Covariate-Assisted Bayesian Graph Learning for Heterogeneous Data

Yabo Niu, Yang Ni, Debdeep Pati & Bani K. Mallick

Taylor & Francis
Taylor & Francis Group

Check for updates

Covariate-Assisted Bayesian Graph Learning for Heterogeneous Data

Yabo Niu[a] , Yang Ni[b], Debdeep Pati[b], and Bani K. Mallick[b]

[a]Department of Mathematics, University of Houston, Houston, TX; [b]Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

In a traditional Gaussian graphical model, data homogeneity is routinely assumed with no extra variables affecting the conditional independence. In modern genomic datasets, there is an abundance of auxiliary information, which often gets under-utilized in determining the joint dependency structure. In this article, we consider a Bayesian approach to model undirected graphs underlying heterogeneous multivariate observations with additional assistance from covariates. Building on product partition models, we propose a novel covariate-dependent Gaussian graphical model that allows graphs to vary with covariates so that observations whose covariates are similar share a similar undirected graph. To efficiently embed Gaussian graphical models into our proposed framework, we explore both Gaussian likelihood and pseudo-likelihood functions. For Gaussian likelihood, a G-Wishart distribution is used as a natural conjugate prior, and for the pseudo-likelihood, a product of Gaussian-conditionals is used. Moreover, the proposed model has large prior support and is flexible to approximate any ν-Hölder conditional variance-covariance matrices with $\nu \in (0, 1]$. We further show that based on the theory of fractional likelihood, the rate of posterior contraction is minimax optimal assuming the true density to be a Gaussian mixture with a known number of components. The efficacy of the approach is demonstrated via simulation studies and an analysis of a protein network for a breast cancer dataset assisted by mRNA gene expression as covariates. Supplementary materials for this article are available online.

1. Introduction

Graphical models are widely recognized as a powerful tool (Dempster 1972) to uncover conditional independence relationship in multivariate observations. They have found widespread applications across many fields, including genomics, causal inference, speech recognition, and computer vision. For instance, in systems biology, graphical models have been used to reverse engineer molecular networks from multi-omic data (Dobra et al. 2004; Friedman 2004; Telesca et al. 2012).

Typically, a graphical model assumes a graph-dependent probability distribution $\mathcal{P}_G(y)$ as the sampling distribution for a q-dimensional random vector $y = (y_1, \ldots, y_q)$, for example, a centered Gaussian distribution with a sparse precision or inverse-covariance matrix. We assume that the conditional independence relationships in y are encoded via an undirected graph $G = (V, E)$ that consists of a set of nodes $V = (1, \ldots, q)$ and a set of undirected edges E such that $s - t \notin E$ if y_s and y_t are conditionally independent given all other variables. One common assumption of many existing graphical model approaches is that the observations are homogeneous and a single graph G is sufficient to characterize the conditional dependency structure of y. However, such an assumption can be violated in many modern applications such as cancer genomic studies where patients are known to be highly heterogeneous and may have distinct molecular networks whose structures are modulated by various individualized factors such as genetic markers and prognostic factors (Dahl 2008; Lohr et al. 2014;

Bolli et al. 2014). Existing statistical approaches (briefly reviewed below) do not adequately capture such network plasticity for heterogeneous populations and hence the goal of this article is to fill in the gap by developing novel covariate-dependent undirected graphical models. We illustrate the idea of the proposed model with a toy example in Figure 1 where a 4-node undirected graph changes with covariate x in both its structure and edge strength.

1.1. Existing Literature on Heterogeneous Graphical Models

Guo et al. (2011), Danaher, Wang, and Witten (2014), Peterson, Stingo, and Vannucci (2015), and Ha, Baladandayuthapani, and Do (2015) considered heterogeneous graphical models without covariates to estimate multiple group-specific undirected graphs where the groups are pre-fixed. The approaches depend on the specific criteria used to segment heterogeneous data, which are often subjective and not necessarily always available, and may result in groups that are still heterogeneous. In another line of work, Rodriguez, Lenkoski, and Dobra (2011) and Talluri, Baladandayuthapani, and Mallick (2014) explored mixtures of graphical models where the population is objectively clustered into homogeneous groups with each group having different graph structures. The clustering is mainly driven by the separation of group-specific means instead of graphs. As demonstrated in our simulation studies, covariate information
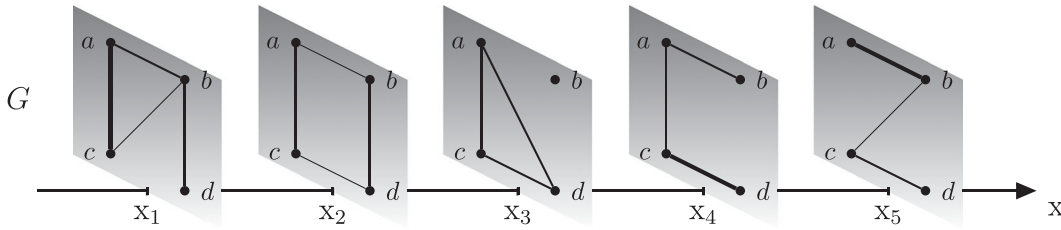
Figure 1. A toy example of an undirected graphical model with four nodes (*a*, *b*, *c*, and *d*) and one continuous covariate x. The edge thickness is proportional to its strength. Both the structure and the edge strength vary with the covariate.

can be beneficial to the estimation of multiple graphs even when there is no clear separation in the first moment of the data. Refer also to Yin and Li (2011), Lee and Liu (2012), Bhadra and Mallick (2013), Cai et al. (2013), Ni, Ji, and Müller (2018), Deshpande, Ročková, V., and George (2019), Niu et al. (2020), and Samanta, Khare, and Michailidis (2022) for graphical models with covariate-dependent mean structure. In essence, these approaches can be written as a multivariate regression model with residuals following a graphical model. While they are effective in adjusting covariates for graph estimation, the graph structure is still assumed to be homogeneous across all observations.

As an extension from covariate dependency through the first moment, a second order covariance regression framework (Hoff and Niu 2012; Fox and Dunson 2015) models the covariance matrix Σ_x as a function of covariates, $\Sigma_x = \Lambda_x \Lambda_x^T + \Psi$ for some positive diagonal matrix Ψ. Matrix Σ_x is clearly positive-definite at any value of *x* by construction. Although, in principle, the covariance regression framework can be extended to graph or inverse-covariance-regression by assuming $\Omega_x = \Sigma_x^{-1} = \Lambda_x \Lambda_x^T + \Psi$, it is not immediately clear how to induce sparsity in Ω_x, a key feature of undirected Gaussian graphical models (GGMs). For instance, in order to set the (i, j)th element of Ω_x to zero, Λ_x needs to be carefully chosen so that the inner product of the *i*th and *j*th rows of Λ_x is exactly zero. Such systematic control calls for a tedious modeling exercise, which becomes even more challenging in relatively sparse high dimensional Ω_x.

There has been some recent developments for covariate-dependent graphs and sparse precision matrices. Liu et al. (2010) developed a graph-valued regression model that partitions the covariate space into rectangles by classification and regression trees and fits GGMs separately to each region. However, the estimated graphs may become unstable and lack similarity for similar covariates due to the separate graph estimation, as reported in Cheng et al. (2014). Several kernel-based methods (Kolar, Parikh, and Xing 2010a, 2010b; Zhou, Lafferty and Wasserman 2010) have been developed for conditional precision matrix estimation. However, due to the curse of dimensionality, they all focused on a univariate covariate. Recently, Ni, Stingo, and Baladandayuthapani (2019) proposed a graphical regression method that estimates directed acyclic graphs as functions of covariates and allows the graph structure to vary continuously with covariates. However, it is difficult to extend their approach to undirected graphs because as mentioned earlier, formulating a sparse positive-definite matrix directly as a function of covariates is highly nontrivial. To the best of our knowledge, there is no existing method that can model continuously varying undirected graphs as functions of general covariates.

1.2. Outline of the Proposed Approach

We propose a novel Bayesian covariate-dependent graphical model that circumvents the challenges described above. Specifically, we avoid directly parameterizing the graphs or sparse precision matrices as functions of covariates by introducing an intermediate layer of latent variables and exploiting the conditional independence of graph and covariates given the intermediate latent variables. We choose this layer to be a random partition that serves as a hidden link between the graphs and covariates. We then specify a model for the graphs given the random partition and a model for the random partition given the covariates, both of which are significantly simpler than directly modeling sparse precision matrices as functions of covariates. Despite the discrete nature of the random partition, one can still arrive at a continuously varying graph or precision matrix by marginalizing out the covariate-dependent partition. We take note here that this cannot be done in non-probabilistic partition-based approaches such as Liu et al. (2010). Even in applications where a clear partition of the population is deemed plausible, marginalizing out the partition can still be interesting and provides natural uncertainty quantification of the graph estimation assisted by covariates. While our focus in this article is on undirected graphs, the proposed model can be easily extended to other graphs such as directed acyclic graphs.

Moreover, we study the theoretical support of our proposed prior by showing that our model is flexible to approximate any ν-Hölder conditional variance-covariance functions $\nu \in (0, 1]$. It is well-known that such functions are dense (in an L_1 sense) in the class of all continuous functions, attesting to the large support. We also establish optimal rates of fractional posterior contraction by using a specific choice of the cohesion function that builds connection between the conditional and joint densities, which allows us to leverage on the existing results on posterior convergence rates of mixture models.

1.3. Novelties of the Proposed Approach

(a) Structurally varying graphs. The existing methods can only handle the covariate-dependent undirected graphs in a discrete fashion (Liu et al. 2010). Our approach offers the ability to model a continuously varying graph by marginalizing out the discrete partition so that discretely and continuously varying

graphs are unified seamlessly under one Bayesian framework, which has not been achieved before. *(b) Theoretical results.* The large support result of the proposed prior shows that although the prior for the conditional precision matrix is continuous in covariates through the multinomial logit transformation in probabilities, it assigns positive probabilities to any L_1 neighborhood of a ν-Hölder continuous conditional precision matrix ($0 < \nu \leq 1$, see its definition in Section 3.4.1). To the best of our knowledge, this is the first result on the flexibility of covariate-dependent graphical models on the space of piecewise conditional covariance functions. In addition, our posterior contraction result demonstrates that our posterior for the conditional density is rate-optimal. *(c) Generality.* Our approach is a general class of models that includes some of the methods mentioned above as special cases. For example, by using certain cohesion function, Rodriguez, Lenkoski, and Dobra (2011) can be considered as a special case of our approach without any covariates. *(d) Generalizability.* Some methods (Liu et al. 2010) can be adapted to different types of graphs but not without major modifications, and some methods (Ni, Stingo, and Baladandayuthapani 2019) cannot easily move beyond its specification. Our approach offers an easy adaptation to any kind of graphs without changing anything significant and it can be easily done by replacing the graph-related likelihood only.

The rest of this article is organized as follows. First, we introduce and formulate the general problem of interest in Section 2. We present our proposed model in Section 3, which contains prior and likelihood specification, model averaging and prediction, and theoretical results of the prior support and the fractional posterior consistency. Posterior inference including a fast blocked Gibbs sampler and some details of model comparison are defered to Section S.4. In Sections 4–5, we illustrate the proposed method using simulations with different dimensions of covariates and graphs, and an application of protein networks in breast cancer with gene expressions as covariates. We conclude this article with discussions in Section 6.

2. General Formulation

Let $y_1, \ldots, y_n$ be n realizations of a q-dimensional random vector $y = (y_1, \ldots, y_q)$ of primary interest and let $x_1, \ldots, x_n$ be the realizations of p-dimensional secondary covariates $x = (x_1, \ldots, x_p)$. The goal is to build a graphical model that characterizes the conditional independencies of y conditional on x. Specifically, we assume a covariate-dependent graphical model $y \sim \mathcal{P}\{y \mid \Omega(x)\}$ where $\Omega(x)$ encodes the covariate-dependent graph structure. For GGMs, $\Omega(x)$ is the covariate-dependent precision matrix. As mentioned in the introduction, directly modeling a covariate-dependent precision matrix can be a challenging problem: it is difficult to ensure that the precision matrix is both sparse and positive-definite while being smoothly varying with the covariates.

To address this challenge, we introduce a discrete latent variable—a covariate-dependent partition—to avoid explicitly define the precision matrix as a function of covariates. It serves as an additional layer between the precision matrix and covariates. This is natural in a Bayesian hierarchical model, and it can

take advantage of marginalization as we shall show in the next section. The covariate-dependent partition ensures the precision matrix function to have only a finite number of values according to the partition; thus, a finite number of graphs as well. To maintain positive definiteness and sparsity in a finite number of precision matrix is substantially easier and many models have been developed for this purpose.

Let $\boldsymbol{\rho} = (S_1, \ldots, S_{K_\rho})$ denote a random partition of $[n] = \{1, \ldots, n\}$, that is, $S_j \cap S_{j'} = \emptyset$ for $j \neq j'$ and $\cup_{j=1}^{K_\rho} S_j = [n]$, where K_ρ is the size of partition $\boldsymbol{\rho}$. In other words, the set S_j contains the indices of the observations that belong to cluster j and K_ρ is the number of clusters. For each cluster $j = 1, 2, \ldots, K_\rho$, we let Ω_j denote its cluster-specific precision matrix and $\Omega = \{\Omega_1, \ldots, \Omega_{K_\rho}\}$. The probability distribution of $y_1, \ldots, y_n$ given the partition and cluster-specific precision matrices can be written as

$$\mathcal{P}(y_1, \ldots, y_n \mid \boldsymbol{\rho}, \Omega) = \prod_{j=1}^{K_\rho} \prod_{i \in S_j} \mathcal{P}(y_i \mid \Omega_j). \quad (2.1)$$

Notice, by introducing the cluster specific indicators $z = \{z_1, \ldots, z_n\}$, where $z_i = j$ if $i \in S_j$ for $i = 1, \ldots, n$ and $j = 1, \ldots, K_\rho$, an equivalent representation of (2.1) can be obtained as

$$\mathcal{P}(y_1, \ldots, y_n \mid z, \Omega) = \prod_{i=1}^{n} \mathcal{P}(y_i \mid z_i, \Omega) = \prod_{i=1}^{n} \mathcal{P}(y_i \mid \Omega_{z_i}).$$

Hence, y_i's are independent given z and Ω, that is, $y_i \mid z_i, \Omega \sim \mathcal{P}(y_i \mid \Omega_{z_i})$. Therefore, the true data-generating distribution of y can be written as a mixture distribution,

$$\mathcal{P}(y \mid \boldsymbol{\pi}, \Omega) = \sum_{j=1}^{K_\rho} \pi_j \mathcal{P}(y \mid \Omega_j), \quad (2.2)$$

where $\pi_j = P(z_i = j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, K_\rho$ and $\sum_{j=1}^{K_\rho} \pi_j = 1$, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{K_\rho}\}$.

For the choice of graphical models (i.e., the choices of probability distribution $\mathcal{P}(y_i \mid \Omega_j)$ and the graphical prior $p(\Omega_j)$), we focus on undirected GGMs. We adopt two approaches in this article. First, we consider a Gaussian likelihood function with the G-Wishart prior (Roverato 2002) as the natural conjugate prior for the precision matrix. It is a well-studied model with many existing efficient procedures (Atay-Kayis and Massam 2005; Jones et al. 2005; Lenkoski and Dobra 2011; Dobra, Lenkoski, and Rodriguez 2011; Wang and Li 2012; Lenkoski 2013; Mohammadi and Wit 2015; Uhler, Lenkoski, and Richards 2018). However, this natural Bayesian formulation does not scale well with high-dimensional graphs. Traditionally, to scale up, researchers often restrict graphs to be decomposable (Dawid and Lauritzen 1993); see also some recent theoretical developments (Xiang, Khare, and Ghosh 2015; Lee and Cao 2021). Another scalable approach, which does not assume decomposability, is the pseudo-likelihood implementation of the neighborhood selection method for GGMs (Meinshausen and Bühlmann 2006). Several Bayesian adaptations have been proposed in recent years (Atchade 2015; Lin et al. 2017). The pseudo-likelihood
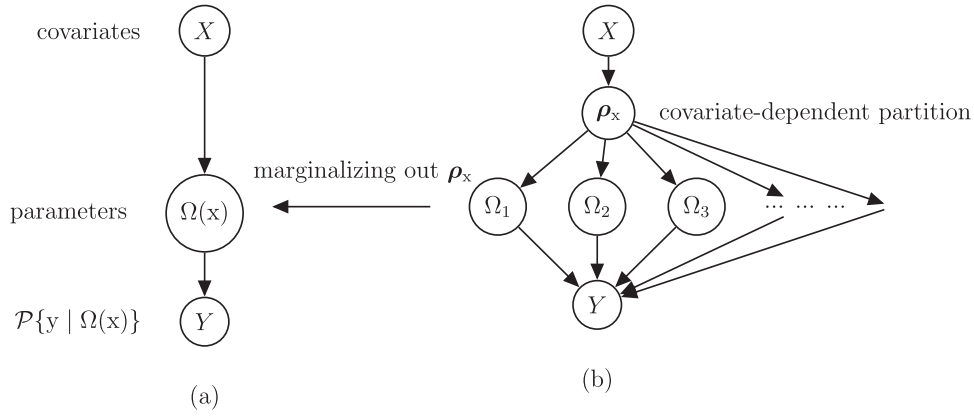
Figure 2. (a) A direct model. (b) The proposed model with an extra layer.

approach, in addition to scalability, allows the use of any existing regression variable selection methods such as the spike-and-slab prior (Mitchell and Beauchamp 1988; Ishwaran and Rao 2005; Narisetty and He 2014), which we adopt in this article.

To incorporate covariates $x_1, \ldots, x_n$, we consider covariate-dependent random partition prior models $p(\rho \mid x_1, \ldots, x_n)$. In this article we choose to use the product partition model with covariates (PPMx, Müller, Quintana, and Rosner 2011). But several other Bayesian models that incorporate covariate information can be used as well, such as dependent Dirichlet process (DDP) models (MacEachern 1999; Sethuraman 1994; Griffin and Steel 2006; Müller, Erkanli, and West 1996; Barcella, De Iorio, and Baio 2017) and the hierarchical mixture of experts (HME) models (Dasgupta and Raftery 1998; Jordan and Jacobs 1994; Bishop and Svensén 2012). Neither DDP or HME is equivalent to PPMx, and none of them has been extended for graphical models. To the best of our knowledge, our proposed model is the first Bayesian partition-based model adapted for covariate-dependent graphical models with heterogeneous data.

Exploiting marginalization, the covariate-dependent graphical model is defined as

$$
\begin{aligned}
&\mathcal{P}\{y_1, \ldots, y_n \mid \Omega(\cdot), x_1 \ldots, x_n\} \\
&= \sum_{\rho \in \mathcal{B}_n} \mathcal{P}(y_1, \ldots, y_n \mid \rho, \Omega_1, \ldots \Omega_{K_\rho}) p(\rho \mid x_1, \ldots, x_n),
\end{aligned}
$$

where $\mathcal{B}_n$ is the collection of all possible partitions of $[n]$. It is clear from the marginal likelihood that the precision matrix $\Omega(\cdot)$ is a function of covariates.

3. Partition-Based Covariate-Dependent Graphical Model

The proposed partition-based covariate-dependent graphical model (PxG) is a hierarchical model with three levels (see Figure 2): the probability distribution $\mathcal{P}(y \mid \Omega)$, the graphical prior $p(\Omega \mid \rho)$ given the partition ρ, and the covariate-dependent partition prior $p(\rho \mid x)$. We will describe the modeling details of each level.

3.1. Covariate-Dependent Partition Prior

As stated in Section 2, we consider partition-based prior models to incorporate covariates. There is a variety of them, such as PPMx, DDP, and HME, just to name a few. In this article, we consider PPMx, that is, the product partition model with covariates proposed in Müller, Quintana, and Rosner (2011) for its simplicity and ease to use. See Page, Quintana, and Rosner (2021) for another application of PPMx. But this does not limit the choice to only PPMx. In fact, any other partition-based models can be adopted within the same Bayesian framework directly if one prefers, which is a big advantage of the proposed model. The PPMx model (Müller, Quintana, and Rosner 2011) is defined as

$$
p(\rho \mid x_1, \ldots, x_n) \propto \prod_{j=1}^{K_\rho} g(X_j^*) \cdot c(S_j),
$$

where the cohesion function $c(S) \geq 0$ for $S \subseteq [n]$ measures the tightness of the elements in S and the similarity function $g(X_j^*) \geq 0$ for $X_j^* = \{x_i \mid i \in S_j\}$ characterizes the similarity of the x_i's in cluster j. The similarity function need not be a proper probability model of covariates, but for convenience, we construct it by marginalizing out the parameters in the auxiliary probability model of covariates $p(x_i \mid \Gamma_j)$ within each group, where $\Gamma_j \sim p(\Gamma_j)$ is a set of parameters, hence

$$
g(X_j^*) = \int \prod_{i \in S_j} p(x_i \mid \Gamma_j) p(\Gamma_j) d\Gamma_j.
$$

Let $z_1, \ldots, z_n \in \{1, \ldots, K_\rho\}$ denote the cluster indicator such that $z_i = j$ if $i \in S_j$. The covariates in our later application are continuous, and therefore we use the following default auxiliary probability model of covariates for cluster $j = 1, \ldots, K_\rho$,

$$
\begin{aligned}
\{x_i\}_{z_i=j} \mid \mu_j, \sigma_j^2 &\overset{\text{iid}}{\sim} \prod_{i \in S_j} N_p(\mu_j, \sigma_j^2 I_p), \\
\mu_j \mid \sigma_j^2 &\sim N_p(\mu_0, \sigma_0^2 \sigma_j^2 I_p), \\
\sigma_j^2 &\sim IG(b_1, b_2).
\end{aligned} \tag{3.1}
$$

The similarity function can be constructed in the same fashion for discrete covariates. Through similarity functions, the PxG model encourages that similar covariates lead to similar

clustering, which in turn leads to similar graph structures. The choices of cohesion function $c(\cdot)$ are plenty. For simplicity, we use $c(S_j) = \alpha(n_j - 1)!$ where $n_j = |S_j|$ is the size of cluster j. With this choice of cohesion function, Dirichlet process (DP) mixture model is a special case of the PxG model (Ferguson 1973; Antoniak 1974), and Rodriguez, Lenkoski, and Dobra (2011) is equivalent to the PxG model with no covariates. See Section S.4 for details of the blocked Gibbs sampler for the PxG model.

3.2. Probability Distribution and Graphical Prior

We focus on undirected GGMs although our methodology is applicable to a much larger class of graphical models. Specifically, conditional on the partition $\boldsymbol{\rho}$, we independently assign an undirected graph and precision matrix to each cluster. That is, $\Omega(x_i) = \Omega_j$ if $i \in S_j$. We effectively reduce an infinite-dimensional functional $\Omega(\cdot)$ to a finite set of parameters $\Omega_1, \ldots, \Omega_{K_\rho}$. Let $Y_j^* = (y_{i_1}^T, \ldots, y_{i_{n_j}}^T)$ with $n_j = |S_j|$ and $S_j = \{i_1, \ldots, i_{n_j}\}$. We consider two approaches for GGMs, which correspond to two different choices of the probability distribution $\mathcal{P}(Y_j^* \mid \Omega_j)$ and graphical prior $p(\Omega_j, G_j)$ for $j = 1, \ldots, K_\rho$.

3.2.1. Approach One: Gaussian-G-Wishart Model
The probability distribution is a centered multivariate Gaussian distribution, $\mathcal{P}(Y_j^* \mid \Omega_j) = \prod_{i \in S_j} N_q(y_i \mid 0, \Omega_j^{-1})$ where the precision matrix $\Omega_j = \{\omega_{st}^j\}_{s,t=1}^q$ encodes the graph structure of $G_j = (V, E_j)$ such that $\omega_{st}^j \neq 0$ if and only if $s - t \in E_j$. Conditional on G_j, we assume a conjugate G-Wishart prior $\Omega_j \sim$ G-Wishart$_{G_j}(b, D)$ with density,

$$p(\Omega_j \mid G_j) = I_{G_j}^{-1}(b, D)|D|^{(b-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(\Omega_j^{-1}D)\right\},$$

where $I_{G_j}^{-1}(b, D)$ is the analytically intractable normalizing constant and $\text{tr}(\cdot)$ is the trace of a matrix. The graphical prior is completed with a prior distribution on graph G. Let g_{st}^j be a binary indicator variable such that $g_{st}^j = 1$ if $s - t \in E_j$ and 0 otherwise. We assume g_{st}^j's are independent Bernoulli, $g_{st}^j \sim$ Bernoulli(α_G), for $1 \leq s < t \leq q$, with prior edge inclusion probability α_G.

Many sampling procedures have been developed for Gaussian-G-Wishart model (Jones et al. 2005; Dobra, Lenkoski, and Rodriguez 2011; Mohammadi and Wit 2015). To improve the scalability of posterior inference, we also consider a pseudo-likelihood approach.

3.2.2. Approach Two: Pseudo-Likelihood Model
As the graph size increases, the Gaussian-G-Wishart approach becomes computationally expensive. Instead of using proper Gaussian likelihood function that constrains the precision matrix to be positive definite, we replace it by the pseudo-likelihood function that is a product of conditional likelihood functions (Besag 1975). Let $y_{i,-s}$ be a subvector of y_i without the sth element. Denote the conditional likelihood of $y_{i,s} \mid y_{i,-s}$ as $Q(y_{i,s} \mid y_{i,-s}, \Omega_j)$ for $i \in S_j$, where $y_{i,s}$ is the sth element of y_i.

Then the pseudo-likelihood function is defined as

$$\mathcal{P}(Y_j^* \mid \Omega_j) = \prod_{i \in S_j}\prod_{s=1}^q Q(y_{i,s} \mid y_{i,-s}, \Omega_j).$$

Such pseudo-likelihood approach has been adopted in modeling undirected graphical models (Ji and Seymour 1996; Csiszár and Talata 2006; Ravikumar, Wainwright, and Lafferty 2010; Ekeberg et al. 2013; Barber and Drton 2015; Pensar et al. 2017). For GGMs, the conditional likelihood function is the node-wise linear regression model, that is, $Q(y_{i,s} \mid y_{i,-s}, \Omega_j) \equiv N(y_{i,s} \mid y_{i,-s}^T \boldsymbol{\beta}_s^j, \tau_s^j)$ where $\boldsymbol{\beta}_s^j = (\beta_{s1}^j, \ldots, \beta_{s,s-1}^j, \beta_{s,s+1}^j, \ldots, \beta_{sq}^j)^T$. This choice is motivated by the fact that the conditional density of $N(y_i \mid 0, \Omega_j^{-1})$ is $N(y_{i,s} \mid y_{i,-s}^T \boldsymbol{\beta}_s^j, \tau_s^j)$ with $\beta_{st}^j = -\omega_{st}^j/\omega_{ss}^j$ and $\tau_s^j = 1/\omega_{ss}^j$ (Peng et al. 2009). Therefore, $\omega_{st}^j = 0$ if and only if $\beta_{st}^j = 0$. Consequently, determining the graph structure is equivalent to finding a sparse subset of $\boldsymbol{\beta}_s^j$ in the regression model $N(y_{i,s} \mid y_{i,-s}^T \boldsymbol{\beta}_s^j, \tau_s^j)$.

In essence, the pseudo-likelihood approach converts the graph structure learning task for GGMs into a variable selection problem in a set of independent linear regressions. Thus, any existing variable selection procedures can be adopted here, including both spike-and-slab priors (Mitchell and Beauchamp 1988; Ishwaran and Rao 2005; Narisetty and He 2014) and shrinkage priors (Park and Casella 2008; Carvalho, Polson, and Scott 2010) under a Bayesian framework; we use the former. Specifically, the density of β_{st}^j is given by

$$\beta_{st}^j \sim g_{st}^j N(0, \eta_1 \tau_s^j) + (1 - g_{st}^j)N(0, \eta_0 \tau_s^j), \qquad (3.2)$$

where g_{st}^j is defined in Section 3.2.1 and $\eta_1 \gg \eta_0 > 0$. The graphical prior is completed with a conjugate inverse-gamma prior, $\tau_s^j \sim \text{IG}(a_1, a_2)$ and the same Bernoulli prior on g_{st}^j as in Section 3.2.1.

The pseudo-likelihood approach provides a fast and flexible way in estimating GGMs, especially when the Gaussian assumption cannot be verified. Atchadé (2019) showed that its posterior contraction rate matches the rate of convergence of the frequentist neighborhood selection. Alluding to the slightly faster upper bound on the rate of posterior contraction, Atchadé (2019) remarked that the pseudo-likelihood approach is statistically more efficient than a full likelihood approach (e.g., Gaussian-G-Wishart approach) for high-dimensional graphs. Although we do not consider high-dimensional graphs ($q > n$) in this article, the pseudo-likelihood approach is computational efficient as well. Its computation is considerably faster than the Gaussian-G-Wishart model because (i) there is no intractable normalizing constant in the graphical prior anymore, and (ii) computation of independent linear regression models are trivially parallelizable. To exploit parallel computing, we do not constrain $g_{st}^j = g_{ts}^j$. This asymmetry can be fixed with post-MCMC processing, which will be detailed in Section S.4.1.3.

3.3. Partition Averaging and Graph Prediction

Recall that by introducing the random partition $\boldsymbol{\rho}$, we effectively discretize the continuous precision matrix function $\Omega(\cdot)$ as

$\Omega(x_i) = \Omega_j$ for $i \in S_j$, which greatly simplifies the modeling of covariate-dependent sparse precision matrix. Because ρ is random, we can recover $\Omega(\cdot)$ (hence, $G(\cdot)$) by *partition averaging* the posterior distribution.

3.3.1. Partition Averaging

The full posterior distribution is given by

$$p\{\Omega(\cdot), G(\cdot), \rho \mid y_{1:n}, x_{1:n}\}$$
$$\propto \mathcal{P}\{y_{1:n} \mid \Omega(\cdot)\} \cdot p\{\Omega(\cdot) \mid \rho, G(\cdot)\} \cdot p\{G(\cdot) \mid \rho\} \cdot p(\rho \mid x_{1:n}),$$

where $y_{1:n}$ and $x_{1:n}$ denote $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$, respectively. Marginalizing out partition ρ,

$$p\{\Omega(\cdot), G(\cdot) \mid y_{1:n}, x_{1:n}\}$$
$$= \sum_{\rho \in \mathcal{B}_n} p\{\Omega(\cdot), G(\cdot), \rho \mid y_{1:n}, x_{1:n}\}$$
$$\propto \sum_{\rho \in \mathcal{B}_n} \mathcal{P}\{y_{1:n} \mid \Omega(\cdot)\} \cdot p\{\Omega(\cdot) \mid \rho, G(\cdot)\} \cdot p\{G(\cdot) \mid \rho\} \cdot p(\rho \mid x_{1:n})$$
$$= \sum_{\rho \in \mathcal{B}_n} \left[\mathcal{P}(y_{1:n} \mid \Omega_1, \ldots, \Omega_{K_\rho}) \prod_{j=1}^{K_\rho} \{p(\Omega_j \mid G_j) p(G_j)\} \right] p(\rho \mid x_{1:n}).$$

In summary, (i) through Bayesian hierarchical formulation via random partition, we reduce an infinite-dimensional quantity $\Omega(\cdot)$ as a sparse positive-definite matrix function of x to a finite number of sparse positive-definite matrices Ω_j's; and (ii) through partition averaging, we recover the infinite dimensional functional $\Omega(\cdot)$ from Ω_j.

3.3.2. Graph Prediction

The PxG model also allows for graph structure prediction given a new sample covariate x_{new} through the posterior predictive distribution

$$p\{\Omega(x_{new}), G(x_{new}) \mid y_{1:n}, x_{1:n}, x_{new}\}$$
$$= \sum_{\rho \in \mathcal{B}_n} \sum_{j=1}^{K_{max}} p(x_{new} \in S_j \mid \rho, y_{1:n}, x_{1:n}) \cdot p(\Omega_j, G_j \mid \rho, y_{1:n}, x_{1:n})$$
$$\cdot p(\rho \mid y_{1:n}, x_{1:n}), \tag{3.3}$$

where K_{max} is the maximum number of clusters allowed. This is a new and useful feature as the prediction does not require the access of y_{new}. Notice that the marginalization performed here cannot be done in non-Bayesian partition-based graphical models (e.g., Liu et al. 2010) since the partition is not random.

3.4. Theoretical Properties of the PxG Model

In the following, we first investigate the flexibility of our prior specification for the PxG model, following which we shall investigate the asymptotic properties of the posterior distribution.

3.4.1. Prior Support

Essential to studying the theoretical support of our prior is to look at an equivalent representation of the conditional density $p(y \mid x)$, where (y, x) denotes a generic response-covariate pair. Based on the previous definition in Section 2, Ω is the collection of $q \times q$ precision-matrix atoms. Let $\mu = \{\mu_j\}_{j=1}^{K_\rho}$ and

$\Sigma^x = \{\Sigma_j^x\}_{j=1}^{K_\rho}$ denote the collection of normal mean-covariance atoms. For simplicity, in this section we assume $\Sigma_j^x = \sigma_j^2 I_p$. As in (2.2), an equivalent representation of the joint density of y and x is given by

$$p(y, x \mid \pi, \Omega, \mu, \Sigma^x) = \sum_{j=1}^{K_\rho} \pi_j N_q(y; 0, \Omega_j^{-1}) N_p(x; \mu_j, \sigma_j^2 I_p).$$

By integrating out y, the marginal density of x is

$$p(x \mid \pi, \mu, \Sigma^x) = \sum_{j=1}^{K_\rho} \pi_j N_p(x; \mu_j, \sigma_j^2 I_p),$$

Hence, we consider the specific form of the conditional density of the PxG model,

$$p(y \mid x, \Omega, \mu, \Sigma^x, \pi) = \frac{p(y, x \mid \Omega, \mu, \Sigma^x, \pi)}{p(x \mid \mu, \Sigma^x, \pi)}$$
$$= \sum_{j=1}^{K_\rho} \pi_j(x) N_q(y; 0, \Omega_j^{-1}), \tag{3.4}$$

where

$$\pi_j(x) = \frac{\pi_j N_p(x; \mu_j, \sigma_j^2 I_p)}{\sum_{t=1}^{K_\rho} \pi_t N_p(x; \mu_t, \sigma_t^2 I_p)}. \tag{3.5}$$

Equation (3.4) is a conditional density representation with covariate-dependent weights in (3.5). Interestingly, if $\sigma_j^2 \equiv \sigma^2$, $\pi_j(x)$ depends only on linear functions of x, noting the simplification,

$$\pi_j(x) = \frac{\pi_j \exp\{(-\|\mu_j\|^2/2 + \mu_j^T x)/\sigma^2\}}{\sum_{t=1}^{K_\rho} \pi_t \exp\{(-\|\mu_t\|^2/2 + \mu_t^T x)/\sigma^2\}}, \tag{3.6}$$

where $\|\cdot\|$ is the vector Euclidean norm. Equation (3.6) is reminiscent of multinomial-logit probabilities as one can reparameterize $\pi_j(x)$ in terms of the canonical parameters $(a_j, b_j) \in \mathbb{R} \times \mathbb{R}^p$ of the Gaussian exponential family

$$\pi_j(x) = \frac{\pi_j \exp(a_j + b_j^T x)}{\sum_{t=1}^{K_\rho} \pi_t \exp(a_t + b_t^T x)}.$$

Thus, the PxG model with $\sigma_j^2 \equiv \sigma^2$ can be viewed as an extension of the HME models by Jordan and Jacobs (1994). Norets (2010) obtained approximation results of an arbitrary continuous conditional density using the HME models of the form (3.6) under suitable regularity conditions, demonstrating the flexibility of such models. Norets and Pati (2017) took this a step further and obtained minimax rates of posterior convergence for appropriately smooth conditional densities.

As $y \mid x, \Omega, \mu, \Sigma^x, \pi$ is not Gaussian, the variance-covariance matrix $\text{var}(y \mid x, \Omega, \mu, \Sigma^x, \pi)$ no longer captures the conditional dependency relations among $(y_1, \ldots, y_q)$. Observe that the conditional independency relation here holds true only by further conditioning on the latent clusters. Similar examples of conditional independency conditional on certain latent parameters can be found in the literature on robust graphical modeling

(Finegold and Drton 2011, 2014; Cremaschi et al. 2019). It nevertheless is important to look at the form of var(y | x, Ω, $\boldsymbol{\mu}$, Σ^x, $\boldsymbol{\pi}$) and how it changes with covariates. Observe that

$$\text{var}(y \mid x, \Omega, \boldsymbol{\mu}, \Sigma^x, \boldsymbol{\pi}) = \sum_{j=1}^{K_\rho} \pi_j(x)\Omega_j^{-1}.$$

Notice that the precision matrix of y | x, Ω, $\boldsymbol{\mu}$, Σ^x, $\boldsymbol{\pi}$ is given by $\{\sum_{j=1}^{K_\rho} \pi_j(x)\Omega_j^{-1}\}^{-1}$. Given y and x, one can find a simplified representation of the posterior mean of var(y | x, Ω, $\boldsymbol{\mu}$, Σ^x, $\boldsymbol{\pi}$) which is $\sum_{j=1}^{K_\rho} \mathbb{E}\{\pi_j(x)\Omega_j^{-1} \mid y, x\}$. It is hard to characterize the nature and sparsity structure of $\{\sum_{j=1}^{K_\rho} \pi_j(x)\Omega_j^{-1}\}^{-1}$ in its full generality, but if $\pi_j(x)$ behaves approximately like an indicator function $\mathbb{1}_{\mathcal{X}_j}(x)$ for a finite partition $\{\mathcal{X}_1, \ldots, \mathcal{X}_{K_\rho}\}$ of $\mathcal{X} \subseteq \mathbb{R}^p$, then the precision matrix of y | x would be approximately $\sum_{j=1}^{K_\rho} \mathbb{1}_{\mathcal{X}_j}(x)\Omega_j$. In this case the precision matrix approximately reflects the conditional independence structure of y | x. Theorem 3.1 shows that if indeed all entries of the true conditional variance var(y | x, Ω, $\boldsymbol{\mu}$, Σ^x, $\boldsymbol{\pi}$) is ν-Hölder continuous, $0 < \nu \le 1$ (see the definition of ν-Hölder continuity in the lemma below), then the prior assigns positive probability to appropriate integrated neighborhood of the true function of the covariance matrix $\Omega_0^{-1}(\cdot)$. For simplicity, we only consider the case when the covariate x is scalar for the following lemma and theory in this section and use notation x for scalar x. Before introducing Theorem 3.1, we first show the fact that ν-Hölder continuous matrix can be approximated by a piecewise constant function.

Lemma 3.1. Denote the true conditional covariance matrix of y given x as $\Omega_0^{-1}(x) = [\sigma_{ij}^0(x)]_{q \times q}$. Let $\mathcal{H}^\nu$ be the space of uniformly ν-Hölder continuous functions, $0 < \nu \le 1$, that is,

$$\mathcal{H}^\nu = \left\{ f : [a, b] \to \mathbb{R}; \; \|f\|_{\mathcal{H}^\nu} \equiv \sup_{x, x' \in [a,b]} \frac{|f(x) - f(x')|}{|x - x'|^\nu} < \infty \right\},$$

where $\|f\|_{\mathcal{H}^\nu}$ is the ν-Hölder coefficient. Assume $\sigma_{ij}^0(x) \in \mathcal{H}^\nu$, $i, j = 1, \ldots, q$. Therefore, for any $\epsilon > 0$, there exists $K_\epsilon \in \mathbb{N}^+$ and a set of positive definite matrices $\{\Omega_{0k}\}_{k=1}^{K_\epsilon}$ where $\Omega_{0k}^{-1} = (\sigma_{ij}^{0k})_{q \times q}$ and a set of positive real numbers $a \le a_1 < \cdots < a_{K_\epsilon+1} \le b$ such that

$$\left\| \Omega_0^{-1}(\cdot) - \sum_{j=1}^{K_\epsilon} \mathbb{1}_{(a_j, a_{j+1}]}(\cdot)\Omega_{0j}^{-1} \right\|_1 < \epsilon,$$

where Ω_{0j}^{-1} and a_j depend only on $\Omega_0^{-1}(\cdot)$ and ϵ, $\|A(\cdot)\|_1 = \int_a^b \|A(x)\| dx$ for any matrix norm $\|\cdot\|$.

Proof. See Supplementary Materials Section S.1. □

Theorem 3.1. (Prior large support). Assume that the true conditional density of y given x satisfies var(y | x) = $\Omega_0^{-1}(x)$, where the precision matrix $\Omega_0(\cdot)$ is a continuous function of a one-dimensional covariate $x \in (a, b)$ for $-\infty < a < b < \infty$. Assume that all entries of the true covariance matrix $\Omega_0^{-1}(x)$ are in $\mathcal{H}^\nu$, $0 < \nu \le 1$. Then there exists $K_\epsilon' \in \mathbb{N}^+$ such that for

any absolutely continuous prior distributions on $(\Omega_j, \mu_j, \Sigma_j^x) \in \mathcal{S}_{q \times q}^+ \times \mathbb{R} \times \mathbb{R}^+$ and $(\pi_1, \ldots, \pi_{K_\epsilon'})$ in the simplex $\Delta^{K_\epsilon'-1}$,

$$\Pi \left\{ \left\| \Omega_0^{-1}(\cdot) - \sum_{j=1}^{K_\epsilon'} \pi_j(\cdot)\Omega_j^{-1} \right\|_1 < \epsilon \right\} > 0,$$

where $\mathcal{S}_{q \times q}^+$ is the cone of $q \times q$ positive definite matrices and $\pi_j(\cdot)$ is in the form of (3.5).

This specific choice of ν-Hölder continuous true conditional variance is motivated by the good performance of the simulation study in Sections 4.1 and 4.2. Refer to Supplementary Materials Section S.2 for a proof of Theorem 3.1. Observe that the proof proceeds in two steps: (i) Invoking Lemma 3.1, we first approximate a ν-Hölder continuous using a piecewise constant function. (ii) Approximating a piecewise constant density using densities of type (3.4). This is in contrast with Norets (2010) and Norets and Pati (2017) who considered approximating only smooth conditional densities. Another difference with Norets (2010) and Norets and Pati (2017) is that we obtain prior support in an integrated metric whereas Norets (2010) obtain pointwise approximation bounds. Observe that for $\nu \in (0, 1]$ we approximate the ν-Hölder covariance function with a piecewise constant function in an integrated metric. This does not hold for $\nu > 1$ in a stronger metric (such as the supremum norm). Refer to Theorem 2 of Castillo (2014), which requires $\alpha > 1$ for the supremum-norm posterior contraction results to hold.

One of our key contributions is to recognize the connection between conditional density in (3.4) to the mixture of experts representation in Norets and Pati (2017) under a special choice of parameters. This helps us adapt results on prior large support, and later in Section 3.4.2 allows us to leverage on the posterior contraction rate for the joint density (Shen, Tokdar, and Ghosal 2013).

In addition to the large prior support property, we also develop posterior contraction rates in Section 3.4.2 where it is important to assume smoothness. In fact, as we shall see below in Section 3.4.2, our derivation for rate of contraction assumes true joint density to be a mixture of Gaussians. As shown in Ghosal, Ghosh, and Van Der Vaart (2000), this is a fairly broad class of densities and can approximate any smooth density with desirable accuracy. In addition, we characterize the role of the number of mixture components K in rate, but assume the dimension to be fixed.

3.4.2. Posterior Contraction Rates

In this section, we investigate the asymptotic properties of the posterior distribution associated with the PxG model. For technical simplicity, we shall consider a fractional likelihood introduced in Bhattacharya, Pati, and Yang (2019), although we anticipate analogous results with the original likelihood. Let p_Θ be the joint density of x and y in the PxG model and p_{Θ_0} be the true model, where Θ and Θ_0 are the model parameters and the true parameters, respectively. Instead of fitting p_Θ, we shall work with p_Θ^α for $\alpha \in (0, 1)$ and investigate the convergence rate of the posterior $\Pi_{n,\alpha}$ given by

$$\Pi_{n,\alpha}(p_\Theta \mid x_{1:n}, y_{1:n}) = \frac{p_\Theta^\alpha(x_{1:n}, y_{1:n})\Pi(\Theta)}{\int p_\Theta^\alpha(x_{1:n}, y_{1:n})\Pi(\Theta)d\Theta}$$

for some prior Π on the parameters Θ. Next we shall state two types of assumptions on the true density p_{Θ_0} and assumed prior Π, respectively.

Assumption 3.1. (Assumptions on the conditional density and marginal density of covariates). Assume the true conditional density of the q-dimensional response y given the p-dimensional covariate x has the following form

$$p(y \mid x, \Theta_0) = \sum_{j=1}^{K} \pi_j(x \mid \Theta_0) \cdot N_q(y; 0, \Sigma_{G_{0j}}^0)$$

$$= \sum_{j=1}^{K} \pi_j(x \mid \Theta_0) \cdot \phi_{\Sigma_{G_{0j}}^0},$$

and the true marginal density of covariates is

$$p(x \mid \Theta_0) = \sum_{j=1}^{K} \pi_{0j} \cdot N_p(x; \boldsymbol{\mu}_{0j}, \Sigma_{0j}^x) = \sum_{j=1}^{K} \pi_{0j} \cdot \phi_{\boldsymbol{\mu}_{0j}, \Sigma_{0j}^x},$$

where

$$\pi_j(x \mid \Theta_0) = \frac{\pi_{0j} \cdot \phi_{\boldsymbol{\mu}_{0j}, \Sigma_{0j}^x}}{\sum_{j=1}^{K} \pi_{0j} \cdot \phi_{\boldsymbol{\mu}_{0j}, \Sigma_{0j}^x}},$$

$\Theta_0 = \{(\pi_{0j}, \Sigma_{G_{0j}}^0, \boldsymbol{\mu}_{0j}, \Sigma_{0j}^x)\}_{j=1}^K$, K is a known constant and $\Sigma_{0j}^x = \text{diag}(\sigma_{0j1}^2, \ldots, \sigma_{0jp}^2)$ is a diagonal matrix with diagonal entries σ_{0ji}^2, $i = 1, \ldots, p$, $j = 1, \ldots, K$. Let $p_{\Theta_0} = p(y \mid x, \Theta_0) \cdot p(x \mid \Theta_0)$ be the true joint density and p_Θ be any joint density with parameters $\Theta = \{(\pi_j, \Sigma_{G_j}, \boldsymbol{\mu}_j, \Sigma_j^x)\}_{j=1}^K$, where $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jp})$ and $\Sigma_j^x = \text{diag}(\sigma_{j1}^2, \ldots, \sigma_{jp}^2)$. Moreover, let $0 < \lambda_{j1}^0 \leq \cdots \leq \lambda_{jq}^0$ be the eigenvalues of $\Sigma_{G_{0j}}^0$. Assume $0 < \lambda_m^0 \leq \lambda_{j1}^0 \leq \cdots \leq \lambda_{jq}^0 \leq \lambda_M^0$ and $\sigma_m^2 \leq \sigma_{0ji}^2 \leq \sigma_M^2$, for $j = 1, \ldots, q$, $i = 1, \ldots, p$, where $\lambda_m^0, \lambda_M^0, \sigma_m^2, \sigma_M^2$ are global constants.

Assumption 3.2. (Assumptions on Π).
1. Prior for $(\pi_1, \ldots, \pi_K)$ is Dirichlet $(\pi_0/K, \ldots, \pi_0/K)$, $\pi_0 > 0$.
2. Prior for μ_{ji} is any normal density, $j = 1, \ldots, K$, $i = 1, \ldots, p$.
3. Prior for σ_{ji}^2 is any inverse-Gamma density, $j = 1, \ldots, K$, $i = 1, \ldots, p$.
4. Prior for Σ_{G_j} given G_j is any G-Wishart density and edges in G_j follow independent Bernoulli density, $j = 1, \ldots, K$.

To measure the closeness between p_{Θ_0} and p_Θ, define the α-divergence with respect to the true probability measure $\mathbb{P}_{\Theta_0}$ as

$$D_{p_{\Theta_0}, \alpha}(p_\Theta, p_{\Theta_0}) := \frac{1}{\alpha - 1} \log \int \left(\frac{p_\Theta}{p_{\Theta_0}}\right)^\alpha p_{\Theta_0} d\mu.$$

Using the two assumptions above, we derive the main theorem of the fractional posterior distribution around the true density.

Theorem 3.2. (Contraction of fractional posterior distributions). Fix $\alpha \in (0, 1)$ and let $\epsilon_n = \sqrt{\frac{K(p+q)\log n}{n}}$. Suppose Assumptions 3.1 and 3.2 hold. Then, for any $D \geq 2$ and $t > 0$,

$$\Pi_{n,\alpha}\left(D_{p_{\Theta_0}, \alpha}(p_\Theta, p_{\Theta_0}) \geq \frac{D + 3t}{1 - \alpha}\epsilon_n^2 \mid x_{1:n}, y_{1:n}\right) \leq e^{-t\epsilon_n^2}$$

holds with $\mathbb{P}_{\Theta_0}$ probability at least $1 - 2/\{(D-1+t)^2\epsilon_n^2\}$, where $\Pi_{n,\alpha}(\cdot \mid x_{1:n}, y_{1:n})$ is the posterior probability and $x_{1:n}, y_{1:n}$ are the sample covariates and responses, respectively.

Proof. See supplementary materials section S.3. □

Compared to the regular posterior contraction results, the fractional posterior result in Theorem 3.2 only requires the prior to be sufficiently concentrated around Θ_0. Observe that we have worked with the simplifying assumption when the true density p_{Θ_0} is itself a mixture of K-Gaussians and our PxG model assumed the knowledge of the number K. Indeed the convergence rate $\sqrt{K(p+q)\log n/n}$ is minimax optimal (Tsybakov 2008). When the number of mixtures $K = n^{\frac{p+q}{2\beta+p+q}}, \beta > 0$, one can recover the contraction rates of posterior for estimating $(p+q)$-variate β-smooth densities (Shen, Tokdar, and Ghosal 2013; Norets and Pati 2017).

4. Simulation Study

In this section, we present three simulated examples. The first two simulation studies consider the entries of precision matrix to be linear or piecewise linear functions of covariates whereas the third simulation assume discrete functions (i.e., piecewise constant precision matrices). We demonstrate that the proposed PxG model is able to handle both types of precision matrix functions. To show the necessity of using covariate information, we compare the PxG model with two following "partial" models: covariate-only model where graph is assumed to be constant among sample; graph-only model where graph varies with samples but does not depend on covariates. Model comparison among them is conducted via DIC in this section.

4.1. Example 1: Piecewise Linear Precision Matrix

We consider an example with a three-dimensional GGM of which the precision matrix is a piecewise linear function of the given one-dimensional covariate x. Define the true covariate-dependent precision matrix as $\Omega(x) = [\Omega_{st}(x)]_{s,t=1,2,3}$ where $\Omega_{ss}(x) = 1.2$ and

$-1 < x < -0.33$, $\Omega_{12}(x) = 0$, $\Omega_{13}(x) = -0.75x + 0.25$,
$\Omega_{23}(x) = 0.75x + 1.25$; $-0.33 \leq x < 0.33$,
$\Omega_{12}(x) = 0.75x + 0.75$, $\Omega_{13}(x) = 0$, $\Omega_{23}(x) = -0.75x + 0.75$;
$0.33 \leq x < 1$, $\Omega_{12}(x) = -0.75x + 1.25$,
$\Omega_{13}(x) = 0.75x + 0.25$, $\Omega_{23}(x) = 0$.

Thus, the piecewise linear function of the precision matrix defines three clusters with different graphs. They are $G_1 = (V, E_1)$ for $-1 < x < -0.33$ with $E_1 = \{1 - 3, 2 - 3\}$; $G_2 = (V, E_2)$ $-0.33 < x < 0.33$ with $E_2 = \{1 - 2, 2 - 3\}$; $G_3 = (V, E_3)$ for $0.33 \leq x < 1$ with $E_3 = \{1 - 2, 1 - 3\}$, where $V = \{1, 2, 3\}$. We draw 100 realizations of the covariate uniformly from each of the three clusters, that is, the total sample size is 300. Then the responses y are drawn from the corresponding covariate-dependent normal distribution $N_3\{0, \Omega^{-1}(x)\}$. We simulate 1000 MCMC samples using the blocked Gibbs sampler for the Gaussian-G-Wishart model with 500 burn-in samples.
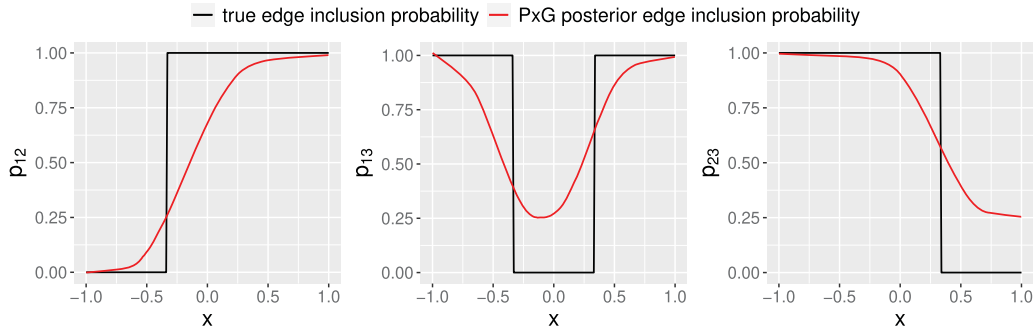
Figure 3. Edge inclusion probability as a function of covariate x. Black: true edge inclusion probability. Red: PxG posterior estimate of edge inclusion probability.
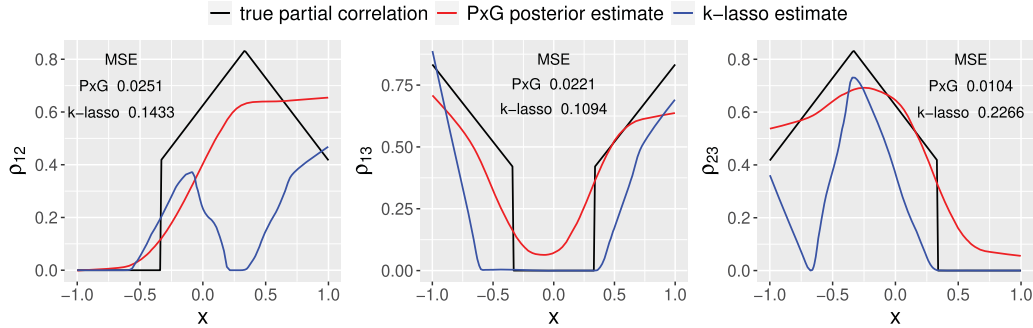

Figure 4. Partial correlation as a function of covariate x. Black: true partial correlation. Red: PxG posterior estimate of partial correlation. Blue: k-lasso estimate of partial correlation.

We plot the posterior edge inclusion probability as a function of x for all three edges in Figure 3. The black lines are the true edge inclusion and the red lines are the estimated posterior edge inclusion probabilities. The posterior edge inclusion probability is captured by averaging out all MCMC samples. We can also calculate the posterior partial correlation matrix as a function of the covariate x based on the MCMC samples; see Figure 4. From both figures, the posterior estimates capture the overall trends of the true edge inclusion and the true piecewise linear precision matrix.

Moreover, we compare the PxG model with the kernel graphical lasso approach (hereafter, k-lasso) described in Liu et al. (2010), which estimates a covariate-dependent covariance matrix via kernel smoothing. Let $S(x)$ be the covariance matrix varying with a covariate x, then

$$S(x) = \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right)(y_i - \mu(x))(y_i - \mu(x))^T \bigg/ \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right)$$

with

$$\mu(x) = \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right) y_i \bigg/ \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right).$$

Here $h > 0$ is the bandwidth and $K(\cdot)$ is a Gaussian kernel. We tune h on a $[0.1,1]$ grid with the Akaike information criterion. To obtain a sparse precision matrix Ω_i for each sample, glasso (Friedman, Hastie, and Tibshirani 2008) is applied, $\hat{\Omega}_i = \arg\min_{\Omega}\left\{\log|\Omega| - \operatorname{tr}(S(x_i)\Omega) - \lambda_i\|\Omega\|_1\right\}$. The corresponding blue curves in Figure 4 shows the partial correlation estimates using k-lasso approach. We can see that it does not capture the basic shape of the piecewise linear partial correlation functions. And k-glasso approach does not yield posterior estimates of edge

inclusion probabilities. Mean squared errors (MSE) of PxG and k-lasso for each partial correlation are shown in Figure 4 as well. By comparing the proposed PxG model (DIC=2193) with the covariate-only model (DIC=2585) and the graph-only model (DIC=3190) and via the deviance information criterion (DIC) (see supplementary materials for details), the PxG model has the smallest DIC value, indicating the best model fitting.

4.2. Example 2: Linear Precision Matrix with Constant Graph

In this scenario, we demonstrate that the PxG model can recover the precision matrix as a function of covariates even when the graph induced by the precision matrix is constant. We compare PxG with approaches proposed by Bhadra and Mallick (2013) (hereafter, BM13) and Deshpande, Ročková, V., and George (2019) (hereafter, mSSL) to show the necessity of using covariate information in graph estimates and to show that under this case model misspecification leads to incorrect graph estimates. We consider a five-dimensional covariate-dependent precision matrix $\Omega(x)$ as a linear function of a scalar covariate x. It is defined as follows: $\Omega_{ss}(x) = 1.4$ for $s = 1, 2, 3, 4, 5$; $\Omega_{s,s+1}(x) = x$ for $s = 1, 2, 3, 4$; the rest entries in $\Omega(x)$ are 0 and $x \in (-0.8, 0) \cup (0, 0.8)$. Thus, 4 entries in $\Omega(x)$ are linear functions of x. The constant graph induced by $\Omega(x)$ is a 5-node chain graph $G = (V, E)$ with 4 edges, where $V = \{1, 2, 3, 4, 5\}$ and $E = \{1 - 2, 2 - 3, 3 - 4, 4 - 5\}$. Covariate x is drawn uniformly from its range with sample size $n = 300$, and the corresponding responses y are generated from $N_5\{0, \Omega^{-1}(x)\}$. We simulate 2000 MCMC samples using the blocked Gibbs sampler of the Gaussian-G-Wishart model with 1000 burn-in samples.
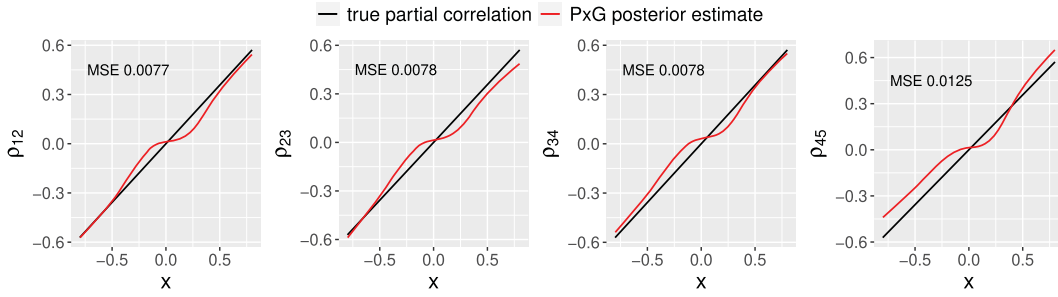
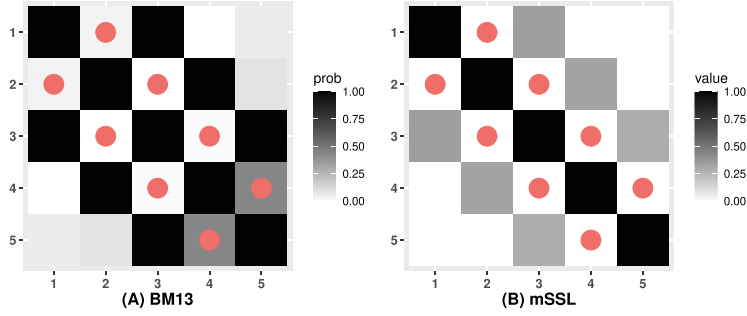**Figure 5.** PxG posterior estimate of partial correlations.

Figure 6. (A) BM13 posterior estimate of edge inclusion probability. (B) mSSL estimate of absolute value of partial correlation. Red dots indicate true edges in $\Omega(x)$.
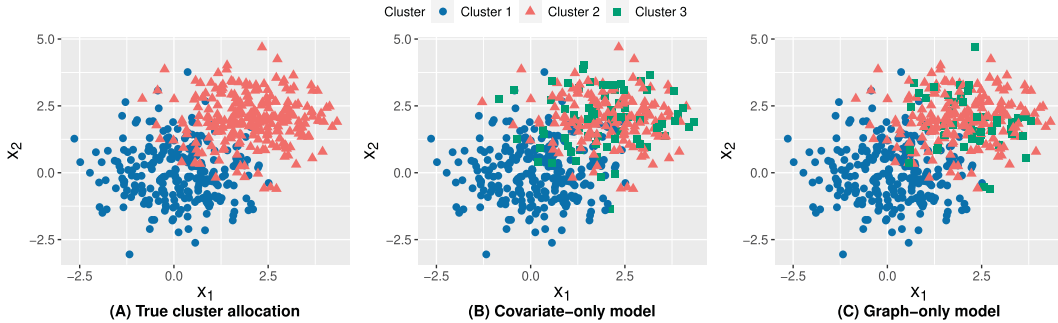

Figure 7. (A) indicates the true cluster allocation, which is also the estimated cluster allocation of PxG (i.e., no misclassification). (B) and (C) are the posterior estimates of cluster allocations of the covariate-only model and the graph-only model, respectively.

We plot the posterior estimate of partial correlation as a function of x for the four edges in Figure 5 indicated by red line along with the MSE for each partial correlation estimate. The black line is the true partial correlation. Thanks to partition averaging, the PxG model captures the linear precision matrix well even though all observations share the same graph and there is no cluster induced from the underlying data generating process. Figure 6 shows the results from BM13 and mSSL. For BM13, the posterior estimate of edge inclusion probability is shown in Figure 6(A); for mSSL, we plot the absolute value of the estimate of partial correlation matrix in Figure 6(B). Red dots in both plots are true edges. Both methods fail to identify the true edges induced by $\Omega(x)$ due to model mis-specification since BM13 and mSSL assume a constant graph and a constant covariance matrix, both of which are independent from covariates. Without using covariate information, the graph estimates from BM13 and mSSL seem to be dictated by the sample partial correlations of responses. In fact, we observe that sample partial correlations of edges 1–3, 2–4, and 3–5 are about 4 times larger than the rest (including the true edges).

4.3. Example 3: Piecewise Constant Precision Matrix

In this example, we consider the precision matrix to be a piece-wise constant function of a set of covariates. Thus, clusters can be induced from the precision matrix. Assume $\Omega(x)$ to be a 50-dimensional covariate-dependent precision matrix where x is a 10-dimensional covariate. The piecewise constant precision matrix of $\Omega(x)$ is defined as $\Omega(x) = \Omega_1$ for $x \sim N_{10}(\mathbf{0}_{10}, I_{10})$ and $\Omega(x) = \Omega_2$ for $x \sim N_{10}(\mathbf{2}_{10}, I_{10})$, where $\mathbf{0}_{10} = 0 \cdot \mathbf{1}_{10}$ and $\mathbf{2}_{10} = 2 \cdot \mathbf{1}_{10}$. Two clusters are induced by $\Omega(x)$, and Ω_1 and Ω_2 are the cluster-specific constant precision matrices. Let G_1 and G_2 be the corresponding sparse graphs induced from Ω_1 and Ω_2, respectively. They are selected randomly with approximately 1% sparsity, and then Ω_1 and Ω_2 are drawn from G-Wishart distribution with the degree of freedom equal to 3 and scale matrix equal to the identity matrix given G_1 and G_2. We generate 250 realizations of covariate x for each cluster and the corresponding responses are drawn from $N_{50}\{\mathbf{0}_{10}, \Omega^{-1}(x)\}$.

Figure 7 (A) shows the first two dimensions of the 10-dimensional covariates with cluster allocation. The variances of both clusters are relatively large comparing to the distance
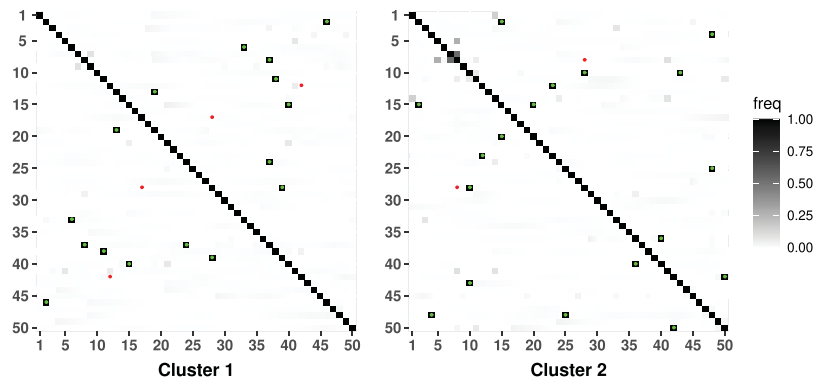
Figure 8. PxG posterior edge inclusion probability for Cluster 1 and 2. Red and green dots in both plots are the true edges in G_1 and G_2 with red indicating false negative edges and green for true positive edges.
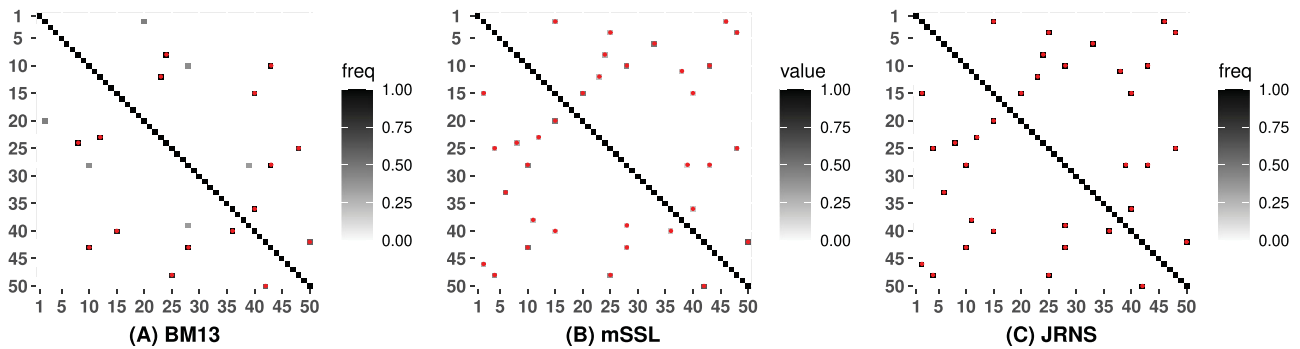

Figure 9. (A) BM13 posterior estimate of edge inclusion probability. (B) mSSL estimate of absolute value of partial correlation. (C) JRNS posterior estimate of edge inclusion probability. Red dots indicate edges selected by each approach.

between their means. Thus, we observe that there is significant overlap between these two clusters. Due to the relatively large dimension of graph in this example, we use the pseudo-likelihood model with the blocked Gibbs sampler. The Gaussian-G-Wishart model is too slow for such application due to the computational burden of calculating the normalizing constant of G-Wishart density. In the pseudo-likelihood approach, g_{st}^j in (3.2) is not necessary the same as g_{ts}^j since they are two separate parameters in two independent conditional likelihood functions that make up the pseudo-likelihood function. Therefore, the posterior edge inclusion probabilities of g_{ts}^j and g_{st}^j, denoted as p_{ts}^j and p_{st}^j, are unlikely to be the same. We use $\max(p_{ts}^j, p_{st}^j)$ as the final posterior edge inclusion probability for edge $s - t$ (similar in spirit to the union approach in the frequentist neighborhood selection method). This ensures the graph estimates are symmetric. We simulate 1000 MCMC samples with 500 burn-in samples. For comparison, we follow the same procedure for the covariate-only model and the graph-only model.

The PxG model correctly identifies both clusters with no misclassification. Edge inclusion probabilities for each cluster are show in Figure 8. Red and green dots are false negative edges and true positive edges, respectively, according to the true graph in each cluster. By selecting edges with posterior edge inclusion probabilities greater than 0.5, its posterior graph estimate of G_1 has 2 false negative edges and no false positive edges; for G_2 it has 1 false negative edge and 1 false positive edge. Due to the closeness of covariates between the two clusters, the covariate-only model is not able to correctly identify both clusters as shown

in Figure 7 (B). Figure 7(C) indicates that, without utilizing any covariate information, the response information is not enough for the graph-only model to identify Cluster 2 correctly, hence the true piecewise precision matrix as well. This demonstrates the necessity of using covariate information to assist graph estimates in heterogeneous GGMs. The DIC values for the PxG, covariate-only, and graph-only models are 104166, 155453, and 106138, respectively, with the PxG having the smallest value among them.

We compare the PxG model with BM13, mSSL, and a recent approach by Samanta, Khare, and Michailidis (2022) (hereafter, JRNS) in this example. We plot the posterior edge inclusion probability for BM13 in Figure 9(A), the estimates of absolute values of partial correlations using mSSL in Figure 9(B), and the posterior edge inclusion probability for JRNS in Figure 9(C) with red dots indicating edges selected by each approach. For BM13 and JRNS, the red dots are edges with posterior edge inclusion probability greater than 0.5 and for mSSL they are edges corresponding to nonzero entries in the estimate of precision matrix. Since the graph and the precision matrix are assumed to be constant among all observations, these three approaches suffer from model misspecification and do not identify the true edges in the graph. The graph estimates of mSSL and JRNS are identical, both having 17 edges. Compared with the true G_2, they have 1 false negative edge and 8 false positive edges. It is worse when compared to true G_1. The posterior graph estimate of BM13 has 8 edges whose posterior edge inclusion probabilities are greater than 0.5. It has 5 false negatives and 3 false positives compared with the true G_2.
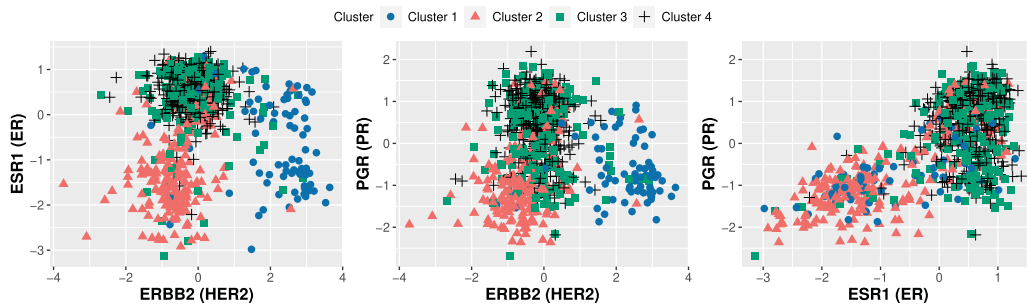
Figure 10. Posterior estimates of cluster allocation of the PxG model shown on pairwise scatterplots of 3 mRNA covariates: ERBB2, ESR1, and PGR.
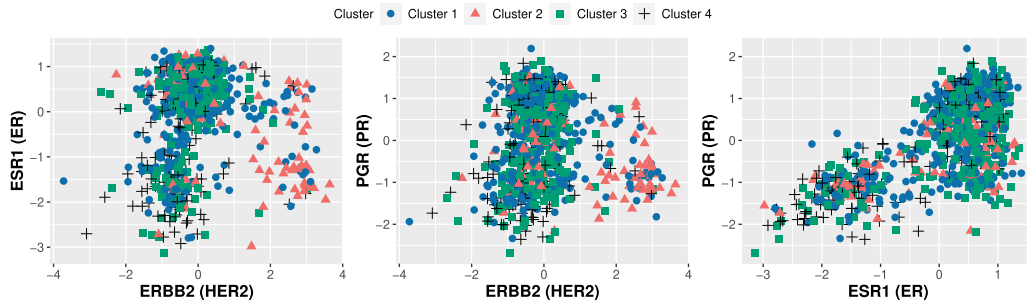

Figure 11. Posterior estimates of cluster allocation of the graph-only model shown on pairwise scatterplots of 3 mRNA covariates: ERBB2, ESR1, and PGR.

5. Real Data Analysis: Breast Cancer Data

In this section, we apply the proposed PxG model to a breast cancer dataset. The mRNA and protein data of breast cancer are downloaded from The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) through the R program `TCGA-Assembler` (Wei et al. 2018). The central dogma of genetics states that genetic information flows from mRNA to protein. It is natural that we use mRNA expressions as covariates and protein expressions as responses in our approach. We model the protein network of breast cancer as a mRNA-dependent graph. The PxG model allows us to uncover how interactions between proteins change with different mRNA expression. Different cancer treatments take advantages of various properties of protein–protein interactions. Therefore, discovering how those interactions vary with mRNA expression is one of the key steps in developing new targeted, personalized cancer therapy.

It has been known that human epidermal growth factor receptor 2 (HER2), estrogen receptor (ER), and progesterone receptor (PR) are important in determining breast cancer subtypes (Onitilo et al. 2009). They are related to treatment, survival rate, and cancer cell metabolism of breast cancer (Kennecke et al. 2010; Lobbezoo et al. 2013; Rimawi, Schiff, and Osborne 2015). We focus on the corresponding mRNAs of these three receptors, which are ERBB2 for HER2, ESR1 for ER and PGR for PR. For proteins, in addition to the protein expression of the three markers, we also include 12 proteins from the mTOR pathway, which is known to play a significant role in breast cancer (Hynes and Boulay 2006; Cidado and Park 2012; McAuliffe et al. 2010). In total, 883 observations are available for this analysis.

We apply the PxG model with pseudo-likelihood approach and generate 10,000 MCMC samples with 1000 burn-in samples. For comparison, the graph-only model is applied, which only uses protein information. Figure 10 shows the posterior estimates of cluster allocation under the PxG model on pairwise scatterplots of 3 mRNA covariates, and Figure 11 shows the posterior estimates of cluster allocation for the graph-only model. Under PxG, there is a clear separation among 3 of the 4 posterior clusters which can be observed from the first two scatterplots in Figure 10, whereas the same separation is not observed under the graph-only results on Figure 11. This suggests that the covariates influence but do not dominate the clusters in the PxG model.

Additionally, we use the total sum of squares (TSS) of covariates to measure the overall variance under a cluster allocation. It is defined as $\text{TSS} = \sum_{j=1}^{\hat{K}} \sum_{i \in \hat{S}_j} \|x_i - \hat{\mu}_j\|^2$, where $\hat{\mu}_j$ is the sample covariate mean of the jth cluster, $\hat{S}_j$ is the index set of the estimated cluster j, and $\hat{K}$ is the estimated number of clusters. The TSS for the PxG and graph-only models are 509,029 and 645,585, respectively, which again indicates that the covariate information assists the clustering results for heterogeneous GGMs under PxG. The DIC values (37,531 for the PxG model and 40,798 for the graph-only model) confirm this conclusion as well.

For protein network estimation, we plot the posterior edge inclusion probability for each cluster under the PxG model in Figure 12, where red dots indicate the included edges (inclusion probability greater than 0.5). According to the posterior estimate of the cluster allocation of the PxG model in Figure 10, ERBB2 is highly expressed (relatively large values) in cluster 1 and it is suppressed (relatively small values) in cluster 2; ESR1 is highly expressed in both cluster 3 and 4. We observe that when ESR1 is highly expressed, the connectivity among proteins intensifies based on the protein networks from cluster 3 and 4 in Figure 12. The connection between protein expressions of ERBB2 and PGR seems independent of the mRNA gene expression of ERBB2 since the edge between them presents in both cluster 1 and 2. Other than few exceptions, most edges among clusters do not
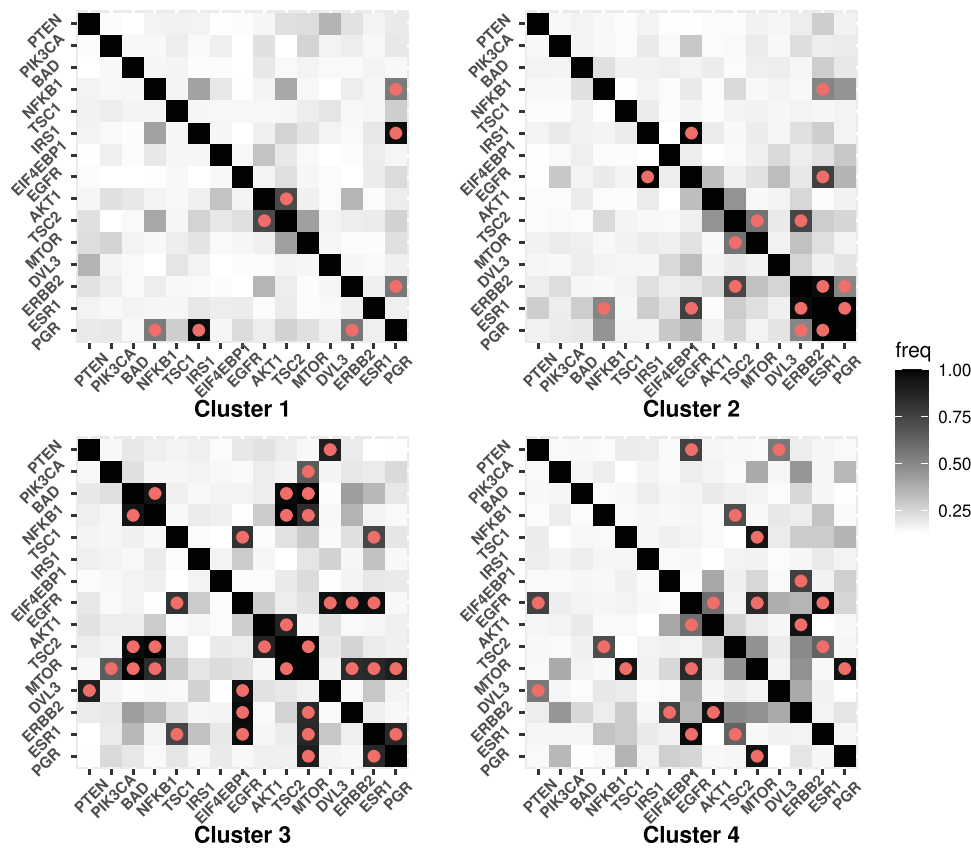
Figure 12. Posterior edge inclusion probability for each cluster under the PxG model. Red dots indicate posterior graph selection.

overlap. Since cluster 3 and 4 have a similar mRNA profiles among ERBB2, ESR1 and PGR, they share the most edges than other pairwise edge comparison among clusters. The results in Figure 12 illustrate how protein connectivity change with certain mRNAs.

6. Discussion

We have introduced a novel Bayesian model, PxG, for heterogeneous graphical models with covariates. We have focused on undirected Gaussian graphical models with two different approaches, that is, the Gaussian-G-Wishart approach and the pseudo-likelihood approach. The modularity of the proposed PxG model allows straightforward extension to other graphical models such as directed acyclic graphs. Despite the discreteness of the random partition, we have shown both theoretically and empirically that PxG is capable of recovering graphs that continuously change with covariates. One of the important components of the proposed PxG is the PPMx model. Recently Page, Quintana, and Rosner (2021) uses PPMx in regression problems to discover the interaction effects of predictors, which are mainly categorical, on an univariate response. Note that our approach uses PPMx for a completely different purpose, that is, we focus on learning the precision matrix and graph structure (conditional independencies) of a set of multivariate Y-variables assisted by a set of covariates X. Therefore, the dependencies considered in PxG and Page, Quintana, and Rosner (2021) are fundamentally different. Future direction of this work includes investigating other nonparametric Bayesian methods

for covariate-dependent random partition models as well as a more direct approach in modeling the continuous relationship between graph/precision matrix and covariates.

Supplementary Materials

Supplementary Materials include all the proofs, the posterior inference procedure, and the code that implements the proposed method.

Acknowledgments

We thank the Editor, the Associate Editor, and anonymous reviewers for their constructive comments, which led to significant improvements of this article.

Disclosure Statement

There are no conflict of interests to declare.

ORCID

Yabo Niu ⬤ http://orcid.org/0000-0001-8087-6747

References

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174. [5]

Atay-Kayis, A., and Massam, H. (2005), "A Monte Carlo Method for Computing the Marginal Likelihood in Nondecomposable Gaussian Graphical Models," *Biometrika*, 92, 317–335. [3]

Atchade, Y. (2015), "A Scalable Quasi-Bayesian Framework for Gaussian Graphical Models," arXiv preprint arXiv:1512.07934. [3]

Atchadé, Y. F. (2019), "Quasi-Bayesian Estimation of Large Gaussian Graphical Models," *Journal of Multivariate Analysis*, 173, 656–671. [5]

Barber, R. F., and Drton, M. (2015), "High-Dimensional Ising Model Selection with Bayesian Information Criteria," *Electronic Journal of Statistics*, 9, 567–607. [5]

Barcella, W., De Iorio, M., and Baio, G. (2017), "A Comparative Review of Variable Selection Techniques for Covariate Dependent Dirichlet Process Mixture Models," *Canadian Journal of Statistics*, 45, 254–273. [4]

Besag, J. (1975), "Statistical Analysis of Non-lattice Data," *Journal of the Royal Statistical Society*, Series D, 24, 179–195. [5]

Bhadra, A., and Mallick, B. K. (2013), "Joint High-Dimensional Bayesian Variable and Covariance Selection with an Application to eQTL Analysis," *Biometrics*, 69, 447–457. [2,9]

Bhattacharya, A., Pati, D., and Yang, Y. (2019), "Bayesian Fractional Posteriors," *The Annals of Statistics*, 47, 39–66. [7]

Bishop, C. M., and Svensén, M. (2012), "Bayesian Hierarchical Mixtures of Experts," arXiv preprint arXiv:1212.2447. [4]

Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., et al. (2014), "Heterogeneity of Genomic Evolution and Mutational Profiles in Multiple Myeloma," *Nature Communications*, 5, 1–13. [1]

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013), "Covariate-Adjusted Precision Matrix Estimation with an Application in Genetical Genomics," *Biometrika*, 100, 139–156. [2]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [5]

Castillo, I. (2014), "On Bayesian Supremum Norm Contraction Rates," *The Annals of Statistics*, 42, 2058–2091. [7]

Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014), "A Sparse Ising Model with Covariates," *Biometrics*, 70, 943–953. [2]

Cidado, J., and Park, B. H. (2012), "Targeting the pi3k/akt/mtor Pathway for Breast Cancer Therapy," *Journal of Mammary Gland Biology and Neoplasia*, 17, 205–216. [12]

Cremaschi, A., Argiento, R., Shoemaker, K., Peterson, C., and Vannucci, M. (2019), "Hierarchical Normalized Completely Random Measures for Robust Graphical Modeling," *Bayesian Analysis*, 14, 1271–1301. [7]

Csiszár, I. and Talata, Z. (2006), "Consistent Estimation of the Basic Neighborhood of Markov Random Fields," *The Annals of Statistics*, 34, 123–145. [5]

Dahl, D. B. (2008), "Distance-based Probability Distribution for Set Partitions with Applications to Bayesian Nonparametrics," in *JSM Proceedings. Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, Alexandria, VA. [1]

Danaher, P., Wang, P., and Witten, D. M. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation across Multiple Classes," *Journal of the Royal Statistical Society*, Series B, 76, 373–397. [1]

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial Point Processes with Clutter via Model-based Clustering," *Journal of the American Statistical Association*, 93, 294–302. [4]

Dawid, A. P., and Lauritzen, S. L. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *The Annals of Statistics*, 21, 1272–1317. [3]

Dempster, A. P. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [1]

Deshpande, S. K., Ročková, V., and George, E. I. (2019), "Simultaneous Variable and Covariance Selection with the Multivariate Spike-and-Slab Lasso," *Journal of Computational and Graphical Statistics*, 28, 921–931. [2,9]

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]

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, 1418–1433. [3,5]

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013), "Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models," *Physical Review E*, 87, 012707. [5]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [5]

Finegold, M., and Drton, M. (2011), "Robust Graphical Modeling of Gene Networks Using Classical and Alternative *t*-distributions," *The Annals of Applied Statistics*, 5, 1057–1080. [7]

——— (2014), "Robust Bayesian Graphical Modeling Using Dirichlet *t*-distributions," *Bayesian Analysis*, 9, 521–550. [7]

Fox, E. B., and Dunson, D. B. (2015), "Bayesian Nonparametric Covariance Regression," *The Journal of Machine Learning Research*, 16, 2501–2542. [2]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [9]

Friedman, N. (2004), "Inferring Cellular Networks Using Probabilistic Graphical Models," *Science*, 303, 799–805. [1]

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000), "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531. [7]

Griffin, J. E., and Steel, M. J. (2006), "Order-based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194. [4]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [1]

Ha, M. J., Baladandayuthapani, V., and Do, K.-A. (2015), "Dingo: Differential Network Analysis in Genomics," *Bioinformatics*, 31, 3413–3420. [1]

Hoff, P. D., and Niu, X. (2012), "A Covariance Regression Model," *Statistica Sinica*, 22, 729–753. [2]

Hynes, N. E., and Boulay, A. (2006), "The mtor Pathway in Breast Cancer," *Journal of Mammary Gland Biology and Neoplasia*, 11, 53–61. [12]

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773. [4,5]

Ji, C., and Seymour, L. (1996), "A Consistent Model Selection Procedure for Markov Random Fields based on Penalized Pseudolikelihood," *The Annals of Applied Probability*, 6, 423–443. [5]

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*, 20, 388–400. [3,5]

Jordan, M. I., and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214. [4,6]

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M. C. U., Voduc, D., Speers, C. H., Nielsen, T. O., and Gelmon, K. (2010), "Metastatic Behavior of Breast Cancer Subtypes," *Journal of Clinical Oncology*, 28, 3271–3277. [12]

Kolar, M., Parikh, A. P., and Xing, E. P. (2010a), "On Sparse Nonparametric Conditional Covariance Selection," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 559–566, Madison, WI, USA. Omnipress. [2]

Kolar, M., Song, L., Ahmed, A., Xing, E. P., et al. (2010b), "Estimating Time-Varying Networks," *The Annals of Applied Statistics*, 4, 94–123. [2]

Lee, K., and Cao, X. (2021), "Bayesian Inference for High-Dimensional Decomposable Graphs," *Electronic Journal of Statistics*, 15, 1549–1582. [3]

Lee, W., and Liu, Y. (2012), "Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood," *Journal of Multivariate Analysis*, 111, 241–255. [2]

Lenkoski, A. (2013), "A Direct Sampler for g-wishart Variates," *Stat*, 2, 119–128. [3]

Lenkoski, A., and Dobra, A. (2011), "Computational Aspects Related to Inference in Gaussian Graphical Models with the g-wishart Prior," *Journal of Computational and Graphical Statistics*, 20, 140–157. [3]

Lin, Z., Wang, T., Yang, C., and Zhao, H. (2017), "On Joint Estimation of Gaussian Graphical Models for Spatial and Temporal Data," *Biometrics*, 73, 769–779. [3]

Liu, H., Chen, X., Wasserman, L., and Lafferty, J. D. (2010), "Graph-valued Regression," in *Advances in Neural Information Processing Systems*, pp. 1423–1431. [2,3,6,9]

Lobbezoo, D. J., van Kampen, R. J., Voogd, A. C., Dercksen, M. W., van den Berkmortel, F., Smilde, T. J., van de Wouw, A. J., Peters, F. P., van Riel, J. M., Peters, N. A., et al. (2013), "Prognosis of Metastatic Breast Cancer Subtypes: The Hormone Receptor/HER2-Positive Subtype is Associated with the Most Favorable Outcome," *Breast Cancer Research and Treatment*, 141, 507–514. [12]

Lohr, J. G., Stojanov, P., Carter, S. L., Cruz-Gordillo, P., Lawrence, M. S., Auclair, D., Sougnez, C., Knoechel, B., Gould, J., Saksena, G., et al. (2014), "Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy," *Cancer Cell*, 25, 91–101. [1]

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science* (Vol. 1), pp. 50–55, Alexandria, VI: American Statistical Association. [4]

McAuliffe, P. F., Meric-Bernstam, F., Mills, G. B., and Gonzalez-Angulo, A. M. (2010), "Deciphering the Role of pi3k/akt/mtor Pathway in Breast Cancer Biology and Pathogenesis," *Clinical Breast Cancer*, 10, S59–S65. [12]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [3]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [4,5]

Mohammadi, A., and Wit, E. C. (2015), "Bayesian Structure Learning in Sparse Gaussian Graphical Models," *Bayesian Analysis*, 10, 109–138. [3,5]

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79. [4]

Müller, P., Quintana, F., and Rosner, G. L. (2011), "A Product Partition Model with Regression on Covariates," *Journal of Computational and Graphical Statistics*, 20, 260–278. [4]

Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817. [4,5]

Ni, Y., Ji, Y., and Müller, P. (2018), "Reciprocal Graphical Models for Integrative Gene Regulatory Network Analysis," *Bayesian Analysis*, 13, 1095–1110. [2]

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019), "Bayesian Graphical Regression," *Journal of the American Statistical Association*, 114, 184–197. [2,3]

Niu, Y., Guha, N., De, D., Bhadra, A., Baladandayuthapani, V., and Mallick, B. K. (2020), "Bayesian Variable Selection in Multivariate Nonlinear Regression with Graph Structures," arXiv preprint arXiv:2010.14638. [2]

Norets, A. (2010), "Approximation of Conditional Densities by Smooth Mixtures of Regressions," *The Annals of Statistics*, 38, 1733–1766. [6,7]

Norets, A., and Pati, D. (2017), "Adaptive Bayesian Estimation of Conditional Densities," *Econometric Theory*, 33, 980–1012. [6,7,8]

Onitilo, A. A., Engel, J. M., Greenlee, R. T., and Mukesh, B. N. (2009), "Breast Cancer Subtypes based on er/pr and her2 Expression: Comparison of Clinicopathologic Features and Survival," *Clinical Medicine & Research*, 7, 4–13. [12]

Page, G. L., Quintana, F. A., and Rosner, G. L. (2021), "Discovering Interactions Using Covariate Informed Random Partition Models," *The Annals of Applied Statistics*, 15, 1–21. [4,13]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [5]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746. [5]

Pensar, J., Nyman, H., Niiranen, J., and Corander, J. (2017), "Marginal Pseudo-Likelihood Learning of Discrete Markov Network Structures," *Bayesian Analysis*, 12, 1195–1215. [5]

Peterson, C., Stingo, F. C., and Vannucci, M. (2015), "Bayesian Inference of Multiple Gaussian Graphical Models," *Journal of the American Statistical Association*, 110, 159–174. [1]

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D.. (2010), "High-Dimensional Ising Model Selection Using l1-regularized Logistic Regression," *The Annals of Statistics*, 38, 1287–1319. [5]

Rimawi, M. F., Schiff, R., and Osborne, C. K. (2015), "Targeting HER2 for the Treatment of Breast Cancer," *Annual Review of Medicine*, 66, 111–128. [12]

Rodriguez, A., Lenkoski, A., and Dobra, A. (2011), "Sparse Covariance Estimation in Heterogeneous Samples," *Electronic Journal of Statistics*, 5, 981–1014. [1,3,5]

Roverato, A. (2002), "Hyper Inverse Wishart Distribution for Non-decomposable Graphs and its Application to Bayesian Inference for Gaussian Graphical Models," *Scandinavian Journal of Statistics*, 29, 391–411. [3]

Samanta, S., Khare, K., and Michailidis, G. (2022), "A Generalized Likelihood-based Bayesian Approach for Scalable Joint Regression and Covariance Selection in High Dimensions," *Statistics and Computing*, 32, 1–23. [2,11]

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650. [4]

Shen, W., Tokdar, S. T., and Ghosal, S. (2013), "Adaptive Bayesian Multivariate Density Estimation with Dirichlet Mixtures," *Biometrika*, 100, 623–640. [7,8]

Talluri, R., Baladandayuthapani, V., and Mallick, B. K. (2014), "Bayesian Sparse Graphical Models and their Mixtures," *Stat*, 3, 109–125. [1]

Telesca, D., Müller, P., Kornblau, S. M., Suchard, M. A., and Ji, Y. (2012), "Modeling Protein Expression and Protein Signaling Pathways," *Journal of the American Statistical Association*, 107, 1372–1384. [1]

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, New York: Springer. [8]

Uhler, C., Lenkoski, A., and Richards, D. (2018), "Exact Formulas for the Normalizing Constants of Wishart Distributions for Graphical Models," *The Annals of Statistics*, 46, 90–118. [3]

Wang, H., and Li, S. Z. (2012), "Efficient Gaussian Graphical Model Determination under g-wishart Prior Distributions," *Electronic Journal of Statistics*, 6, 168–198. [3]

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., and Ji, Y. (2018), "TCGA-assembler 2: Software Pipeline for Retrieval and Processing of TCGA/CPTAC Data," *Bioinformatics*, 34, 1615–1617. [12]

Xiang, R., Khare, K., and Ghosh, M. (2015), "High Dimensional Posterior Convergence Rates for Decomposable Graphical Models," *Electronic Journal of Statistics*, 9, 2828–2854. [3]

Yin, J., and Li, H. (2011), "A Sparse Conditional Gaussian Graphical Model for Analysis of Genetical Genomics Data," *The Annals of Applied Statistics*, 5, 2630–2650. [2]

Zhou, S., Lafferty, J., and Wasserman, L. (2010), "Time Varying Undirected Graphs," *Machine Learning*, 80, 295–319. [2]