apoptosis in Cisplatin-Resistant Human Ovarian Cancer Cells by Increasing Pten Expression," *Journal of Biological Chemistry*, 282, 28609–28618. [10]
- Shojaie, A., and Michailidis, G. (2010), "Penalized Principal Component Regression on Graphs for Analysis of Subnetworks," in *Advances in Neural Information Processing Systems*, pp. 2155–2163. [1]
- Siegel, R., Ma, J., Zou, Z., and Jemal, A. (2014), "Cancer Statistics, 2014," *CA: A Cancer Journal for Clinicians*, 64, 9–29. [10,11]
- Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search* (Vol. 81), Cambridge, MA: The MIT Press. [1]
- Stingo, F., and Marchetti, G. M. (2015), "Efficient Local Updates for Undirected Graphical Models," *Statistics and Computing*, 25, 159–171. [1]
- Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010), "A Bayesian Graphical Modeling Approach to MicroRNA Regulatory Network Inference," *The Annals of Applied Statistics*, 4, 2024–2048. [1]
- Tanner, B., Hasenclever, D., Stern, K., Schormann, W., Bezler, M., Hermes, M., Brulport, M., Bauer, A., Schiffer, I. B., Gebhard, S., et al. (2006), "Erbb-3 Predicts Survival in Ovarian Cancer," *Journal of Clinical Oncology*, 24, 4317–4323. [11]
- The Cancer Genome Atlas (2012), "Comprehensive Molecular Characterization of Human Colon and Rectal Cancer," *Nature*, 487, 330–337. [12]
- Tibes, R., Qiu, Y., Lu, Y., Hennessy, B., Andreeff, M., Mills, G. B., and Kornblau, S. M. (2006), "Reverse Phase Protein Array: Validation of a Novel Proteomic Technology and Utility for Analysis of Primary Leukemia Specimens and Hematopoietic Stem Cells," *Molecular Cancer Therapeutics*, 5, 2512–2521. [10]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 267–288. [7]
- U. S. Cancer Statistics Working Group (2014), *United States Cancer Statistics: 1999–2011 Incidence and Mortality Web-based Report*, Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. Available at www.cdc.gov/uscs. [11]
- van Dyk, D. A., and Jiao, X. (2015), "Metropolis-Hastings Within Partially Collapsed Gibbs Samplers," *Journal of Computational and Graphical Statistics*, 24, 301–327. [6]
- van Dyk, D. A., and Park, T. (2008), "Partially Collapsed Gibbs Samplers: Theory and Methods," *Journal of the American Statistical Association*, 103, 790–796. [2,6]
- Wang, H., and West, M. (2009), "Bayesian Analysis of Matrix Normal Graphical Models," *Biometrika*, 96, 821–834. [1,2,4,13]
- Weichsel, P. M. (1962), "The Kronecker Product of Graphs," *Proceedings of the American Mathematical Society*, 13, 47–52. [4]
- Wermuth, N. (1980), "Linear Recursive Equations, Covariance Selection, and Path Analysis," *Journal of the American Statistical Association*, 75, 963–972. [3,4]
- Whittaker, J. (2009), *Graphical Models in Applied Multivariate Statistics* (1st ed.), New York: Wiley. [1]
- Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1, 80–83. [7]
- Yajima, M., Telesca, D., Ji, Y., and Müller, P. (2015), "Detecting Differential Patterns of Interaction in Molecular Pathways," *Biostatistics*, 16, 240–251. [1]
- Yin, J., and Li, H. (2012), "Model Selection and Estimation in the Matrix Normal Graphical Model," *Journal of Multivariate Analysis*, 107, 119–140. [2]
- Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1]
- Zhou, S. (2014), "Gemini: Graph Estimation With Matrix Variate Normal Instances," *The Annals of Statistics*, 42, 532–562. [2]