Journal of Machine Learning Research 23 (2022) 1-29 Submitted 1/21; Revised 3/22; Published 8/22

Bayesian Covariate-Dependent Gaussian Graphical Models with Varying Structure

Yang Ni YNI@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Francesco C. Stingo FRANCESCOCLAUDIO.STINGO@UNIFI.IT
Department of Statistics, Computer Science, Applications "G. Parenti"
The University of Florence
Florence, Italy

Veerabhadran Baladandayuthapani VEERAB@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Qiang Liu

Abstract

We introduce Bayesian Gaussian graphical models with covariates (GGMx), a class of multivariate Gaussian distributions with covariate-dependent sparse precision matrix. We propose a general construction of a functional mapping from the covariate space to the cone of sparse positive definite matrices, which encompasses many existing graphical models for heterogeneous settings. Our methodology is based on a novel mixture prior for precision matrices with a non-local component that admits attractive theoretical and empirical properties. The flexible formulation of GGMx allows both the strength and the sparsity pattern of the precision matrix (hence the graph structure) change with the covariates. Posterior inference is carried out with a carefully designed Markov chain Monte Carlo algorithm, which ensures the positive definiteness of sparse precision matrices at any given covariates' values. Extensive simulations and a case study in cancer genomics demonstrate the utility of the proposed model.

Keywords: Covariate-dependent graphs; Markov random fields; Random thresholding; Subject-level inference; Undirected graphs

1. Introduction

Undirected Gaussian graphical models (GGMs), also known as Gaussian Markov random fields, are one of the common tools to analyze multivariate data with complex structure and find many useful applications across biomedicine, finance, and public health. A GGM can simply be expressed as a multivariate Gaussian distribution with a sparse precision (inverse-covariance) matrix. The zero entries of the precision matrix have probabilistic interpretation of conditional independence between the Gaussian random variables (nodes of a graph). Moreover, all the conditional independence relationships can be directly estimated from the accompanying undirected graph for which a zero entry in the precision matrix corresponds

to a missing edge in the graph. This equivalency, essentially reduces the problem of graph structure learning in GGMs to finding zeros in the precision matrix.

Many existing GGM approaches (Dobra et al., 2004; Sudderth et al., 2004; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Scott and Carvalho, 2008; Dobra et al., 2011; Green and Thomas, 2013; Drton and Maathuis, 2017; Khare et al., 2018; Massam, 2018; Gan et al., 2019) assume an independent and identically distributed (i.i.d.) sampling scheme $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip}) \sim N(0, \boldsymbol{\Omega}^{-1})$ for $i = 1, \ldots, n$ where $\boldsymbol{\Omega}$ is the precision matrix. However, the independence assumption does not hold in many applications. For example, observations in multivariate time series data are not independent and exhibit temporal correlations; similarly for spatial data with spatial correlation. In addition, the assumption of identical distribution implies homogeneity across observations and is often violated as well. For instance, tumor heterogeneity is a well-known characteristic in cancer: patients with the same cancer-type can be rather different in their genetic/genomic architecture. Forcing the same GGM (i.e., the same precision matrix $\boldsymbol{\Omega}$) onto every patient is a restrictive modeling assertion, when modeling cancer genomic networks.

Attempts have been made to extend GGMs or other types of graphical models beyond i.i.d. data. If there is a natural grouping of the observations, *multiple graphical models* (Guo et al., 2011; Danaher et al., 2014; Oates et al., 2014; Peterson et al., 2015; Yajima et al., 2015; Xie et al., 2016; Ni et al., 2018; Shaddox et al., 2018) can be applied to learn group-specific graphs assuming observations within each group are i.i.d.. Another line of work incorporates additional covariates $\boldsymbol{x}_i$ in estimating graphs. *Conditional Gaussian graphical models* (Rothman et al., 2010; Yin and Li, 2011; Bhadra and Mallick, 2013) are multivariate linear regression models with the error terms following an i.i.d. GGM (can be viewed as chain graphical models). While graph estimation is conditional on the covariates, they only enter the model via the mean structure. As a consequence, the graph topology and the precision matrix stay the same across observations. In this paper, we are taking a more direct approach, in the sense that the latent graph and hence the sparse precision matrix are explicit functions of covariates.

There are a few recent work in this direction. Liu et al. (2010a) proposed a tree-based method that partitions the covariate space into a finite number of subspaces by classification and regression trees and fits GGMs separately to subsets of data. However, the estimated graphs may be unstable and lack similarity for similar covariates due to the separate graph estimation, as reported by Cheng et al. (2014). Kolar et al. (2010) proposed a penalized kernel smoothing approach that allows the precision matrix to vary with covariates. Cheng et al. (2014) developed a conditional Ising model for binary data where the dependencies are linear functions of covariates. Although the methods of Kolar et al. (2010) and Cheng et al. (2014) allow edge strength to vary with covariates, graph structure is assumed to be constant across all observations. Recently, Ni et al. (2019) proposed a graphical regression framework that allows both edge strength and graph structure to vary with covariates in (directed) Bayesian networks. They assumed there exists a natural ordering of the nodes. Given this assumption, Bayesian networks can be written as systems of recursive linear regressions. A conditional independence function was then introduced to connect regression coefficients with covariates.

In this paper, we consider a general problem of estimating *undirected* GGMs conditional on covariates (GGMx). GGMx allows not only the edge strength (i.e., off-diagonal elements

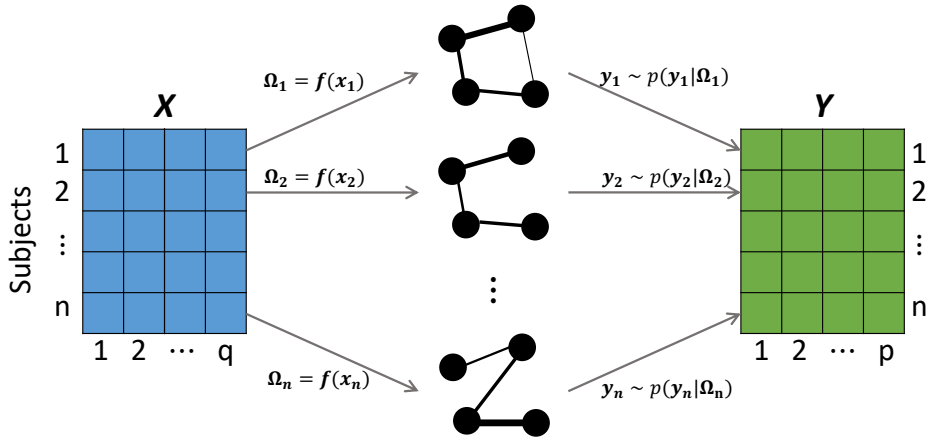


Figure 1: Illustration of GGMx. Subject-level sparse precision matrices $\boldsymbol{\Omega}_i$ and graphs G_i of $\boldsymbol{Y}$ vary with covariates $\boldsymbol{X}$. The edge thickness is proportional to its strength of association ω_{ijk}. Both edge strength and graph structure change with $\boldsymbol{X}$. GGMx can also be viewed as a generative model: $\boldsymbol{X}$ generate graphs which in turn generate $\boldsymbol{Y}$.

of precision matrix) but also the graph structure (i.e., sparsity pattern of precision matrix) to vary as functions of covariates, which is illustrated in Figure 1 with graphs with four nodes and two covariates. Figure 1 also illustrates the generative mechanism underlying GGMx: covariates $\boldsymbol{x}_i$ generate sparse precision matrices $\boldsymbol{\Omega}_i$ (hence the graphs G_i), which in turn generate responses $\boldsymbol{y}_i$. The major challenge in this context is the positive definiteness constraint of precision matrices – a *sine qua non* for GGMs – in the presence of covariates. We propose a simple strategy by specifying a matrix-valued function $\boldsymbol{f}(\cdot)$, such that $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ is a positive definite matrix for any $\boldsymbol{x}_i$ *almost surely*; along with the function $\boldsymbol{f}(\cdot)$ consisting of a random thresholding component that encourages sparse precision matrix estimation, specifically enforcing the required zero-pattern that corresponds to missing edges. The sparse functional relationship between $\boldsymbol{\Omega}_i$ and $\boldsymbol{x}_i$ allows for novel graph interpolation for an unseen observation at covariates $\boldsymbol{x}^*$. We show that the random thresholding gives rise to a discrete mixture of *non-local priors (Johnson and Rossell, 2010) for precision matrices*. We also carefully design a Markov chain Monte Carlo (MCMC) algorithm for posterior inference, which guarantees to propose positive definite precision matrices for any $\boldsymbol{x}_i$. GGMx allows for subject-level inference on unknown graphs. Moreover, GGMx is a general class of graphical models, which subsumes at least five special cases including standard GGMs, group-specific GGMs (Guo et al., 2011; Danaher et al., 2014; Peterson et al., 2015), time-varying GGMs (Zhou et al., 2010), covariate-dependent GGMs (Kolar et al., 2010), and context-specific GGMs (Nyman et al., 2017). Extensive simulation studies show strong and robust performance of GGMx compared with competing methods. Using a cancer genomics case study, we demonstrate how GGMx can be used to infer subject-specific gene networks, which can facilitate deeper investigations in the genomic foundation of precision medicine.

The rest of this article is organized as follows. We introduce the background and notations in Section 2. We present the proposed GGMx in Section 3 and discuss the link between the random thresholding prior and non-local priors in Section 4. We summarize

the posterior inference and graph interpolation in Section 5. We demonstrate the utility and robustness of GGMx with extensive simulation studies in Section 6. GGMx is illustrated by a real data application in Section 7. Section 8 provides our closing discussion.

2. Background and notation

A GGM is a multivariate Gaussian distribution with a sparse precision matrix. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p) \sim N(0, \boldsymbol{\Omega}^{-1})$ be multivariate Gaussian random variables with mean zero and precision matrix $\boldsymbol{\Omega} = [\omega_{jk}]$. Since the off-diagonal elements in $\boldsymbol{\Omega}$ are proportional to partial correlations, a zero entry $\omega_{jk} = 0$ indicates that Y_j and Y_k are conditionally independent given all other variables. A GGM graphically represents the zero patterns of $\boldsymbol{\Omega}$ by an undirected graph. An undirected graph $G = (V, E)$ consists of a set of nodes $V = \{1, \ldots, p\}$ and a set of undirected edges $E \subseteq \{\{j, k\} | j, k \in V\}$. The nodes V represent the variables $\boldsymbol{Y}$ and an edge $\{j, k\}$ is present in the graph if and only if $\omega_{jk} \neq 0$. This is not an arbitrary way of drawing a graph. In fact, the conditional independence relationships that are encoded in the multivariate Gaussian distribution can be directly read off from G using the notion of graph separation. Importantly, learning the graph structure is equivalent to finding the zero patterns of $\boldsymbol{\Omega}$.

Under the Bayesian paradigm, several prior distributions (Roverato, 2002; Wang et al., 2012; Wang, 2015) for sparse precision matrices have been developed, which all take the same general form,

$$\pi(\boldsymbol{\Omega}) = \frac{\tilde{\pi}(\boldsymbol{\Omega}) I(\boldsymbol{\Omega} \in M^+)}{\int \tilde{\pi}(\boldsymbol{\Omega}) I(\boldsymbol{\Omega} \in M^+) d\boldsymbol{\Omega}} \propto \tilde{\pi}(\boldsymbol{\Omega}) I(\boldsymbol{\Omega} \in M^+), \tag{1}$$

with M^+ being the collection of positive definite matrices (PDMs). For example, G-Wishart prior (Roverato, 2002) assumes $\tilde{\pi}(\boldsymbol{\Omega})$ to be a Wishart distribution $Wishart(\cdot|b, \boldsymbol{\Omega}_0)$ and $M^+ := M_G^+$ to be PDMs consistent with a graph G, which leads to $\pi(\boldsymbol{\Omega}|G, b, \boldsymbol{\Omega}_0) \propto Wishart(\boldsymbol{\Omega}|b, \boldsymbol{\Omega}_0) I(\boldsymbol{\Omega} \in M_G^+)$. Bayesian graphical lasso (Wang et al., 2012) assumes $\tilde{\pi}(\boldsymbol{\Omega})$ to be a product of independent exponential priors $Exp(\cdot|\lambda)$ and double-exponential priors $DE(\cdot|\lambda)$ on diagonal and off-diagonal elements of $\boldsymbol{\Omega}$, $\pi(\boldsymbol{\Omega}|\lambda) \propto \prod_{j<k} DE(\omega_{jk}|\lambda)$ $\prod_j Exp(\omega_{ii}|\lambda/2) I(\boldsymbol{\Omega} \in M^+)$. Graphical spike-and-slab prior (Wang, 2015) replaces the double-exponential priors in Bayesian graphical lasso by spike-and-slab priors, $\pi(\boldsymbol{\Omega}|G, v_1, v_0, \lambda)$ $\propto \prod_{\{j,k\} \in E} N(\omega_{jk}|0, v_1) \prod_{\{j,k\} \notin E} N(\omega_{jk}|0, v_0) \prod_j Exp(\omega_{ii}|\lambda/2)\ I(\boldsymbol{\Omega} \in M^+)$ where $v_1 \gg v_0$. Priors on $\boldsymbol{\Omega}$ can be defined either conditionally on the graph G or marginally; in what follows we do not use a model indicator parameter G but will infer the graph structures directly from the zero patterns in the precision matrices.

3. Gaussian graphical models with covariates

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be n realizations of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$. We assume an independent multivariate Gaussian distribution for each observation $\boldsymbol{y}_i \sim p(\boldsymbol{y}_i|\boldsymbol{\Omega}_i) = N(0, \boldsymbol{\Omega}_i^{-1})$ with the precision matrix $\boldsymbol{\Omega}_i = [\omega_{ijk}]$, importantly, indexed by $i = 1, \ldots, n$. A subject-level graph $G_i = (V, E_i)$ is embedded in the subject-level precision matrix $\boldsymbol{\Omega}_i$: $\{j, k\} \in E_i$ if and only if $\omega_{ijk} \neq 0$.

Without further modeling assumptions, $\boldsymbol{\Omega}_i$ cannot be estimated with a single observation i. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n realizations of covariates $\boldsymbol{X} = (1, X_1, \ldots, X_q)$. Note that when $\boldsymbol{X} = 1$ (i.e., there is no covaraites), the proposed GGMx is reduced to standard GGMs; more discussion of special cases of GGMx will be given later. We model $\boldsymbol{\Omega}_i \coloneqq \boldsymbol{f}(\boldsymbol{x}_i)$ through a symmetric matrix-valued function $\boldsymbol{f}(\cdot)$, which is estimable as a population-level parameter shared across all observations.

General construction of covariate-dependent priors. The key is the construction of the function $\boldsymbol{f}(\cdot) = [f_{jk}(\cdot)]$ such that $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ is a PDM for any $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Let $\mathcal{M}^+$ denote the collection of all such functions. This can be achieved by specifying a prior $\boldsymbol{f} \sim \Pi$ that assigns positive mass only on functions that satisfy such requirement, $\Pi(\mathcal{M}^+) = 1$. We consider the following generalization of the prior density in (1) as,

$$\pi(\boldsymbol{f}) = \frac{\tilde{\pi}(\boldsymbol{f}) I(\boldsymbol{f} \in \mathcal{M}^+)}{\int \tilde{\pi}(\boldsymbol{f}) I(\boldsymbol{f} \in \mathcal{M}^+) d\boldsymbol{f}}, \tag{2}$$

where $\tilde{\pi}$ is a distribution on matrix-valued functions. Note that the support of $\tilde{\pi}$ is not limited to $\mathcal{M}^+$, offering great flexibility in the choice of $\tilde{\pi}$. For example, we can start from independent distributions *a priori* such that $\tilde{\pi}(\boldsymbol{f}) = \prod_{j \leq k} \tilde{\pi}(f_{jk})$; using this construction, the marginal distribution $\tilde{\pi}(f_{jk})$ need not be defined with a constrained range. Because of the deterministic relationship $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$, prior $\pi(\boldsymbol{f})$ induces a conditional prior on $\boldsymbol{\Omega}_i$ given $\boldsymbol{x}_i$.

Two additional critical properties are desired for $\boldsymbol{f}(\cdot)$. (i) *Smoothness* — similar inputs should give rise to similar PDMs. Without smoothness, similar subjects may have vastly different networks, which is difficult to interpret in many applications including ours. (ii) *Sparsity* — $\pi(\boldsymbol{f})$ should have positive probability on sparse PDMs. Sparsity is a common assumption in high-dimensional models including GGMs, which improves statistical efficiency and interpretability compared to dense models. In order to encourage sparsity of $\boldsymbol{\Omega}_i$, a positive mass has to be placed on sparse PDMs *a priori* because otherwise there will be zero mass on sparse PDMs *a posteriori* even if data strongly favor sparse PDMs. To equip $\boldsymbol{f}(\cdot)$ with these two properties, we decompose each off-diagonal element $f_{jk}(\cdot)$ of $\boldsymbol{f}(\cdot)$ into two components,

$$f_{jk}(\boldsymbol{x}_i) = g_{jk}(\boldsymbol{x}_i) I(|g_{jk}(\boldsymbol{x}_i)| > t_{jk}), \quad \text{for} \quad j < k, \tag{3}$$

where $g_{jk}(\cdot)$ is some smooth function, the hard thresholding $I(|g_{jk}(\boldsymbol{x}_i)| > t_{jk})$ promotes sparsity in $f_{jk}(\cdot)$, and t_{jk} is a *random* threshold, which can be interpreted as a minimum effect size of ω_{ijk}. Specifically, whenever $g_{jk}(\boldsymbol{x}_i)$ is less than t_{jk} in magnitude, the hard thresholding truncates $f_{jk}(\boldsymbol{x}_i)$ to zero and hence induces a missing edge between nodes j and k for subject i. Our use of a thresholding function to induce sparsity on precision matrices $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ is novel and crucially different from conventional GGM priors including the G-Wishart prior and the graphical spike-and-slab prior (Wang, 2015): in order to construct observation-specific graphs, conventional priors would require a latent indicator for each potential edge and each observation, which would greatly increase the model complexity. For example, in our application with multiple myeloma dataset, conventional priors would need $n \cdot p \cdot (p-1)/2 = 79,728$ latent indicators whereas the proposed GGMx needs much fewer $p \cdot (p-1)/2 = 528$ thresholding parameters. Moreover, as will be introduced later, GGMx

enables undirected graph interpolation for unseen covariates, a new feature that is difficult to obtain with conventional priors. Other choices of thresholding functions are possible such as soft thresholding and nonnegative garrote thresholding. The main motivation of choosing hard thresholding over the alternatives is its theoretical connection with mixture of non-local priors; see Section 4.

For the diagonal elements (inverse-partial-variance, Whittaker 2009) $f_{jj}(\cdot)$ of $\boldsymbol{f}(\cdot)$, we assume the following model to ensure its nonnegativity,

$$f_{jj}(\boldsymbol{x}_i) = \exp\{g_{jj}(\boldsymbol{x}_i)\}. \tag{4}$$

Note that unlike off-diagonal elements in (3), the diagonal element $f_{jj}(\cdot)$ is not subject to thresholding.

Remark 1 *Our formulation encompasses covariate-dependent priors on both the off-diagonal (inverse-covariance) and diagonal (inverse-partial-variance) elements, thus conducting both graphical and inverse-partial-variance regression, simultaneously.*

Remark 2 *The proposed prior has two advantages over the more commonly used G-Wishart prior: (i) the induced prior on $\boldsymbol{\Omega}_i$ from (2) explicitly incorporates covariates $\boldsymbol{x}_i$ and (ii) the normalizing constant of G-Wishart is not a constant with respect to graph G and therefore comparing two graphs requires explicit evaluation of the intractable normalizing constant whereas $\pi(\boldsymbol{f})$, due to the thresholding function, does not have such complication.*

Given $\boldsymbol{f}(\cdot)$ and $\boldsymbol{X}$, the proposed GGMx satisfies *functional Markov properties*, e.g., the pairwise functional Markov property, which is stated formally in the following lemma.

Lemma 1 *If $f_{jk}(\boldsymbol{X}) = 0$, then $Y_j \perp\!\!\!\perp Y_k | \boldsymbol{Y}_{rest}, \boldsymbol{X}$ where $\boldsymbol{Y}_{rest}$ is the subvector of $\boldsymbol{Y}$ without Y_j and Y_k.*

The proof of Lemma 1 directly follows from the fact that $f_{jk}(\boldsymbol{X}) = 0$ implies there is a missing edge between nodes j and k given covariates $\boldsymbol{X}$, which in turn implies that $Y_j \perp\!\!\!\perp Y_k | \boldsymbol{Y}_{rest}, \boldsymbol{X}$ from standard GGM theory.

A natural choice of $g_{jk}(\cdot)$ is a linear function $g_{jk}(\boldsymbol{x}_i) = \boldsymbol{\beta}_{jk}^T \boldsymbol{x}_i$ although, in general, $g_{jk}(\cdot)$ can be any smooth function. Given the limited sample size of the case study, we consider $g_{jk}(\cdot)$ to be linear for parsimony (see Section 8 for a brief discussion on modeling a nonlinear g_{jk}) and interpretability ($\boldsymbol{\beta}_{jk}$ are the rates of changes of ω_{ijk} in $\boldsymbol{x}_i$). If the focus is on learning the graph structure and strength, i.e., the off-diagonal elements of $\boldsymbol{\Omega}_i$, one can further simplify the model by reducing diagonal elements $g_{jj}(\boldsymbol{x}_i)$ to be constant with respect to the covariates.

GGMx is a fairly flexible class of models and has at least five special cases (see Table 1). (i) If $\boldsymbol{X}$ only contains the intercept, then GGMx reduces to the case of the standard GGM because the graph is a function of a constant and hence is constant. (ii) If $\boldsymbol{X}$ is categorical, then GGMx is a multiple graphical model (also known as group-specific GGM) as the categorical covariate defines the groups. (iii) If $\boldsymbol{X}$ is univariate time points, then GGMx can be used for modeling time-varying GGMs (Zhou et al., 2010) by treating time as a covariate[1]. (iv) If the thresholds t_{jk}'s are fixed to 0, then GGMx is a covariate-dependent

1. This is a conceptual statement. Note that existing time-varying GGM methods typically assume graphs to vary non-linearly with time whereas this paper considers linearly-varying graphs.

GGM in which the strength of the graph varies continuously with the covariates but the structure is constant because a non-zero linear function is non-zero almost everywhere. (v) If $\boldsymbol{X}$ is a subset of $\boldsymbol{Y}$, then GGMx can be interpreted as a context-specific GGM (Nyman et al., 2017) where the graph structure varies with (discretized) $\boldsymbol{X}$.

Table 1: Five special cases of GGMx.

Special cases of GGMx	**Conditions**	**Mapping $\boldsymbol{X} \mapsto \boldsymbol{\Omega}$**	
Standard GGM	$\boldsymbol{x}_i = 1$	$g_{jk}(\boldsymbol{x}_i) = \beta_{jk}$	$\omega_{ijk} = \beta_{jk} I(\lvert\beta_{jk}\rvert > t_{jk})$
Group-specific GGM	$\boldsymbol{x}_i = c$, $c \in \{1, \ldots, C\}$	$g_{jk}(\boldsymbol{x}_i) = \beta_{jkc}$	$\omega_{ijk} = \beta_{jkc} I(\lvert\beta_{jkc}\rvert > t_{jk})$
Time-varying GGM	$\boldsymbol{x}_i = x$, time $x \in \mathbb{R}$	$g_{jk}(x) = x \cdot \beta_{jk}$	$\omega_{ijk} = x \cdot \beta_{jk} I(\lvert x \cdot \beta_{jk}\rvert > t_{jk})$
Covariate-dependent GGM	$t_{jk} = 0$	$g_{jk}(\boldsymbol{x}_i) = \boldsymbol{\beta}_{jk}^T \boldsymbol{x}_i$	$\omega_{ijk} = \boldsymbol{\beta}_{jk}^T \boldsymbol{x}_i$
Context-specific GGM	$\boldsymbol{x}_i = y_{i1}$	$g_{jk}(\boldsymbol{x}_i) = y_{i1} \cdot \beta_{jk}$	$\omega_{ijk} = y_{i1} \cdot \beta_{jk} I(\lvert y_{i1} \cdot \beta_{jk}\rvert > t_{jk})$

Priors. We assign priors to $\boldsymbol{\beta}_{jk}$ and t_{jk}, which in turn define $\tilde{\pi}(\boldsymbol{f})$. We assume an independent multivariate Gaussian prior $\boldsymbol{\beta}_{jk} \sim \pi(\boldsymbol{\beta}_{jk}) = N(\boldsymbol{\beta}_{jk} | 0, \tau_{jk} \boldsymbol{I}_q)$. The thresholding parameter t_{jk} can be interpreted as the minimum size of off-diagonal elements of $\boldsymbol{\Omega}_i$. Since its value is usually unknown in practice, we assign a truncated normal prior $t_{jk} \sim \pi(t_{jk}) = N(\mu_t, \sigma_t^2) I(t_{jk} > 0)$ to reflect the uncertainty. As we will show in the next section, the priors of $\boldsymbol{\beta}_{jk}$ and t_{jk} induce a mixture of non-local priors on $\boldsymbol{\Omega}_i$.

To complete the prior formulation, for the hyperparamter τ_{jk}, we assign a hyperprior

$$\boldsymbol{\tau} = \{\tau_{jk}\}_{j \le k} \sim \pi(\boldsymbol{\tau}) = \frac{C_{\boldsymbol{\tau}} \prod_{j \le k} IG(\tau_{jk} | a_\tau, b_\tau)}{\int C_{\boldsymbol{\tau}} \prod_{j \le k} IG(\tau_{jk} | a_\tau, b_\tau) d\boldsymbol{\tau}}$$

where $IG(a, b)$ denotes an inverse-gamma density with shape a and scale b, and $C_{\boldsymbol{\tau}}$ is the normalizing constant in (2),

$$C_{\boldsymbol{\tau}} = \int \tilde{\pi}(\boldsymbol{f}) I(\boldsymbol{f} \in \mathcal{M}^+) d\boldsymbol{f}.$$

Including $C_{\boldsymbol{\tau}}$ in the prior of $\pi(\boldsymbol{\tau})$ serves to cancel out $C_{\boldsymbol{\tau}}^{-1}$ in (2) so that the full conditional of τ_{jk} is inverse-gamma. Similar cancellation trick has been used and thoroughly investigated in Bayesian graphical lasso (Wang et al., 2012).

A schematic representation of the proposed GGMx is provided in Figure 2.

4. Theoretical Properties

We establish a general result of the connection between the proposed prior of precision matrices induced by (2) and (3) and non-local alternative priors in GGM. A non-local prior assigns a vanishing density (under the alternative hypothesis) to the neighborhood of the null hypothesis. In variable selection contexts, this density vanishes around 0 and therefore shrinks small effect to zero, which is appealing because we are interested in a parsimonious estimation of the graph (i.e., a sparse network). Non-local priors have been shown, both theoretically and empirically, to have superior performance over local priors in

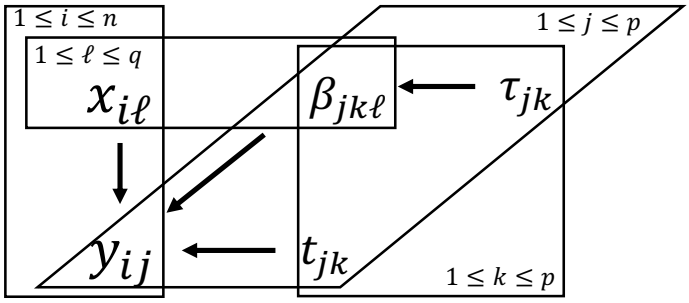


Figure 2: A schematic representation of GGMx.

various applications including hypothesis testing, high-dimensional sparse regression, and Bayesian networks (Johnson and Rossell, 2010, 2012; Altomare et al., 2013; Rossell and Telesca, 2017; Shin et al., 2018; Ni et al., 2019). However, to the best of our knowledge, all existing priors of sparse precision matrices in GGM (G-Wishart, Bayesian graphical lasso, and stochastic search structure learning prior) are local, i.e., $\pi(\boldsymbol{\Omega})$ does not approach 0 as $\omega_{jk} \to 0$ for $(j,k) \in E$. Conceptually, local priors have a seemingly "contradictory" representation of one's prior belief. On the one hand, $(j,k) \in E$ suggests ω_{jk} is non-zero. But on the other hand, local priors fail to assign zero mass at $\omega_{jk} = 0$; in fact, local priors often assign the maximum mass at zero. The practical implication of such "contradiction" is that local priors tend to favor denser models and be more susceptible to false discoveries compared to non-local priors especially for high-dimensional models like GGMx.

Let π_θ and π_t generically denote the priors for θ_{jk} and t_{jk}, $\theta_{jk} \sim \pi_\theta(\theta_{jk})$ and $t_{jk} \sim \pi_t(t_{jk})$. Let $\boldsymbol{T} = [t_{jk}]$. We now show the connection between non-local priors and the proposed prior of the following general form,

$$\pi(\boldsymbol{\Omega}|\boldsymbol{T}) = \frac{\tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{T}) I(\boldsymbol{\Omega} \in M^+)}{\int \tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{T}) I(\boldsymbol{\Omega} \in M^+) d\boldsymbol{\Omega}},$$

and

$$\tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{T}) = \prod_{j=1}^{p} \pi_d(\omega_{jj}) \prod_{j<k} \pi_{\omega|t}(\omega_{jk}|t_{jk}),$$

where

$$\omega_{jk} = \theta_{jk} I(|\theta_{jk}| > t_{jk}), \text{ for } j < k.$$

Note that the equations above have no reference to covariates. We deliberately do so for clarity and generality; all the following theoretical results apply to the marginal distribution $\pi(\boldsymbol{\Omega}_i)$ in GGMx by letting $\theta_{jk} = g_{jk}(\boldsymbol{x}_i) = \boldsymbol{\beta}_{jk}^T \boldsymbol{x}_i$ and $\omega_{jj} = \exp\{g_{jj}(\boldsymbol{x}_i)\}$. Conditional on t_{jk}, the prior π_θ induces a spike-and-slab mixture distribution,

$$\pi_{\omega|t}(\omega_{jk}|t_{jk}) = \rho \delta_0(\omega_{jk}) + (1-\rho) \tilde{\pi}_{\omega|t}(\omega_{jk}|t_{jk})$$

where the mixture weight $\rho = Pr(|\omega_{jk}| < t_{jk}|t_{jk})$ is computed under the conditional distribution of ω_{jk} induced by $\pi_\theta(\cdot)$ and hence is a function of t_{jk} (not ω_{jk}), and the slab is a truncated distribution,

$$\tilde{\pi}_{\omega|t}(\omega_{jk}|t_{jk}) = \frac{\pi_\theta(\omega_{jk}) I(|\omega_{jk}| > t_{jk})}{Pr(|\omega_{jk}| > t_{jk}|t_{jk})}.$$

Slightly abusing the notations, let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M) = (\omega_{12}, \ldots, \omega_{1p}, \omega_{23}, \ldots, \omega_{2p}, \ldots, \omega_{p-1,p})$ be an M-dimensional vector containing upper-triangular elements of $\boldsymbol{\Omega}$ with $M = \binom{p}{2}$. Let $S \subseteq \{1, \ldots, M\}$ denote the indices of non-zeros elements in $\boldsymbol{\Omega}$ (or equivalently in $\boldsymbol{\omega}$), i.e. $\omega_m = 0$ if and only if $m \in S^c$. Then the conditional prior of $\boldsymbol{\Omega}$ given $\boldsymbol{T}$ can be written as a mixture over all possible subsets S,

$$\pi(\boldsymbol{\Omega}|\boldsymbol{T}) = \frac{1}{g(\boldsymbol{T})} I(\boldsymbol{\Omega} \in M^+) \prod_{j=1}^{p} \pi_d(\omega_{jj}) \tag{5}$$

$$\times \sum_{S \in 2^{\{1,\ldots,M\}}} \prod_{m \in S} \pi_\theta(\omega_m) I(|\omega_m| > t_m) \prod_{m \in S^c} Pr(|\omega_m| < t_m | t_m) \delta_0(\omega_m), \tag{6}$$

where $g(\boldsymbol{T}) = \int \tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{T}) I(\boldsymbol{\Omega} \in M^+) d\boldsymbol{\Omega}$ is the normalizing constant and $2^{\{1,\ldots,M\}}$ is the power set of $\{1, \ldots, M\}$. Our main theorem shows that under very mild conditions, the marginal prior $\pi(\boldsymbol{\Omega})$ is a discrete mixture of non-local priors. Before we present the main theorem, we first state a lemma that is useful in proving the theorem.

Lemma 2 *$E[1/g(\boldsymbol{T})] < \infty$ if the distribution $\pi_\theta(\cdot)$ of θ_{jk} has positive mass around zero, i.e., there exists $\delta > 0$ such that for any $0 < \delta' < \delta$, $\int_{-\delta'}^{\delta'} \pi_\theta(\theta) d\theta > 0$, and the distribution $\pi_d(\cdot)$ of ω_{jj} is not a point mass at zero, i.e., $\pi_d(\cdot) \neq \delta_0(\cdot)$.*

Proof Consider

$$\begin{aligned} g(\boldsymbol{T}) &= \int \tilde{\pi}(\boldsymbol{\Omega}|\boldsymbol{T}) I(\boldsymbol{\Omega} \in M^+) d\boldsymbol{\Omega} = Pr(\boldsymbol{\Omega} \in M^+ | \boldsymbol{T}) \\ &> Pr(\{\omega_{jj} > (p-1)\lambda\}_{j=1}^p, \{|\omega_{jk}| \leq \lambda\}_{j<k} | \boldsymbol{T}), \quad \forall \lambda \geq 0 \\ &> Pr(\{\omega_{jj} > (p-1)\lambda\}_{j=1}^p, \{|\theta_{jk}| \leq \lambda\}_{j<k}) \\ &= \prod_{j=1}^{p} Pr(\omega_{jj} > (p-1)\lambda) \prod_{j<k} Pr(|\theta_{jk}| \leq \lambda) \stackrel{\text{def}}{=} L(\lambda). \end{aligned}$$

The first inequality holds because diagonally dominant symmetric matrix is positive definite and the second inequality is true because $|\theta_{jk}| \leq \lambda$ implies $|\omega_{jk}| \leq \lambda$ by design and $\boldsymbol{T}$ is independent of ω_{jj} and θ_{jk}. If π_θ has positive mass around zero and π_d is not a point mass at zero, we can pick a sufficiently small (but positive) $\lambda^\star > 0$ such that the lower bound $L(\lambda^\star)$ of $g(\boldsymbol{T})$ is positive. Then it follows that $E[1/g(\boldsymbol{T})] < \infty$. ■

Theorem 1 *The marginal prior $\pi(\boldsymbol{\Omega})$ is given by*

$$\pi(\boldsymbol{\Omega}) = \sum_{S \in 2^{\{1,\ldots,M\}}} \rho_S \pi_S(\boldsymbol{\Omega}),$$

where $\pi_S(\boldsymbol{\Omega})$ is the prior under the hypothesis $H_S : \omega_m \neq 0, m \in S$ and $\omega_m = 0, m \in S^c$. Moreover, $\pi_S(\boldsymbol{\Omega})$ is a non-local prior for any $S \in 2^{\{1,\ldots,M\}} \backslash \emptyset$, that is, $\pi_S(\boldsymbol{\Omega}) \to 0$ as $\omega_m \to 0$ for $m \in S$, provided (i) $Pr(t = 0) = 0$, (ii) $\pi_\theta(\cdot)$ is bounded and has positive mass near 0, and (iii) $\pi_d(\cdot) \neq \delta_0(\cdot)$.

Proof The marginal distribution of $\boldsymbol{\Omega}$ is given by

$$\pi(\boldsymbol{\Omega}) = \int \pi(\boldsymbol{\Omega}|\boldsymbol{T})\pi_t(\boldsymbol{T})d\boldsymbol{T} = I(\boldsymbol{\Omega} \in M^+) \prod_{j=1}^{p} \pi_d(\omega_{jj}) \int \frac{1}{g(\boldsymbol{T})} \prod_{j<k} \pi_{\omega|t}(\omega_{jk}|t_{jk})\pi_t(\boldsymbol{T})d\boldsymbol{T}.$$

Let $m(\boldsymbol{\Omega}) = I(\boldsymbol{\Omega} \in M^+) \prod_{j=1}^{p} \pi_d(\omega_{jj})$, then

$$\begin{aligned}
\pi(\boldsymbol{\Omega}) &= \int m(\boldsymbol{\Omega}) \frac{1}{g(\boldsymbol{T})} \prod_{j<k} \pi_{\omega|t}(\omega_{jk}|t_{jk})\pi_t(\boldsymbol{T})d\boldsymbol{T} \\
&= \int m(\boldsymbol{\Omega}) \frac{1}{g(\boldsymbol{T})} \prod_{j<k} \{Pr(|\omega_{jk}| < t_{jk}|t_{jk})\delta_0(\omega_{jk}) + \pi_\theta(\omega_{jk})I(|\omega_{jk}| > t_{jk})\}\pi_t(t_{jk})d\boldsymbol{T} \\
&= \int m(\boldsymbol{\Omega}) \frac{1}{g(\boldsymbol{T})} \sum_{S \in 2^{\{1,\ldots,M\}}} \prod_{m \in S} \pi_\theta(\omega_m) I(|\omega_m| > t_m)\pi_t(t_m) \prod_{m \in S^c} Pr(|\omega_m| < t_m|t_m)\delta_0(\omega_m)\pi_t(t_m)d\boldsymbol{T} \\
&= \sum_{S \in 2^{\{1,\ldots,M\}}} m(\boldsymbol{\Omega}) E_{\boldsymbol{T}} \left[\frac{1}{g(\boldsymbol{T})} \prod_{m \in S} I(|\omega_m| > t_m) \prod_{m \in S^c} Pr(|\omega_m| < t_m|t_m) \right] \prod_{m \in S} \pi_\theta(\omega_m) \prod_{m \in S^c} \delta_0(\omega_m) \\
&\stackrel{\text{def}}{=} \sum_{S \in 2^{\{1,\ldots,M\}}} h_S(\boldsymbol{\Omega}) = \sum_{S \in 2^{\{1,\ldots,M\}}} \int h_S(\boldsymbol{\Omega})d\boldsymbol{\Omega} \times \frac{h_S(\boldsymbol{\Omega})}{\int h_S(\boldsymbol{\Omega})d\boldsymbol{\Omega}} \stackrel{\text{def}}{=} \sum_{S \in 2^{\{1,\ldots,M\}}} \rho_S \times \pi_S(\boldsymbol{\Omega}).
\end{aligned}$$

We will show that for any $m \in S$ and any sequence $\omega_m^{(n)} \to 0$ as $n \to \infty$, $\pi_S(\boldsymbol{\Omega}^{(n)}) \to 0$ as $n \to \infty$ where $\boldsymbol{\Omega}^{(n)}$ contains $\omega_m^{(n)}$ as an element. Note that

$$\frac{1}{g(\boldsymbol{T})} \prod_{m \in S} I(|\omega_m| > t_m) \prod_{m \in S^c} Pr(|\omega_m| < t_m|t_m) \leq \frac{1}{g(\boldsymbol{T})}.$$

Since $E[1/g(\boldsymbol{T})] < \infty$ due to conditions (ii) - (iii) and Lemma 2, and $\lim_{n\to\infty} I(|\omega_m^{(n)}| > t_m) = 0$ almost surely due to condition (i), then by dominated convergence theorem, we have

$$\begin{aligned}
&\lim_{n\to\infty} E_{\boldsymbol{T}} \left[\frac{1}{g(\boldsymbol{T})} I(|\omega_m^{(n)}| > t_m) \prod_{m' \in S, m' \neq m} I(|\omega_{m'}| > t_{m'}) \prod_{m \in S^c} Pr(|\omega_m| < t_m|t_m) \right] \\
=&E_{\boldsymbol{T}} \left[\frac{1}{g(\boldsymbol{T})} \left\{ \lim_{n\to\infty} I(|\omega_m^{(n)}| > t_m) \right\} \prod_{m' \in S, m' \neq m} I(|\omega_{m'}| > t_{m'}) \prod_{m \in S^c} Pr(|\omega_m| < t_m|t_m) \right] \\
=&0
\end{aligned}$$

Finally, condition (ii) renders $\pi_S(\boldsymbol{\Omega}^{(n)}) \to 0$. ■

Conditions (i) - (iii) in Theorem 1 are very mild and satisfied by a wide range of π_t, π_θ, and π_d. Condition (i) is trivially satisfied if π_t is continuous (e.g., gamma, inverse-gamma, log-normal, and truncated normal distributions). Condition (ii) holds for Cauchy, normal, and most of the scale mixtures of normal distributions such as Laplace, normal-gamma, and t distributions. Condition (iii) only excludes point mass at zero $\delta_0(\cdot)$ from all the possible choices of $\pi_d(\cdot)$.

A simple illustrative example As a concrete example, $\pi_S(\mathbf{\Omega})$ is non-local under the prior distributions specified in Section 3, namely, $\pi_\theta(\theta_{jk}) = N(0, \tau)$, $\pi_d(\omega_{jj}) = \text{log-normal}(0, \tau)$, and $\pi_t(t_{jk}) = N(\mu_t, \sigma_t^2)I(t_{jk} > 0)$. To visualize the proposed non-local prior, we consider a small precision matrix with $p = 3$ and perform a prior simulation to generate $\mathbf{\Omega}$ from $\pi(\mathbf{\Omega})$, the procedure of which is a special case of the posterior simulation procedure (ignoring the likelihood) to be described in Section 5. We visualize $\pi_S(\mathbf{\Omega})$ for $S = \{1, \ldots, M\}$, i.e., a complete graph. The marginal densities of pairs of off-diagonal elements of $\mathbf{\Omega}$ (normalized to partial correlations) are depicted in the top panel of Figure 3 which show vanishing density as ω_{jk} approaches 0. By contrast, a local prior on $\mathbf{\Omega}$ (simulated by fixing $t_{jk} = 0$) has an increasing density as ω_{jk} approaches 0 as shown in the bottom panel of Figure 3.

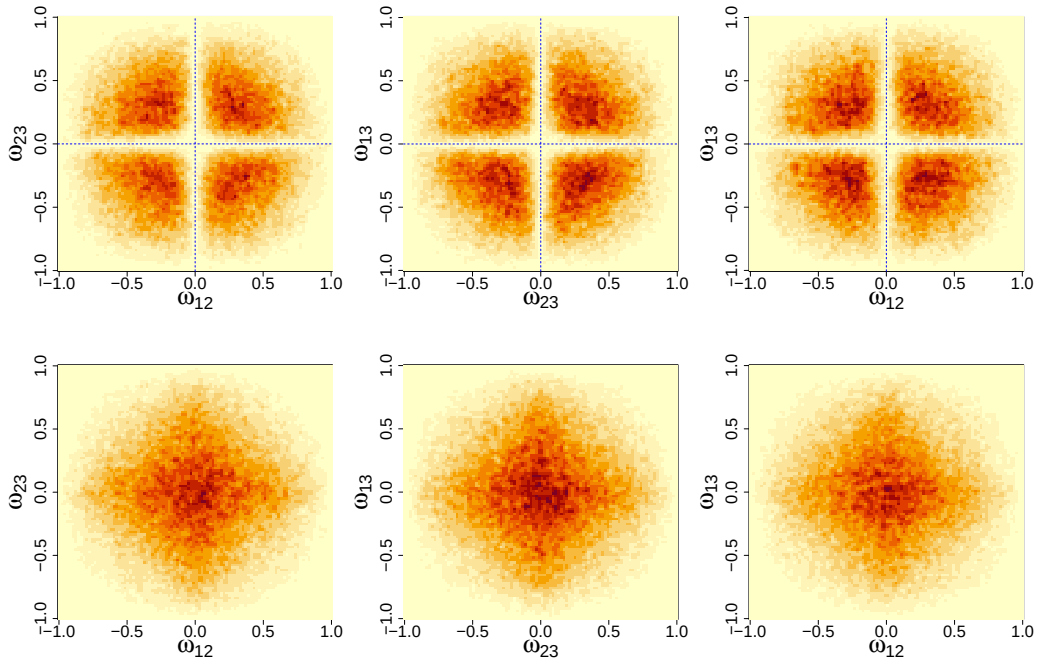


Figure 3: Non-local (top) and local (bottom) prior distributions of $\mathbf{\Omega}$.

Remark The connection between non-local priors and random thresholding has been investigated in the regression context (Rossell and Telesca, 2017; Ni et al., 2019). We make a nontrivial extension to precision matrix estimation for undirected GGMs. One major difference between our theory and those in Rossell and Telesca (2017) and Ni et al. (2019) is the complexity of the intractable prior normalizing constant $g(\boldsymbol{T})$ in (5). Intractable prior normalizing constant is a common challenge in standard Bayesian GGMs (Dobra et al., 2011; Wang et al., 2012; Wang, 2015), both theoretically and computationally. In order to show the equivalence between non-local priors and random thresholding for GGMs, we make extra assumptions, i.e., $\pi_\theta(\cdot)$ has positive mass around zero and $\pi_d(\cdot) \neq \delta_0(\cdot)$, in order to bound $E[1/g(\boldsymbol{T})]$. These mild assumptions are not required in previous works. Also note that Rossell and Telesca (2017) truncates probability density whereas we threshold the random variables. Consequently, the resulting marginal prior of Rossell and Telesca (2017)

is a non-local prior while ours is a discrete mixture of the non-local prior and point mass at 0. Computationally, the issue of intractable normalizing constant is resolved by a carefully designed MCMC algorithm, which will be discussed in the next section.

5. Posterior Inference

The proposed GGMx is parameterized by three sets of parameters $\{\boldsymbol{\beta}_{jk}\}_{j\leq k}$, $\{t_{jk}\}_{j<k}$, and $\{\tau_{jk}\}_{j\leq k}$. The joint posterior distribution of these parameters is given by,

$$\begin{aligned}&p(\{\boldsymbol{\beta}_{jk}\}_{j\leq k}, \{t_{jk}\}_{j<k}, \{\tau_{jk}\}_{j\leq k}|\{\boldsymbol{y}_i, \boldsymbol{x}_i\}_{i=1}^n)\\ \propto&\prod_{i=1}^n N(\boldsymbol{y}_i|0, \boldsymbol{\Omega}_i)\prod_{j<k} N(t_{jk}|\mu_t, \sigma_t^2)I(t_{jk}>0)\prod_{j\leq k} N(\boldsymbol{\beta}_{jk}|0, \tau_{jk}\boldsymbol{I}_q)IG(\tau_{jk}|a_\tau, b_\tau),\end{aligned}$$

where the right-hand side of this equation depends on $\boldsymbol{x}_i$ through $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ and $\boldsymbol{f}(\cdot)$ is defined by $\{\boldsymbol{\beta}_{jk}\}_{j\leq k}$ and $\{t_{jk}\}_{j<k}$. The posterior inference of the model parameters is carried out by MCMC. We need to carefully choose a proposal distribution that can propose $\boldsymbol{f} \in \mathcal{M}^+$ efficiently. This is not a trivial task because the probability that we generate $\boldsymbol{f} \in \mathcal{M}^+$ is practically zero if we propose $\boldsymbol{\beta}_{jk}$ and t_{jk} from naive proposals such as standard random walks. Here, we introduce a proposal that always proposes $\boldsymbol{f} \in \mathcal{M}^+$.

For illustration, suppose we are currently updating the (j,k)th element of $\boldsymbol{\Omega}_i$. Let $\boldsymbol{\omega}_{i,-k,k}$ denote the kth column of $\boldsymbol{\Omega}_i$ without the kth row and let $\boldsymbol{\Omega}_{i,-k,-k}$ denote the submatrix of $\boldsymbol{\Omega}_i$ without the kth row and column. Let $\phi_{ik} = \omega_{ikk} - u_{ik}$ with $u_{ik} = \boldsymbol{\omega}_{i,-k,k}^T \boldsymbol{\Omega}_{i,-k,-k}^{-1}\boldsymbol{\omega}_{i,-k,k}$. We first propose new $\beta_{jk\ell}^*$ and t_{jk}^* from some proposal densities $q_\beta(\beta_{jk\ell}^*|\beta_{jk\ell})$ and $q_t(t_{jk}^*|t_{jk})$ such as random walks for $\ell = 1, \ldots, q+1$. The resulting new values of $\boldsymbol{\omega}_{i,-k,k}$ and u_{ik} are denoted by $\boldsymbol{\omega}_{i,-k,k}^*$ and u_{ik}^*. Notice that $\boldsymbol{\Omega}_i$ is positive definite if and only if $\phi_{ik} > 0$ for $k = 1, \ldots, p$. This is due to the Sylvester's criterion that a symmetric matrix is positive definite if and only if all of the leading principal minors are positive. Without loss of generality, assuming k is the last column and all previous principal minors are positive, and assuming covariates $\boldsymbol{x}_i$ are positive. Then the last leading principal minor $\det(\boldsymbol{\Omega}_i) = (\omega_{ikk} - u_{ik})\det(\boldsymbol{\Omega}_{i,-k,-k})$ is positive if and only if $\omega_{ikk} - u_{ik} > 0$. Therefore, in order to ensure positive definiteness of $\boldsymbol{\Omega}_i, \forall i$ when updating its (j,k)th element, we will additionally propose a new $\beta_{kk\ell}^*$ such that $\omega_{ikk}^* = \exp\{g_{kk}^*(\boldsymbol{x}_i)\} > u_{ik^*}$ where $g_{kk}^*(\boldsymbol{x}_i) = x_{i\ell}\beta_{kk\ell}^* + \sum_{\ell'\neq\ell} x_{i\ell'}\beta_{kk\ell'}$. The solution to this inequality for all i is the constraint that the proposal of $\beta_{kk\ell}^*$ needs to respect. Specifically, we will propose $\beta_{kk\ell}^* \sim q_\beta(\beta_{kk\ell}^*|\beta_{kk\ell})I(\beta_{kk\ell}^* \in S_{k\ell}^*)$ where

$$S_{k\ell}^* = \left\{\beta \middle| \beta > \max_i \left(\frac{\log(u_{ik}^*) - \sum_{\ell'\neq\ell} x_{i\ell'}\beta_{kk\ell'}}{x_{i\ell}}\right)\right\}.$$

We summarize the property of the proposal density in the following proposition, of which the proof is given by the proceeding paragraph.

Proposition 1 *The proposal density* $q(\beta_{jk\ell}^*, t_{jk}^*, \beta_{kk\ell}^*|\beta_{jk\ell}, t_{jk}, \beta_{kk\ell}) = q_\beta(\beta_{jk\ell}^*|\beta_{jk\ell})q_t(t_{jk}^*|t_{jk})$ $q_\beta(\beta_{kk\ell}^*|\beta_{kk\ell})I(\beta_{kk\ell}^* \in S_{k\ell}^*)$ *and the full conditional density* $p(\beta_{jk\ell}^*, t_{jk}^*, \beta_{kk\ell}^*|\cdot)$ *have the same support.*

We now provide the MCMC (Metropolis-within-Gibbs) algorithm below; its validity is guaranteed by Proposition 1 and standard MCMC theory.

The MCMC Algorithm. Initialize model parameters. Repeat the following steps until practical convergence.

(I) Update precision matrices $\mathbf{\Omega}_i$. Scanning through each column $k = 1, \ldots, p$, each row $j \neq k$, and each covariate $\ell = 1, \ldots, q+1$, we propose $\beta^*_{jk\ell}$, t^*_{jk}, and $\beta^*_{kk\ell}$ from $q_\beta(\beta^*_{jk\ell}|\beta_{jk\ell})$, $q_t(\log t^*_{jk}|\log t_{jk})$, and $q_\beta(\beta^*_{kk\ell}|\beta_{kk\ell})I(\beta^*_{kk\ell} \in S^*_{k\ell})$ where $q_t(\log t^*_{jk}|\log t_{jk}) = N(\log t^*_{jk}|\log t_{jk}, \eta_t^2)$, $q_\beta(\beta^*_{jk\ell}|\beta_{jk\ell}) = N(\beta^*_{jk\ell}|\beta_{jk\ell}, \eta_\beta^2)$, and $q_\beta(\beta^*_{kk\ell}|\beta_{kk\ell}) = N(\beta^*_{kk\ell}|\beta_{kk\ell}, \eta_\beta^2)$. We accept the proposal with probability $\min(1, \alpha)$ where

$$\alpha = \frac{\prod_{i=1}^n p(\boldsymbol{y}_i|\mathbf{\Omega}^*_i)\pi(\beta^*_{jk\ell})\pi(t^*_{jk})\pi(\beta^*_{kk\ell})q_\beta(\beta_{jk\ell}|\beta^*_{jk\ell})q_t(t_{jk}|t^*_{jk})q_\beta(\beta_{jk\ell}|\beta^*_{jk\ell})I(\beta_{kk\ell} \in S_{k\ell})}{\prod_{i=1}^n p(\boldsymbol{y}_i|\mathbf{\Omega}_i)\pi(\beta_{jk\ell})\pi(t_{jk})\pi(\beta_{kk\ell})q_\beta(\beta^*_{jk\ell}|\beta_{jk\ell})q_t(t^*_{jk}|t_{jk})q_\beta(\beta^*_{kk\ell}|\beta_{kk\ell})I(\beta^*_{kk\ell} \in S^*_{k\ell})}.$$

The proposal standard deviations η_t^2 and η_β^2 can be set to achieve desired acceptance rate (say, 20%-40%).

(II) Update the hypervariances τ_{jk} from the inverse-gamma full conditional, $\tau_{jk} \sim IG(a_\tau + 1/2, b_\tau + \beta^2_{jk\ell}/2)$.

Graph estimation. A point estimate of G_i can be obtained by thresholding the posterior probability of inclusion. Specifically, we select $\{j,k\} \in E_i$ if $Pr(\{j,k\} \in E_i|\boldsymbol{y}_i, \boldsymbol{x}_i) > c$ where $c \in [0,1]$ is the probability cutoff[2] . The posterior probability of inclusion can be approximated by the MCMC samples,

$$Pr(\{j,k\} \in E_i|\boldsymbol{y}_i, \boldsymbol{x}_i) = Pr(\omega_{ijk} \neq 0|\boldsymbol{y}_i, \boldsymbol{x}_i) \approx \frac{1}{R}\sum_{r=1}^R I\{\omega^{(r)}_{ijk} \neq 0\},$$

where the superscript (r) indexes the posterior samples.

Graph interpolation. Since the precision matrix $\mathbf{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ is modeled as a function of $\boldsymbol{x}_i$, we can interpolate a graph $G^* = (V, E^*)$ for an unseen observation at covariates $\boldsymbol{x}^*$. It is achieved through the posterior predictive distribution of $\boldsymbol{f}(\cdot)$, which can be approximated by the MCMC samples,

$$Pr(\{j,k\} \in E^*|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}^*) = Pr\{f_{jk}(\boldsymbol{x}^*) \neq 0|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{x}^*\} \approx \frac{1}{R}\sum_{r=1}^R I\{f^{(r)}_{jk}(\boldsymbol{x}^*) \neq 0\}.$$

Graph interpolation requires covariates $\boldsymbol{x}^*$ only, since the right-hand side of the equation above does not depend on $\boldsymbol{y}^*$. In practice, this is a desirable property. For example, one can predict the gene network for new patients without sequencing the whole genome; the measurement of covariates (e.g., blood biomarkers) will suffice.

6. Simulations

6.1 Simulation Setup

We assessed the utility and operating characteristics of GGMx in seven simulation scenarios with different levels of sparsity and types of covariates. The same size of the dataset in

2. Note that the probability cutoff c is different from the random threshold t_{jk}. The random threshold is a model parameter used to induce sparsity whereas the probability cutoff is introduced to obtain posterior point estimate of graphs.

application was used: $n = 151$, $p = 33$, and $q = 2$ (q was set to 1 for the last scenario). Note that even with a moderate dataset, the number of parameters $(\boldsymbol{\beta}_{jk}, t_{jk}, \tau_{jk})$ that need to be estimated is $\frac{p(p+1)(q+2)}{2} + \frac{p(p-1)}{2}$ =2,772, which is substantially larger than the sample size. We focused on graph structure learning in the first five scenarios by assuming constant diagonal elements $g_{jj}(\cdot)$ for simplicity; non-constant case (i.e., simultaneous inverse-partial-variance and graphical regression) will be considered in the last two scenarios. We fixed the probability cutoff c to be 0.5 in all scenarios.

Scenario I. We generated the simulated data from our model. We randomly set 2% of $\beta_{jk\ell}$ for $j < k$ to be ± 1 with equal probability. We set $t_{jk} = 0.5$ and all the diagonal elements of $\boldsymbol{\Omega}_i$ to be 1. The covariate x_{ij} was generated from an uniform distribution $x_{ij} \overset{iid}{\sim} U(-1, 1)$. The resulting precision matrix $\boldsymbol{\Omega}_i$ might not be positive definite for all observations $i = 1, \ldots, n$. We repeated the process until $\boldsymbol{\Omega}_i > 0, \forall i$. Then the observation $\boldsymbol{y}_i$ was drawn from normal $\boldsymbol{y}_i \overset{ind}{\sim} N(0, \boldsymbol{\Omega}_i^{-1})$. Using the same procedure, we generated a similar independent dataset with sample size 50 for testing graph interpolation of GGMx.

Scenario II. The procedure in Scenario I was inefficient to generate a denser network. In addition, it may not mimic well the data in application. In this scenario, we used one posterior draw from GGMx applied to the multiple myeloma data as simulation truth. The true $\beta_{jk\ell}$'s are shown as heatmaps in Figure 4a where $\ell = 1$ corresponds to the intercept and $\ell = 2, 3$ correspond to the two covariates. Since the heatmap of β_{jk1} is denser than those of β_{jk2} and β_{jk3}, there were more nearly constant edges than highly varying edges. The true t_{jk}'s are shown in Figure 4b. The covariates $\boldsymbol{x}_i$ of the multiple myeloma dataset was used. And $\boldsymbol{y}_i$ was drawn from the model $\boldsymbol{y}_i \overset{ind}{\sim} N(0, \boldsymbol{\Omega}_i^{-1})$ with $\boldsymbol{\Omega}_i = \boldsymbol{f}(\boldsymbol{x}_i)$.

Scenario III. This scenarios considered a simulation truth from an ordinary GGM, i.e. $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}, \forall i$. We generated a true $\boldsymbol{\Omega}$ as follows.

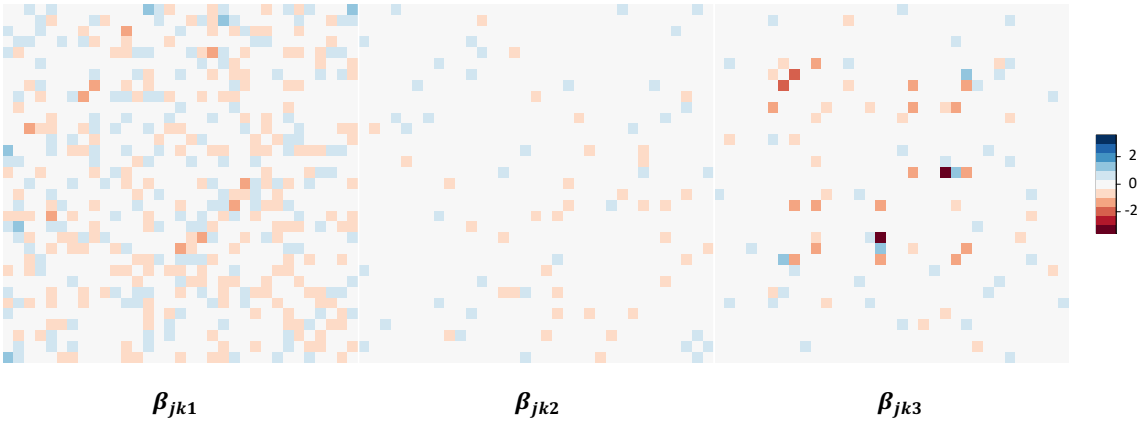
1. Generate an Erdös-Rényi graph G with connecting probability 5%.

2. Set the diagonal entries of $\boldsymbol{\Omega}$ to 1. For each edge $\{j, k\}$ in G, draw corresponding off-diagonal entrie ω_{jk} uniformly in $[-1, -0.5] \cup [0.5, 1]$.

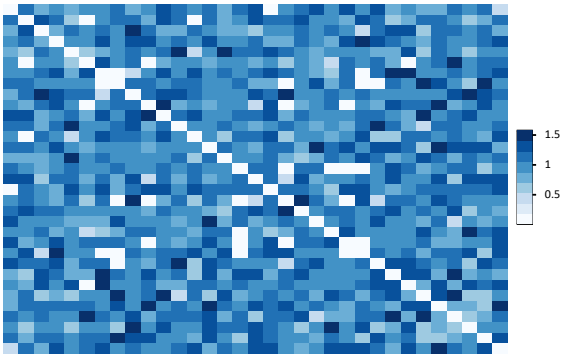
3. Since $\boldsymbol{\Omega}$ might not be positive definite, we kept adding $0.1\boldsymbol{I}$ to $\boldsymbol{\Omega}$ until $\boldsymbol{\Omega}$ became positive definite. The resulting partial correlations were less than 0.4 in magnitude. Then we simulated $\boldsymbol{y}_i \overset{iid}{\sim} N(0, \boldsymbol{\Omega}^{-1})$ and $x_{ij} \overset{iid}{\sim} U(-1, 1)$. GGMx took the independently generated x_{ij} as covariates, which were pure "noises" for constructing the graph of $\boldsymbol{y}_i$.

Scenario IV. We extended Scenario III to multiple graphs with $C = 3$ groups. The sample size of each group was $n_1 = 50, n_2 = 50$, and $n_3 = 51$. Graph G_1 was generated as an Erdös-Rényi graph with connecting probability 10%, which led to 63 edges. We randomly turned 3 edges on and 3 edges off from G_1 to obtain G_2 and similarly constructed G_3 from G_2. As a result, each pair of (G_1, G_2) and (G_2, G_3) shares about 90% edges whereas (G_1, G_3) shares about 80% edges. Then given graphs, the precision matrices and observations $\boldsymbol{y}_i$ were generated in the same way as Scenario III. To apply GGMx in this setting, we let x_{ij} be a binary indicator such that $x_{ij} = 1$ if observation i belongs to group j for $j = 1, 2$ and $x_{ij} = 0$ for $j = 1, 2$ if observation i belongs to group 3.

Scenario V. We have considered continuous covariates (Scenarios I-II), a discrete covariate (Scenarios IV), or no relevant covariates (Scenario III). Here, we included a scenario with one continuous covariate and one discrete covariate. We generated the data by following



(a)



(b)

Figure 4: Simulation truths for Scenario II. Heatmaps of (a) true β_{jkl}'s and (b) true t_{jk}. They are one posterior draw from GGMx applied to the multiple myeloma data.

Scenario I with one covariate replaced by a $Bernoulli(0.5)$ variable and the corresponding coefficients $\beta_{jk\ell}$'s set to ± 0.5 with equal probability.

Scenario VI. We considered a scenario without assuming ω_{ijj} to be a constant; instead we set $g_{jj}(\boldsymbol{x}_i) = 0.1 + 0.2x_{i1} + 0.2x_{i2}$ and $\omega_{ijj} = \exp\{g_{jj}(x_i)\}$. For off-diagonal elements, we randomly included 2% of the edges and the corresponding $\beta_{jk\ell}$ for $j < k$ was set to be 0.7. The covariate $x_{i\ell}$ was generated from $x_{i\ell} \overset{iid}{\sim} 2Beta(2, 1)$. The resulting precision matrix $\boldsymbol{\Omega}_i$ might not be positive definite for all observations $i = 1, \ldots, n$. We repeated the process until $\boldsymbol{\Omega}_i > 0, \forall i$. Then the observation $\boldsymbol{y}_i$ was drawn from normal $\boldsymbol{y}_i \overset{ind}{\sim} N(0, \boldsymbol{\Omega}_i^{-1})$.

Scenario VII. To illustrate GGMx can be used to recover time-varying GGM, we reduced the number of covariate to $q = 1$ from Scenario VI.

6.2 Methods under Consideration

We compared the proposed GGMx with six competing methods: Bayesian Gaussian graphical models (Mohammadi et al., 2015), graphical lasso (Friedman et al., 2008), kernel graphical lasso (Liu et al., 2010a), fused graphical lasso, group graphical lasso (Danaher et al., 2014), and Bayesian multiple Gaussian graphical model (Shaddox et al., 2018).

Bayesian Gaussian graphical models (BGGMs) assume i.i.d. multivariate Gaussian likelihood and the G-Wishart prior on the precision $\boldsymbol{\Omega} \sim W_G(b, \boldsymbol{D})$ and a uniform prior on the graph G. G-Wishart prior is conjugate to the multivariate Gaussian likelihood. However, due to intractable prior normalizing constant of G-Wishart prior, non-trivial MCMC algorithm is required for posterior inference. We use an efficient trans-dimensional MCMC algorithm proposed by Mohammadi et al. (2015) based on a continuous-time birth-death process.

Graphical lasso (glasso) is a penalized likelihood approach that maximizes the objective function $\log|\boldsymbol{\Omega}| - \text{tr}(\boldsymbol{S}\boldsymbol{\Omega}) - \lambda||\boldsymbol{\Omega}||_1$ where $\boldsymbol{S}$ is the sample covariance matrix. The first two terms are the Gaussian log-likelihood and the last term is an ℓ_1 penalty, which induces sparsity in $\boldsymbol{\Omega}$. The optimization is solved using a coordinate descent algorithm.

Both BGGM and glasso assume i.i.d. sampling and are designed to infer networks that do not change with covariates. For a more fair comparison, we implemented the kernel graphical lasso (k-glasso) approach outlined in Liu et al. (2010a). K-glasso is a modification of glasso with the sample covariance matrix $\boldsymbol{S}$ replaced by a covariate-dependent covariance matrix via kernel smoothing. Specifically, let

$$\boldsymbol{S}(\boldsymbol{x}) = \sum_{i=1}^{n} K\left(\frac{||\boldsymbol{x}-\boldsymbol{x}_i||}{h}\right)(\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}))(\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{x}))^T \Big/ \sum_{i=1}^{n} K\left(\frac{||\boldsymbol{x}-\boldsymbol{x}_i||}{h}\right),$$

with

$$\boldsymbol{\mu}(\boldsymbol{x}) = \sum_{i=1}^{n} K\left(\frac{||\boldsymbol{x}-\boldsymbol{x}_i||}{h}\right)\boldsymbol{y}_i \Big/ \sum_{i=1}^{n} K\left(\frac{||\boldsymbol{x}-\boldsymbol{x}_i||}{h}\right),$$

where $||\cdot||$ is the Euclidean norm, $h > 0$ is the bandwidth, and $K(\cdot)$ is a Gaussian kernel. Then a sparse estimate of $\boldsymbol{\Omega}_i$ is obtained by applying glasso with $\boldsymbol{S} = \boldsymbol{S}(\boldsymbol{x}_i)$, $\widehat{\boldsymbol{\Omega}}_i = \arg\min_{\boldsymbol{\Omega}}\{\log|\boldsymbol{\Omega}| - \text{tr}(\boldsymbol{S}(\boldsymbol{x}_i)\boldsymbol{\Omega}) - \lambda_i||\boldsymbol{\Omega}||_1\}$.

As pointed out in Section 3, the proposed GGMx is a multiple graphical model when the covariates are categorical. Multiple graphical models assume that observations are divided into C groups. The goal is to jointly estimate group-specific sparse precision matrices $\boldsymbol{\Omega}^{(c)}$, $c = 1, \ldots, C$. Since the grouping of observations can be represented by a categorical variable, GGMx is able to learn group-specific graphs. For comparison, we consider three alternative multiple graphical model approaches, the two penalized approaches proposed in Danaher et al. (2014), fused graphical lasso (FGL) and group graphical lasso (GGL), and the Bayesian multiple Gaussian graphical model (MGGM) proposed by Shaddox et al. 2018. Both penalized algorithms maximize the following objective with respect to positive definite matrices $\{\boldsymbol{\Omega}^{(c)}\}_{c=1}^{C}$,

$$\sum_{c=1}^{C} n_c\{\log|\boldsymbol{\Omega}^{(c)}| - \text{tr}(\boldsymbol{S}^{(c)}\boldsymbol{\Omega}^{(c)})\} - P(\{\boldsymbol{\Omega}^{(c)}\}_{c=1}^{C}),$$

where n_c is the sample size of group c, $\boldsymbol{S}^{(c)}$ is the sample covariance matrix of group c, and $P(\cdot)$ is a penalty that encourages sparsity and similarity of $\{\boldsymbol{\Omega}^{(c)}\}_{c=1}^{C}$. The penalty is chosen to be $\lambda_1 \sum_{c=1}^{C} \sum_{j \neq k} |\omega_{jk}^{(c)}| + \lambda_2 \sum_{c<c'} \sum_{j,k} |\omega_{jk}^{(c)} - \omega_{jk}^{(c')}|$ for FGL and $\lambda_1 \sum_{c=1}^{C} \sum_{j \neq k} |\omega_{jk}^{(c)}| + \lambda_2 \sum_{j \neq k} \sqrt{\sum_{c=1}^{C} \omega_{jk}^{(c)2}}$ for GGL.

Finally, MGGM uses local priors on sparse precision matrices (Wang, 2015) and can be thought as the local prior counterpart of the proposed method for the multiple graphs setting; comparisons with this method only pertain to Scenario IV.

For GGMx, we set the hyperparameters, $a_\tau = b_\tau = 10^{-1}$, $\mu_t = 1$, and $\sigma_t = 0.2$; these choices will be tested in sensitivity analyses at the end of this section. Both GGMx and BGGM were run for 10,000 iterations with 5,000 burn-in. The regularization parameter of glasso was selected by the stability approach (Liu et al., 2010b) implemented in the R package `huge`. The tuning parameters λ_1 and λ_2 of FGL and GGL were selected based on the approximated Akaike Information Criterion (AIC) as suggested by Danaher et al. (2014). A 20×20 grid evenly spaced between 0.05 and 0.5 for λ_1, and between 0.001 and 0.01 for λ_2, was used. Likewise, the tuning parameters λ_i and h_i of k-glasso were also selected based on AIC on a 20×20 grid $[0.1, 1] \times [0.1, 1]$ for each observation $i = 1, \ldots, n$. All results were based on 50 repeat simulations.

6.3 Simulation Results

To assess the graph recovery performance, we computed true positive rate (TPR), false discovery rate (FDR), and Matthews correlation coefficient (MCC),

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}},$$

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

where TP, FP, TN, and FN stand for true positives, false positives, true negatives, and false negatives. MCC takes value between -1 and 1 with 1 being perfect graph recovery and 0 being random guess. In addition, we scrutinized the edges with inclusion probability that is considerably affected by the covariates' value. Hence, we introduced another three measures: partial TPR (pTPR), partial FDR (pFDR), and partial MCC (pMCC) which are simply TPR, FDR, and MCC restricted to the edges with true frequency of inclusion across observations between 0.1 and 0.9. We report all the metrics in Figure 5. Overall, GGMx had robust, superior performance (with high true positive and low false discovery rates) across all scenarios.

In Scenario I, GGMx clearly outperformed BGGM and glasso in all six measures. This was expected because the data were generated from the proposed model and all edges were associated with covariates. BGGM and glasso assume i.i.d. sampling and therefore did not perform well. Although k-glasso was much better than BGGM and glasso, it is clear that GGMx performed significantly better than k-glasso in all metrics. In addition, GGMx can interpolate graph structure given new covariates. The results of graph interpolation (not shown) were very similar to those of graph estimation.

In Scenario II, it appeared that BGGM was comparable to GGMx in TPR, FDR and MCC. This is because in the simulation truth, there were much more nearly constant edges

than highly varying edges. In many cases including the application, it is interesting to focus on highly varying edges as they are most differential across observations. Not surprisingly, GGMx had favorable performance compared to BGGM and glasso in terms of pTPR, pFDR, and pMCC. K-glasso was not able to pick up the signals in this scenario, which mimicked the real data.

In Scenario III where there was no relationship between graph and covariates, BGGM outperformed GGMx, glasso, and k-glasso. But GGMx still had a reasonably good performance with the lowest FDR and substantially better overall performance than glasso and k-glasso.

In Scenario IV (multiple graphical models), GGL, FGL, and MGGM had higher TPR compared to GGMx, however, at the price of higher FDR. Consequently, GGMx outperformed GGL, FGL, and MGGM in terms of FDR and the overall measure MCC.

In Scenario V, GGL and FGL were applied ignoring the continuous covariate whereas k-glasso was applied ignoring the discrete covariate. GGMx was able to simultaneously incorporate both continuous and discrete covariates in estimating graphs and therefore as expected it had the best performance compared to BGGM, glasso, k-glasso, GGL, and FGL in practically all measures.

In Scenario VI where the diagonal elements of $\mathbf{\Omega}_i$ were not constrained to be constants, the results were consistent with those in Scenarios I-V. BGGM and glasso outperformed k-glasso overall but k-glasso was much better with respect to the selection of the edges that have substantial variability (measured by pTPR, pFDR, and pMCC). The proposed GGMx was clearly the best in both overall measures and partial measures. For example, GGMx had considerably higher MCC as well as pMCC than all the competing methods. In addition, we also evaluated the estimation accuracy of $\mathbf{\Omega}_i$ by computing the mean squared error (MSE). We again focused on edges with true frequency of inclusion across observations between 0.1 and 0.9. The resulting MSE was 0.10, 1.24, 0.44, and 0.82 for GGMx, BGGM, glasso, and k-glasso, which demonstrated the capability of the proposed GGMx in capturing the heterogeneity in $\mathbf{\Omega}_i$.

In Scenario VII, the main conclusion stays the same as in Scenario VI although k-glasso had significantly reduced FDR, however, at the price of significantly reduced TPR. GGMx, on the other hand, demonstrated its stable performance across all scenarios and all measures.

Lastly, we assessed the sensitivity of GGMx to the choice of all the hyperparameters (a_τ, b_τ) and (μ_t, σ_t). We picked Scenario VII and varied the hyperparameters in the following range, $(a_\tau, b_\tau) \in \{(10^{-2}, 10^{-2}), (10^{-3}, 10^{-3}), (10^{-4}, 10^{-4})\}$ and $(\mu_t, \sigma_t) \in \{(1.0, 0.5), (1.0, 1.0), (1.5, 1.0)\}$[3]. The performance of GGMx with different hyperparameters is reported in Table 2, which shows GGMx is robust within the considered range.

7. Application in Multiple Myeloma

We present an application of GGMx in modeling transcriptomic regulation in multiple myeloma (MM) which is a late-stage malignancy of plasma cells. Recent research has shifted the focus from traditional "one size fits all" therapies to precision medicine strate-

3. The resulting prior means (variances) of the minimum effect size t_{jk} are 1.0 (0.22), 1.3 (0.63), and 1.6 (0.77).

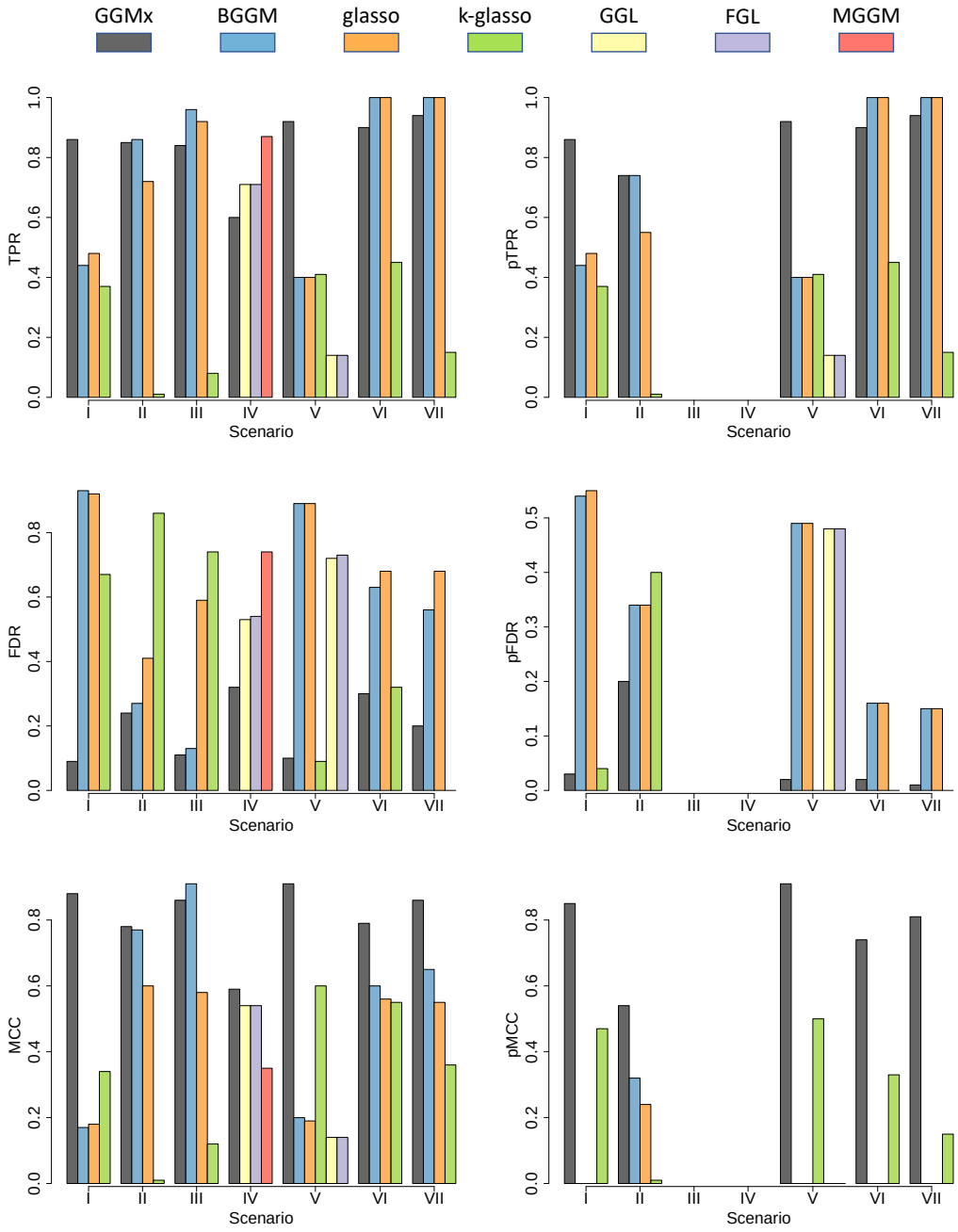


Figure 5: Simulations. Operating characteristics averaged over 50 repeat simulations under seven scenarios. Graph interpolation is not shown as it is similar to graph estimation. pMCC is 0 when no missing edge is detected. pTPR, pFDR, and pMCC are not available for Scenarios III and IV.

gies because MM is a highly heterogeneous genetic disease at an individual level (Hervé et al., 2011). To find better personalized treatment and more accurate prescriptive recom-

Table 2: Sensitivity Analysis. Operating characteristics for simulations under six alternative hyperparameter settings. The numbers are calculated on the basis of 50 repetitions; standard deviations are within parentheses. The first row shows the performance of GGMx in Scenario VII with default hyperparameter setting $(a_\tau, b_\tau) = (10^{-1}, 10^{-1})$ and $(\mu_t, \sigma_t) = (1, 0.2)$.

		TPR	FDR	MCC	pTPR	pFDR	pMCC
Default Parameter Setting		0.94 (0.04)	0.20 (0.12)	0.86 (0.07)	0.94 (0.04)	0.01 (0.01)	0.81 (0.09)
(a_τ, b_τ)							
	$(10^{-2}, 10^{-2})$	0.93 (0.04)	0.15 (0.10)	0.89 (0.06)	0.93 (0.04)	0.01 (0.01)	0.81 (0.09)
	$(10^{-3}, 10^{-3})$	0.93 (0.04)	0.10 (0.08)	0.91 (0.05)	0.93 (0.04)	0.01 (0.01)	0.80 (0.08)
	$(10^{-4}, 10^{-4})$	0.93 (0.04)	0.07 (0.07)	0.93 (0.04)	0.93 (0.04)	0.01 (0.01)	0.80 (0.08)
(μ_t, σ_t)							
	$(1.0, 0.5)$	0.97 (0.03)	0.33 (0.12)	0.80 (0.08)	0.97 (0.03)	0.02 (0.02)	0.82 (0.07)
	$(1.0, 1.0)$	0.98 (0.02)	0.30 (0.13)	0.82 (0.08)	0.98 (0.02)	0.03 (0.02)	0.81 (0.08)
	$(1.5, 1.0)$	0.97 (0.03)	0.19 (0.11)	0.88 (0.07)	0.97 (0.03)	0.03 (0.02)	0.81 (0.08)

mendations to MM patients, there needs to be a better understanding of the heterogeneity based on genomically defined pathways (Lohr et al., 2014). We use data generated by the Multiple Myeloma Research Consortium, a multi-institutional collaborative research effort collected data (among others) on gene expressions and clinical parameters from MM patients (Chapman et al., 2011).

We focus our analyses on the genes mapped to one of the most important pathways in MM, NF-κB signaling pathway. Activation of the NF-κB pathway has been implicated in MM, but the genomic foundation of such activation is only partially understood (Demchenko et al., 2010; Roy et al., 2018). Clinical information includes measurements of two important prognostic factors, serum beta-2 microglobulin (Sβ_2M) and serum albumin. The International Staging System (Greipp et al. 2005) uses these two prognostic factors to stage MM: stage I, Sβ_2M < 3.5 mg/L and serum albumin ≥ 3.5 g/dL; stage II, neither stage I nor III; and stage III, Sβ_2M ≥ 5.5 mg/L. The observed values of Sβ_2M and serum albumin, and the staging partition are depicted in Figure 6. We use these two prognostic factors as covariates ($q = 2$).

The goal of this study was to infer subject-level gene expression networks whose structures are modified by the prognostic factors. After removing outliers and samples with missing gene expression or clinical information, we had $n = 151$ samples and $p = 33$ genes. We ran two separate MCMCs, each with 50,000 iterations, discarded the first 50% as burn-in and saved every 50th sample after burn-in. To check MCMC convergence, we calculated the potential scale reduction factor (PSRF, Gelman et al. 1992) for each entry in $\mathbf{\Omega}_i$, $i = 1, \ldots, n$. The median PSRF was 1.00 with interquartile range 0.01, which showed no lack of convergence. We then concatenated the two chains and all subsequent inference was based on the combined Monte Carlo samples. The probability cutoff c was chosen to control the posterior expected FDR at 1%.

Population-level inference The estimated graphs had 30 edges per subject on average with minimum 20 edges (from a stage III patient) and maximum 37 edges (from a stage I patient). We summarized a population-level gene expression network $G = (V, E)$ as the

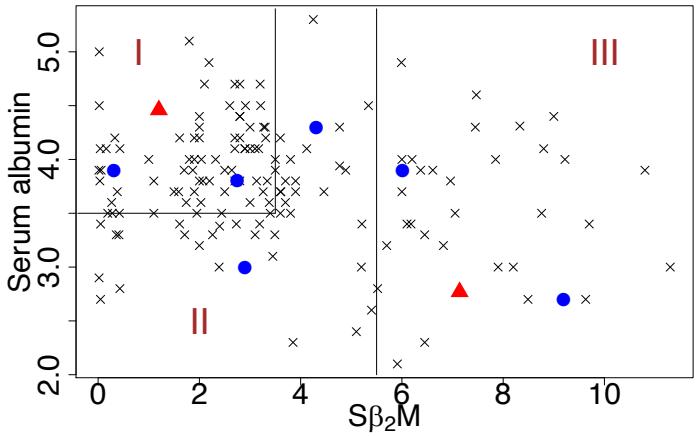


Figure 6: Observed prognostic factors are shown as crosses and dots. Dots are chosen as representative cases for network visualization in Figure 8. Triangles will be used to interpolate networks for unseen patients shown in Figure 9. The prognostic covariates space are partitioned into Stages I, II, and III, according to the International Staging System for multiple myeloma.

union of all networks across subjects $E = \cup_{i=1}^{n} E_i$. There were $|E| = 42$ edges in G. To visualize the graph variability, we computed the variance of edge inclusion. Specifically, let $\boldsymbol{e}_{jk} = (e_{1jk}, \ldots, e_{njk})$ be a binary vector such that $e_{ijk} = 1$ if $\{j, k\} \in E_i$. Then for edge $\{j, k\}$, the variance of edge inclusion was defined as the sample variance of $\boldsymbol{e}_{jk}$. The population-level network was reported in Figure 7, with the edge width proportional to edge inclusion variability. We found 14 out of 42 edges with variance greater than 0.2 (note the maximum variance is 0.25 for Bernoulli random variable). These 14 edges appeared in about 30%-70% of the patients. In line with our simulation studies, traditional GGMs are unlikely to accurately capture these differential edges.

Subject-level inference Next, we focus on the subject-level inference. We chose 6 representative patients, 2 from each stage, to show their respective networks in Figure 8. The values of their prognostics factors are represented by the dots in Figure 6. We set the edge width proportional to the absolute value of partial correlation $\rho_{ijk} = -\frac{\omega_{ijk}}{\sqrt{\omega_{ijj}\omega_{ikk}}}$, and use solid lines to represent positive partial correlations and dashed lines negative partial correlations.

We highlight several interesting biological findings. RELB was found to be a highly connected gene across all patients (Figures 7 and 8). RELB is a core member of NF-κB family. Hence it is not surprising that RELB played an import role in NF-κB pathway. In fact, many MM patients have abnormal NF-κB target gene expression, associated with genetic aberration of NFKB1 and NFKB2 (Annunziata et al., 2007). This further confirms our finding that RELB was consistently positively associated with NFKB1 and NFKB2. In addition, NFKBIA is an inhibitor of NF-κB, which is consistent with our findings that NFKBIA was negatively associated with RELB across patients. It is also known that genes in the same family tend to be positively associated with each other. Our study found positive links, for example, BIRC2—BIRC3 and NFKBIA—NFKBIZ. As disease progresses, some

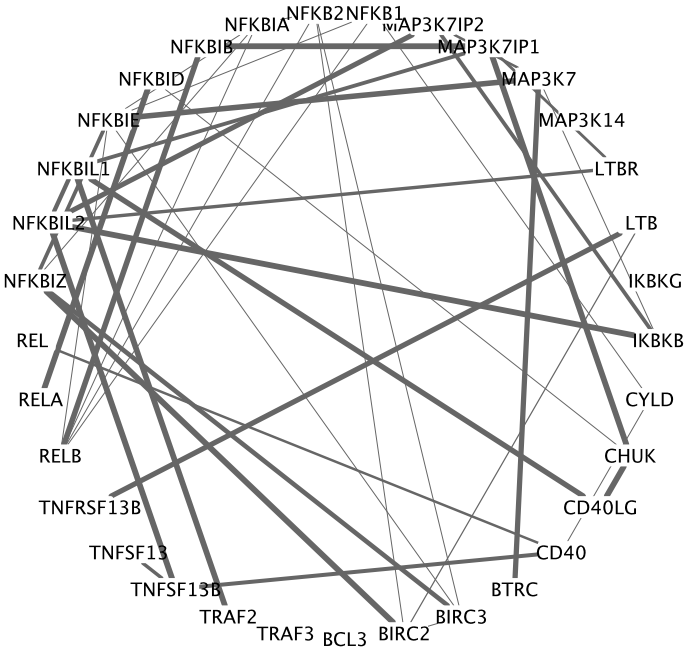


Figure 7: Population-level summary of gene expression network. The network is a union of all networks across subjects. The edge width is proportional to edge inclusion variability.

paths get blocked and some new connections get acquired. Among others, the link between LTB and TNFRSF13B was found in stage III patients but not in stage I patients whereas the link between NFKBIL2 and MAP3K7IP2 was lost in stage III patients. While some of those links are well documented in the biological literature (Liu et al., 2017), their gain and loss mechanisms need further validation and investigation.

Finally, as new patients come into the clinic, GGMx can be used to quickly predict the individualized gene network only based on the blood test results of Sβ_2M and serum albumin without the costly and time-consuming whole genome sequencing. For illustration, we picked two sets of covariates that were unobserved in our collected data; they are represented by triangles in Figure 6. The estimated gene expression network of the two hypothetical patients are shown in Figure 9, which was enabled by the unique feature of graph interpolation of the proposed GGMx.

8. Discussion

In this article, we introduce a general regression framework for (undirected) Gaussian graphical models with covariates (GGMx). This generalization of regular GGM beyond i.i.d. data allows the graph structure and strength to change with covariates and is particularly challenging especially in the undirected graph context due to the positive definiteness constraint of a precision matrix. We have addressed this challenge through a novel prior that is theoretically connected to non-local priors for precision matrices, paired with a carefully designed MCMC algorithm for efficient posterior inference. GGMx includes at least five special cases including standard GGMs, group-specific GGMs, time-varying GGMs, covariate-dependent

GGMs, and context-specific GGMs. We demonstrated the utility and robustness of GGMx through extensive simulations and an application in precision oncology. Our GGMx framework is broadly applicable to many other scientific domains of interest. For example, in brain functional magnetic resonance imaging data, GGMx can be used to study how brain connectivity networks change with covariates such as time and stimuli.

We remark that *covariance regression* (Hoff and Niu, 2012; Fox and Dunson, 2015) is a closely related model. It is, however, fundamentally different from the proposed GGMx in at least two ways. First, covariance regression assumes the covariance matrix rather than the precision matrix to be a function of $\boldsymbol{x}_i$ which takes a specific form, $\boldsymbol{\Sigma}_i = \boldsymbol{\Omega}_i^{-1} = \boldsymbol{\Psi} + \boldsymbol{\Lambda}(\boldsymbol{x}_i)\boldsymbol{\Lambda}(\boldsymbol{x}_i)^T$ for some PDM $\boldsymbol{\Psi}$ and matrix-value function $\boldsymbol{\Lambda}(\boldsymbol{x}_i)$. Second, covariance regression assumes a dense $\boldsymbol{\Sigma}_i$ whereas GGMx allows $\boldsymbol{\Omega}_i$ to be sparse and moreover, the sparsity pattern can change with covariates. Note that zeros in $\boldsymbol{\Lambda}(\boldsymbol{x}_i)$ generally do not translate to zeros in $\boldsymbol{\Sigma}_i$ or $\boldsymbol{\Omega}_i$. Therefore, it is not straightforward to extend the covariance regression framework to allow sparsity.

While our work is a useful first step for undirected graphical regression, there are several extensions and refinements possible. We have chosen the smooth covariate-dependent functions, $g_{jk}(\cdot)$, to be linear for simplicity and parsimony. Same choice has been made by similar papers (Cheng et al., 2014). However, in general, it can be replaced by a nonlinear function. For example, letting $\widetilde{\boldsymbol{x}}$ denote some basis expansion of $\boldsymbol{x}$ such as splines and wavelets, we can model $g_{jk}(\boldsymbol{x}) = \boldsymbol{\beta}_{jk}^T\widetilde{\boldsymbol{x}}$ and the same inference procedure with linear functions applies. We plan to incorporate nonlinearity in our future work. Furthermore, we have worked with a moderate number of variables due to several reasons. First, the number of parameters that need to be estimated in GGMx is on the order of $\frac{p(p+1)(q+2)}{2} + \frac{p(p-1)}{2}$, which can be large even for a moderate number of variables and covariates. Second, from an application perspective, we focus on a specific signalling pathway in multiple myeloma, NF-κB for deeper scientific interpretations. The small sample size (relative to the number of parameters) does not allow for reliable inferences for a much larger number of variables (e.g., the entire transcriptomic profile). Finally, the scalability of the proposed GGMx also limits the number of variables under consideration. The scalability can be potentially improved by adopting more efficient MCMC algorithms such as Metropolis-adjusted Langevin algorithm (Roberts et al., 1996) or Hamiltonian Monte Carlo (Duane et al., 1987). Both algorithms take advantage of gradient information of the target distribution. However, the hard thresholding function in (3) is discontinuous. This difficulty can be potentially overcome by considering a continuous relaxation of the hard thresholding function (Cai et al., 2018). Another potential solution is resorting to variational Bayes algorithms, which approximate the posterior distributions by simpler variational distributions through minimizing the Kullback–Leibler divergence between them. We hope to address the scalability issue in our future work.

Acknowledgement

YN was partially supported by NSF DMS-2112943. VB was partially supported by NIH grants R01CA244845-01A1 and P30 CA-046592 and start-up funds from the U-M Rogel Cancer Center and School of Public Health.

References

Davide Altomare, Guido Consonni, and Luca La Rocca. Objective Bayesian search of Gaussian directed ccyclic graphical models for ordered variables with non-local priors. *Biometrics*, 69(2):478–487, 2013.

Christina M Annunziata, R Eric Davis, Yulia Demchenko, William Bellamy, Ana Gabrea, Fenghuang Zhan, Georg Lenz, Ichiro Hanamura, George Wright, Wenming Xiao, et al. Frequent engagement of the classical and alternative NF-κB pathways by diverse genetic abnormalities in multiple myeloma. *Cancer Cell*, 12(2):115–130, 2007.

Anindya Bhadra and Bani K Mallick. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457, 2013.

Qingpo Cai, Jian Kang, Tianwei Yu, et al. Bayesian network marker selection via the thresholded graph Laplacian Gaussian prior. *Bayesian Analysis*, 2018.

Michael A Chapman, Michael S Lawrence, Jonathan J Keats, Kristian Cibulskis, Carrie Sougnez, Anna C Schinzel, Christina L Harview, Jean-Philippe Brunet, Gregory J Ahmann, Mazhar Adli, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 471(7339):467–472, 2011.

Jie Cheng, Elizaveta Levina, Pei Wang, and Ji Zhu. A sparse Ising model with covariates. *Biometrics*, 2014.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc B*, 76(2):373–397, 2014.

Yulia N Demchenko, Oleg K Glebov, Adriana Zingone, Jonathan J Keats, P Leif Bergsagel, and W Michael Kuehl. Classical and/or alternative NF-κB pathway activation in multiple myeloma. *Blood*, 115(17):3541–3552, 2010.

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

Adrian Dobra, Alex Lenkoski, and Abel Rodriguez. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Emily B Fox and David B Dunson. Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research*, 16(1):2501–2542, 2015.

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

Lingrui Gan, Naveen N Narisetty, and Feng Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114 (527):1218–1231, 2019.

Andrew Gelman, Donald B Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Peter J. Green and Alun Thomas. Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, 100(1):91, 2013.

Philip R Greipp, Jesus San Miguel, Brian GM Durie, et al. International staging system for multiple myeloma. *Journal of Clinical Oncology*, 23(15):3412–3420, 2005.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, page asq060, 2011.

Avet-Loiseau Hervé, Magrangeas Florence, Moreau Philippe, Attal Michel, Facon Thierry, Anderson Kenneth, Harousseau Jean-Luc, Munshi Nikhil, and Minvielle Stéphane. Molecular heterogeneity of multiple myeloma: pathogenesis, prognosis, and therapeutic implications. *Journal of Clinical Oncology*, 29(14):1893–1897, 2011.

Peter D Hoff and Xiaoyue Niu. A covariance regression model. *Statistica Sinica*, pages 729–753, 2012.

Valen E Johnson and David Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

Valen E Johnson and David Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

Kshitij Khare, Bala Rajaratnam, and Abhishek Saha. Bayesian inference for Gaussian graphical models beyond decomposable graphs. *Journal of the Royal Statistical Society: Series B*, 80(4):727–747, 2018.

Mladen Kolar, Ankur P Parikh, and Eric P Xing. On sparse nonparametric conditional covariance selection. In *ICML-10*, pages 559–566, 2010.

Han Liu, Xi Chen, Larry Wasserman, and John D Lafferty. Graph-valued regression. In *Advances in Neural Information Processing Systems*, pages 1423–1431, 2010a.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, pages 1432–1440, 2010b.

Ting Liu, Lingyun Zhang, Donghyun Joo, and Shao-Cong Sun. NF-κB signaling in inflammation. *Signal Transduction and Targeted Therapy*, 2:17023, 2017.

Jens G Lohr, Petar Stojanov, Scott L Carter, et al. Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer Cell*, 25(1):91–101, 2014.

Hélène Massam. Bayesian inference in graphical Gaussian models. *Handbook of Graphical Models*, pages 257–282, 2018.

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

Abdolreza Mohammadi, Ernst C Wit, et al. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

Yang Ni, Peter Müller, Yitan Zhu, and Yuan Ji. Heterogeneous reciprocal graphical models. *Biometrics*, 74(2):606–615, 2018.

Yang Ni, Francesco C Stingo, and Veerabhadran Baladandayuthapani. Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197, 2019.

Henrik Nyman, Johan Pensar, and Jukka Corander. Stratified gaussian graphical models. *Communications in Statistics-Theory and Methods*, 46(11):5556–5578, 2017.

Chris J Oates, Jim Korkola, Joe W Gray, Sach Mukherjee, et al. Joint estimation of multiple related biological networks. *The Annals of Applied Statistics*, 8(3):1892–1919, 2014.

Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110 (509):159–174, 2015.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

David Rossell and Donatello Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.

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

Alberto Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

Payel Roy, Uday Aditya Sarkar, and Soumen Basak. The NF-κB activating pathways in multiple myeloma. *Biomedicines*, 6(2):59, 2018.

James G Scott and Carlos M Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808, 2008.

Elin Shaddox, Francesco C Stingo, Christine B Peterson, Sean Jacobson, Charmion Cruickshank-Quinn, Katerina Kechris, Russell Bowler, and Marina Vannucci. A Bayesian approach for learning gene networks underlying disease severity in COPD. *Statistics in Biosciences*, 10(1):59–85, 2018.

Minsuk Shin, Anirban Bhattacharya, and Valen E Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica*, 28(2):1053, 2018.

Erik B Sudderth, Martin J Wainwright, and Alan S Willsky. Embedded trees: Estimation of gaussian processes on graphs with cycles. *IEEE Transactions on Signal Processing*, 52 (11):3136–3150, 2004.

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

Hao Wang et al. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

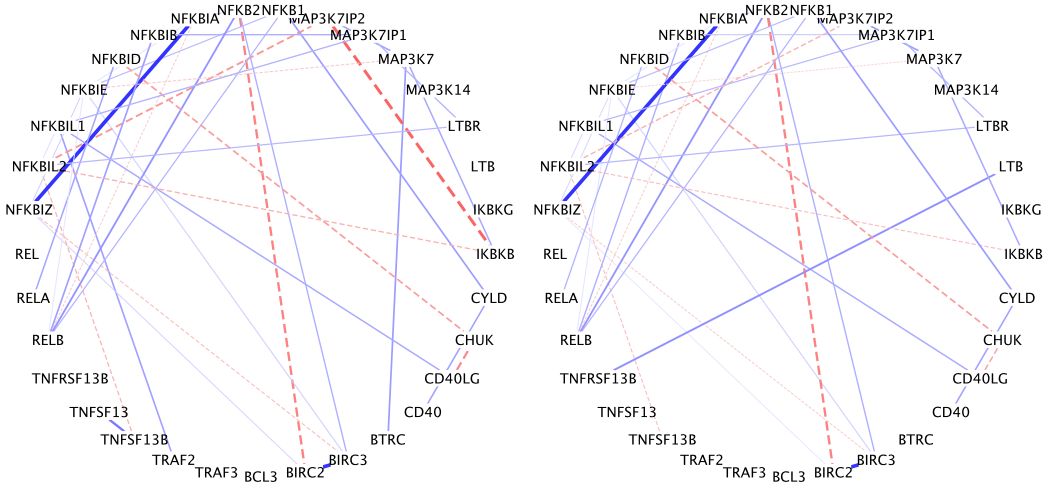
Yuying Xie, Yufeng Liu, and William Valdar. Joint estimation of multiple dependent gaussian graphical models with applications to mouse genomics. *Biometrika*, 103(3):493–511, 2016.

Masanao Yajima, Donatello Telesca, Yuan Ji, and Peter Müller. Detecting differential patterns of interaction in molecular pathways. *Biostatistics*, 16(2):240–251, 2015.

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

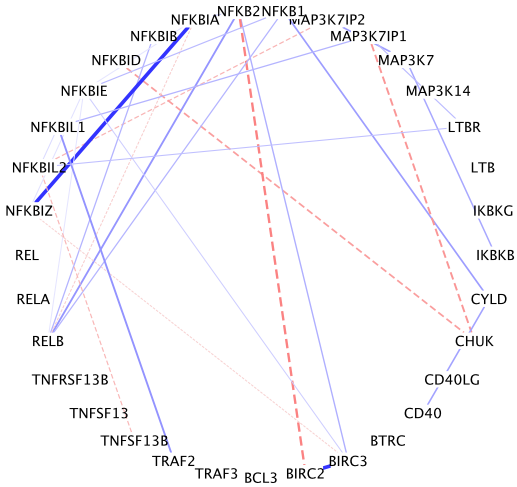
Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Shuheng Zhou, John Lafferty, and Larry Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

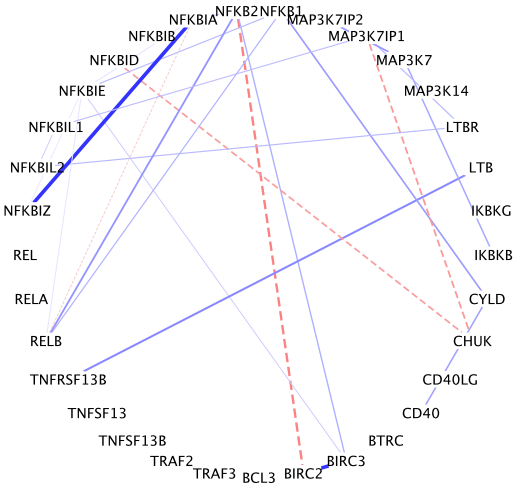


(a) Stage I

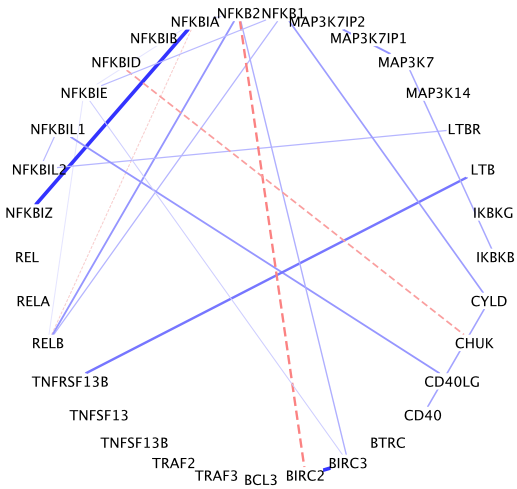
(b) Stage I



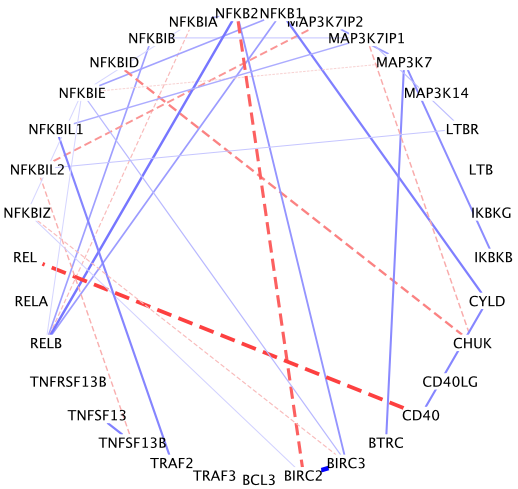
(c) Stage II



(d) Stage II



(e) Stage III



(f) Stage III

Figure 8: Subject-level networks for six representative patients, represented as dots in Figure 6. The edge width is proportional to the absolute value of partial correlation. The sign of partial correlation is represented by line type: + solid line and - dashed line.

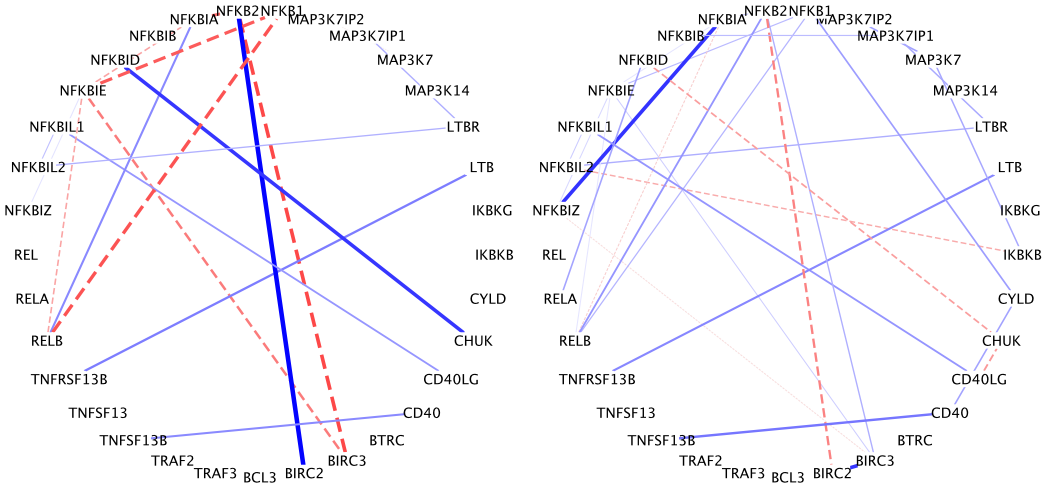


Figure 9: Network interpolation for two sets of unseen prognostic factors, represented as triangles in Figure 6.