Check for updates

Received: 8 November 2021 | Revised: 13 February 2022 | Accepted: 10 March 2022

DOI: 10.1002/wics.1582

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS

WILEY

Joint Gaussian graphical model estimation: A survey

Katherine Tsai[1] | **Oluwasanmi Koyejo**[2] | **Mladen Kolar**[3]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Chicago, Illinois, USA

[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Chicago, Illinois, USA

[3]The University of Chicago Booth School of Business, Chicago, Illinois, USA

Correspondence
Mladen Kolar, The University of Chicago Booth School of Business, Chicago, IL, USA
Email: mladen.kolar@chicagobooth.edu

Funding information
O. Koyejo acknowledges partial funding from a C3.ai Digital Transformation Institute Award, a Jump Arches Award, and an Strategic Research Initiatives award from the University of Illinois at Urbana-Champaign. K. Tsai acknowledges funding from National Science Foundation Graduate Research Fellowships Program. Other authors have no relevant financial or nonfinancial interests to disclose. This work was also funded in part by the following grants: NSF III 2046795 and IIS 1909577, along with computational resources donated by Microsoft Azure.

Edited by: James E. Gentle, Commissioning Editor and Co-Editor-in-Chief and David W. Scott, Review Editor and Co-Editor-in-Chief

Abstract

Graphs representing complex systems often share a partial underlying structure across domains while retaining individual features. Thus, identifying common structures can shed light on the underlying signal, for instance, when applied to scientific discovery or clinical diagnoses. Furthermore, growing evidence shows that the shared structure across domains boosts the estimation power of graphs, particularly for high-dimensional data. However, building a joint estimator to extract the common structure may be more complicated than it seems, most often due to data heterogeneity across sources. This manuscript surveys recent work on statistical inference of joint Gaussian graphical models, identifying model structures that fit various data generation processes.

This article is categorized under:

Data: Types and Structure > Graph and Network Data

Statistical Models > Graphical Models

KEYWORDS

Gaussian graphical model, graphical lasso, high-dimensional estimation, joint network, sparsity

1 | INTRODUCTION

Graphical models are powerful tools for expressing statistical relationships between variables. Examples of practical uses are ubiquitous and include models that characterize the causal relationships between the neurological activity of brain regions, genetic expression across genes, and a variety of other physiological measurements. A variety of applications have illustrated the value of graphical models for analyzing scientific phenomena (Chan et al., 2017; Dondelinger et al., 2013; Felsenstein, 1981; Friedman et al., 2000; Schäfer & Strimmer, 2005). Specifically, graphical models have proven useful for elucidating the mechanisms of brain function (Colclough et al., 2018; Foti & Fox, 2019; Greenewald et al., 2017; Manning et al., 2018; Qiu et al., 2016; Schwab et al., 2018; Skripnikov & Michailidis, 2019). This manuscript outlines joint graphical

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *WIREs Computational Statistics* published by Wiley Periodicals LLC.

models, an extension to standard graphical models that are useful for jointly analyzing data from multiple sources, for example, neurological data measured at multiple timescales, or joint neurological, genetic and phenotypic data. Specifically, this manuscript lays out the representation of joint graphical models (see Figure 1) and some of their properties, then outlines the best practices for estimating joint graphical models. This manuscript provides examples of data generation processes where the joint approach can significantly improve estimates compared with separate estimation.

A graph $G=(V,E)$ consists of a set of p nodes, also known as vertices $V=\{1, \ldots, p\}$ and a set of edges $E \subseteq V \times V$. In a probabilistic graphical model (Lauritzen, 1996), the set of nodes V is associated with coordinates of a random vector $\mathbf{x}=(x_1,\ldots,x_p)^\top$ and the edge set E captures dependency relationships between the components of the vector. In particular, in an undirected probabilistic graphical model, the absence of an edge between nodes a and b indicates that x_a and x_b are conditionally independent given all other variables $\mathbf{x}_{-\{a,b\}}=\{x_c | c \in V \backslash \{a,b\}\}$. In the case when a is a subset of nodes $\{1, \ldots, p\}$ rather than a single node, we will denote $\mathbf{x}_a \in \mathbb{R}^{|a|}$ as the vector whose entries correspond to x_i for $i \in a \subseteq \{1, \ldots, p\}$. In a neuroscience application, the random vector $\mathbf{x}$ could represent, for example, measurements of brain activity in different regions—so the set of edges corresponds to functional brain connectivity. Given n measurements of the vector $\mathbf{x}$, inferring the graph structure corresponds to identifying pairs of coordinates that are conditionally independent given all other variables (Drton & Maathuis, 2017). Inferring the graph structure based on conditional associations is more challenging than inferring the correlation structure between the measurements. However, the conditional independence graphs are generally considered more scientifically meaningful (Dobra et al., 2004).

2 | BACKGROUND: GAUSSIAN GRAPHICAL MODELS

The most widely used examples of probabilistic graphical models are Gaussian graphical models, where $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Omega}^{-1})$ is assumed to be distributed as a multivariate Gaussian vector with the mean vector μ and the precision matrix $\mathbf{\Omega}$ whose

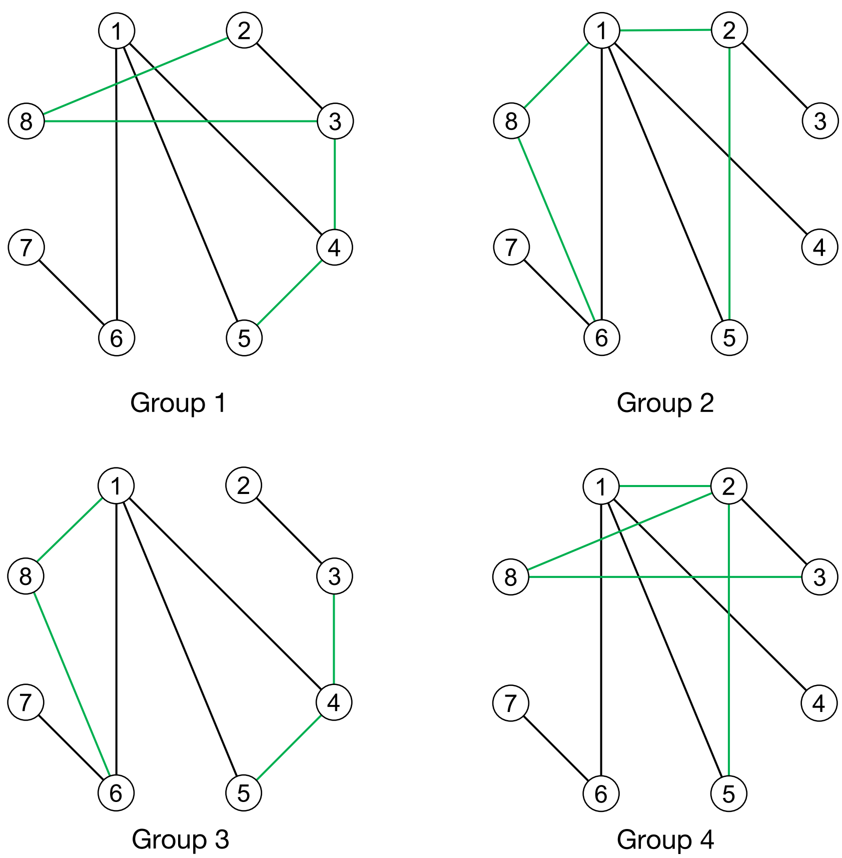

FIGURE 1 Joint graphical model estimation studies a group of graphs that have partially shared edge structures, presented in black, and individually owned edge structures, presented in green. Jointly estimating the shared structures enhances the estimation power while preserving individual structures as well

entries correspond to the partial correlation between the associated variables. In this setting, any two coordinates x_a and x_b are conditionally independent given $\mathbf{x}_{-\{a,b\}}$ if and only if the (a,b) entry of the precision matrix $\mathbf{\Omega}$ is zero (Lauritzen, 1996), and the graph structure can be inferred based on nonzero entries of $\mathbf{\Sigma}^{-1} := \mathbf{\Omega}$, also known as the inverse covariance matrix. Throughout the manuscript, we use the terms inverse covariance matrix and precision matrix interchangeably. In practice, the covariance matrix is not known and the graph structure needs to be estimated using samples drawn from an underlying distribution. For example, in a low-dimensional setting, we can first obtain an estimator of the precision matrix by maximizing the log-likelihood

$$\widehat{\mathbf{\Omega}} = \operatorname{argmax} \; n\left[\frac{1}{2}\log\{\det(\mathbf{\Omega})\} - \frac{1}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{\Omega}\right)\right], \tag{1}$$

where $\det(\cdot)$ is the determinant, $\operatorname{tr}(\cdot)$ is the trace, $\mathbf{x} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ is the empirical mean and $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\top}$ is the empirical covariance matrix. Next, the graph structure is estimated by thresholding small (in absolute value) elements of $\widehat{\mathbf{\Omega}}$ or testing whether they are zero (Drton & Maathuis, 2017; Drton & Perlman, 2004), that is, the graph structure corresponds to the nonzero entries of the resulting thresholded precision matrix.

In a high-dimensional setting, where the number of parameters to estimate, p, is much larger than the number of data points observed, n, maximizing the log-likelihood (1) results in poor quality estimates. In the particular case of Gaussian data, the resulting estimate, that is, the inverse of the covariance matrix, does not exist when $n < p$. Unfortunately, the high-dimensional setting is prevalent in various applications. For example, functional imaging of brain measurements using (standard) 2 mm^3 voxels will result in approximately $p = O(10^6)$ voxels with $n = O(10^2)$ measurements (Hsieh et al., 2013; Poldrack et al., 2011). There are two common problems that neuroscientists are interested in studying: (i) a static graph representing conditional independence between time series (Foti & Fox, 2019) and (ii) time-varying graphs within individuals (Calhoun et al., 2014; Lurie et al., 2020). In the first problem, we estimate a single graph by treating each time point as an i.i.d. sample (Varoquaux et al., 2010), effectively ignoring the temporal dependence. In the second problem, we estimate graphs for different time points or graphs within a time window. We will cover associated methodologies for both problems in Sections 3–4 and Section 6, respectively. We note that in addition to the small sample sizes, applications to fMRI are affected by temporal correlations in the observed data, which can reduce the effective sample size (Qiu et al., 2016). While p is large as compared with n, most entries in $\mathbf{\Omega}$, denoted as $\omega_{i,j}$, $i,j = 1, \ldots, p$, are zero, that is, the inverse covariance matrix is sparse. Hence, a typical strategy to estimate $\mathbf{\Omega}$ in a high-dimensional setting is to add a regularization function, such as the ℓ_1-norm of the parameters, to the log-likelihood function (1), which encourages the graph to be sparse or have other desirable structural biases (Bühlmann & van de Geer, 2011; Yuan & Lin, 2007). Specifically, we estimate $\mathbf{\Omega}$ using the following optimization program

$$\widehat{\mathbf{\Omega}} = \operatorname{argmax} \; n\left[\frac{1}{2}\log\{\det(\mathbf{\Omega})\} - \frac{1}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{\Omega}\right)\right] - \lambda_n \sum_{i \neq j} |\omega_{i,j}|. \tag{2}$$

Note that regularization is not added to the diagonal terms, $\omega_{i,i}$, $i = 1, \ldots, p$, because $\mathbf{\Omega}$ is positive definite and adding penalty on the diagonal entries would introduce additional bias. In this manuscript, we focus on the simultaneous estimation of multiple graphs that are structurally similar. We will illustrate in the following sections that exploiting the common structures will improve the estimation results on every individual graph.

2.1 | Joint Gaussian graphical models

We continue to use the example of brain measurements to demonstrate the idea of joint graphical model estimation. Consider the case of n fMRI scans collected from each of the K subjects. Suppose that we seek to estimate the functional connectivity (in this case, a graphical model) between the p voxels of each subject k, where p is much greater than n ($p \gg n$). Each functional connectivity network shares similarities with other networks, but are not identical. To better estimate the network of the subject k from n MRI scans, we may borrow information from other networks given that they are expected to share similar patterns. One simple approach is to construct a regularization function that

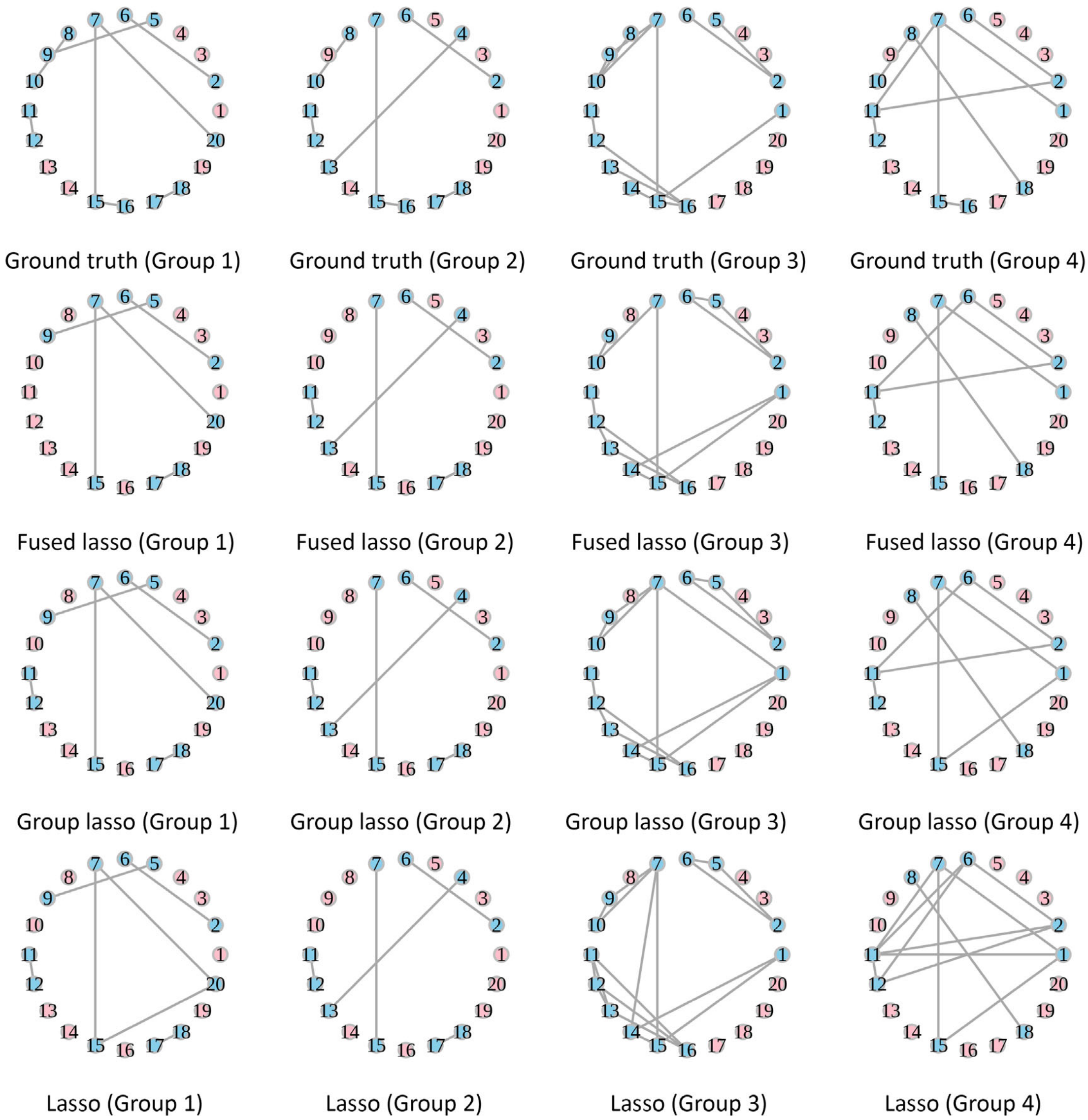

FIGURE 2 Comparison of joint and separate graph estimation procedures with $p = 20$ variables and sample size of $n = 5000$. Pink nodes do not have edges connecting to them, while blue nodes have at least on edge connecting to them. (Top row): Ground truth graphs. (Second row): Jointly estimated graphs using fused lasso. (Third row): Jointly estimated graphs using the group graphical lasso. (Fourth row): Graphs estimated separately using the graphical lasso. Details of all methods are provided in the text. From the figure, it is clear that joint estimation significantly outperforms separate estimation

encourages similarities between graphs, an approach known as joint estimation. Figure 2 shows a promising result when one pools the data across subjects, the resulting estimates better recover the ground truth graphs compared with separate estimations. It has also been shown that joint estimation can increase sensitivity and detect edges that are missing in separate estimation (Chiquet et al., 2011; Peterson et al., 2015). Thus, ignoring the information of other groups may lead to suboptimal solutions (Danaher et al., 2014; Lee & Liu, 2015). Moreover, joint estimation of graphical models has been applied successfully in a number of problems, including metabolite experiments (Tan et al., 2017), cancer networks (Hao et al., 2018; Lee & Liu, 2015; Mohan et al., 2012; Peterson et al., 2015; Saegusa & Shojaie, 2016), biomedical data (Kling et al., 2015; Pierson et al., 2015; Yajima et al., 2014), gene expression (Chun et al., 2015; Lin et al., 2017), text processing (Guo et al., 2011), climate data (Ma & Michailidis, 2016), and fMRI (Colclough et al., 2018; Lukemire et al., 2021; Qiu et al., 2016; Skripnikov & Michailidis, 2019). In all of these problems, data are heterogeneous, but the graphs share similarities.

To rigorously describe the example discussed above, we consider the problem of estimating graph structures $G^{(k)} = (V, E^{(k)})$, $k = 1, \ldots, K$, from K related groups of data. The data for each group are p-variate and share the same set

TABLE 1 Variants of the joint Gaussian graphical model

Category	Method	Model name	Model structure Ω	Penalty function/negative log prior/constraint
Coarse-grained	Penalized MLE	JGL (Guo et al., 2011)—Section 3.1	$\begin{cases} \omega_{i,j}^{(k)} = \theta_{i,j}\gamma_{i,j}^{(k)}, & i \neq j; \\ \omega_{i,i}^{(k)} = \gamma_{i,i}^{(k)}, & i = j. \end{cases}$	$\lambda_1 \sum_{i\neq j}\theta_{i,j} + \lambda_2 \sum_{k=1}^{K}\sum_{i\neq j} \lvert \gamma_{i,j}^{(k)} \rvert$
		JWLGL (Shan et al., 2020)—Section 3.1	$\omega_{i,j}^{(k)} = \theta_{m,m'}^{(k)}\gamma_{i,j}^{m,m\prime,(k)}.\theta_{m,m'}^{(k)} = \begin{cases} \alpha_{m,m'}\beta_{m,m'}^{(k)}, & m \neq m'; \\ 1, & \text{otherwise.} \end{cases}$ $\gamma_{i,j}^{m,m,(k)} = \begin{cases} \iota_{i,j}^{m,m}\rho_{i,j}^{m,m,(k)}, & i \neq j; \\ 1, & i = j. \end{cases}$	$\lambda_1 \sum_{m\neq m'}\alpha_{m,m'} + \lambda_2 \sum_{m \neq m'}\sum_{k=1}^{K}\sum_{i\neq j} \lvert \beta_{m,m'}^{(k)}\gamma_{i,j}^{m,m\prime,(k)} \rvert + \lambda_3 \sum_{m=1}^{M}\sum_{i\neq j}\iota_{i,j}^{m,m} + \lambda_4 \sum_{m=1}^{M}\sum_{i\neq j}\sum_{k=1}^{K} \lvert \rho_{i,j}^{m,m,(k)} \rvert$
		FGL (Danaher et al., 2014)—Section 3.2	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K}\sum_{i\neq j} \lvert \omega_{i,j}^{(k)} \rvert + \lambda_2 \sum_{k<k'}\sum_{i,j} \lvert \omega_{i,j}^{(k)} - \omega_{i,j}^{(k\prime)} \rvert$
		GGL (Danaher et al., 2014)—Section 3.2	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K}\sum_{i\neq j} \lvert \omega_{i,j}^{(k)} \rvert + \lambda_2 \sum_{i\neq j}\left\{\sum_{k=1}^{K}\left(\omega_{i,j}^{(k)}\right)^2\right\}^{\frac{1}{2}}$
		JAGL (Shan & Kim, 2018)—Section 3.2	$\omega_{i,j}^{(k)}$	$\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\neq j}\frac{1}{\left\lvert(1-\pi)\widehat{\iota_{i,j}}+\pi\widehat{s_{i,j}^{(k)}}\right\rvert^{r}} \lvert \omega_{i,j}^{(k)} \rvert$ [a]
		TFRE (Bilgrau et al., 2020)—Section 3.2	$\omega_{i,j}^{(k)}$	$\sum_{k=1}^{K}\frac{\lambda_k}{2}\left\lVert\boldsymbol{\Omega}^{(k)} - \mathbf{T}^{(k)}\right\rVert_F^2 + \sum_{k_1,k_2}^{K}\frac{\lambda_{k_1,k_2}}{4}\left\lVert\left(\boldsymbol{\Omega}^{(k_1)} - \mathbf{T}^{(k_1)}\right) - \left(\boldsymbol{\Omega}^{(k_2)} - \mathbf{T}^{(k_2)}\right)\right\rVert_F^2$
		SCAN (Hao et al., 2018)—Section 3.2	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K}\sum_{i\neq j} \lvert \omega_{i,j}^{(k)} \rvert + \lambda_2 \sum_{i\neq j}\left\{\sum_{k=1}^{K}\left(\omega_{i,j}^{(k)}\right)^2\right\}^{\frac{1}{2}} + \lambda_3 \sum_{k=1}\sum_{i=1}^{p} \lvert \mu_i^{(k)} \rvert$
		RCON (Mohan et al., 2012, Mohan et al., 2014)—Section 5.2	$\omega_{i,j}^{(k)}$	$G_q\left(\boldsymbol{\Omega}^{(1)} - \boldsymbol{\Omega}^{(2)}\right) = \min_{V:\boldsymbol{\Omega}^{(1)}-\boldsymbol{\Omega}^{(2)}=V+V^{\top}} f(V)\, f(V) = \sum_{j=1}^{p}\lVert V_j \rVert_q$
		GFGL (Gibberd & Nelson, 2017)—Section 6.1	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{t=1}^{T}\sum_{i\neq j} \lvert \omega_{i,j}^{(t)} \rvert + \lambda_2 \sum_{t=2}^{T}\left\lVert\boldsymbol{\Omega}_{-ii}^{(t)} - \boldsymbol{\Omega}_{-ii}^{(t-1)}\right\rVert_F$
	CLIME (Cai et al., 2011)	JEMP (Lee & Liu, 2015)—Section 3.1	$\omega_{i,j}^{(k)} = \theta_{i,j} + \gamma_{i,j}^{(k)}$	$\left\lvert\frac{1}{K}\sum_{k=1}^{K}\left\{\widehat{\boldsymbol{\Sigma}}^{(k)}\left(\boldsymbol{\Theta}+\boldsymbol{\Gamma}^{(k)}\right) - I\right\}\right\rvert_{\infty} \leq \lambda_1$ $\left\lvert\widehat{\boldsymbol{\Sigma}}^{(k)}\left(\boldsymbol{\Theta}+\boldsymbol{\Gamma}^{(k)}\right) - I\right\rvert_{\infty} \leq \lambda_2,\ \sum_{k=1}^{K}\boldsymbol{\Gamma}^{(k)} = 0$
		KSE (Qiu et al., 2016)—Section 6.2	$\omega_{i,j}^{(k)}$	$\left\lvert\widehat{S}(u_0)\boldsymbol{\Omega}(u_0) - I\right\rvert_{\infty} \leq \lambda_1$
	Bayesian approach	DSS-JGL (Li et al., 2019)—Section 3.3	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K}\sum_{i=1}^{p} \lvert \omega_{i,i}^{(k)} \rvert + \lambda_2 \sum_{k=1}^{K}\sum_{i\neq j}\frac{\lvert\omega_{i,j}^{(k)}\rvert}{v_{z_{i,j}}} + \lambda_3 \sum_{k<k'}\sum_{i\neq j} v_{(w_{i,j}z_{i,j})}^{-1} \lvert \omega_{i,j}^{(k)} - \omega_{i,j}^{(k\prime)} \rvert$ [b]
		BJEMGM (Gan et al., 2019)—Section 3.3	$\omega_{i,j}^{(k)}$	$\sum_{i=1}^{p}\sum_{k=1}^{K}\lambda_3\omega_{i,i}^{(k)} + \sum_{i<j}\log\left\{\prod_{k=1}^{K}\frac{\lambda}{2\lambda_1}\exp\left(-\lvert\omega_{i,j}^{(k)}\rvert/\lambda_1\right) + \prod_{k=1}^{K}\frac{1-\lambda}{2\lambda_2}\exp\left(-\lvert\omega_{i,j}^{(k)}\rvert/\lambda_2\right)\right\}$

(Continues)

TABLE 1 (Continued)

Category	Method	Model name	Model structure Ω	Penalty function/negative log prior/constraint
Fine-grained	Penalized MLE	LASICH (Saegusa & Shojaie, 2016)—Section 4.2	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \mid \omega_{i,j}^{(k)} \mid + \lambda_1 \lambda_2 \sum_{i \neq j} \left\{ \sum_{k,k'}^{K} W_{k,k'} \left(\omega_{i,j}^{(k)} + \omega_{i,j}^{(k')} \right)^2 \right\}^{\frac{1}{2}}$
		GEN-ISTA (Price et al., 2021)—Section 4.2	$\omega_{i,j}^{(k)}$	$\lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \mid \omega_{i,j}^{(k)} \mid + \lambda_2 \sum_{q=1}^{Q} \frac{1}{\lvert D_q \rvert} \sum_{k,k' \in D_q} \left\lVert \mathbf{\Omega}^{(k)} - \mathbf{\Omega}^{(k)\prime} \right\rVert_F^2$
	Neighborhood selection	JSEM (Ma & Michailidis, 2016)—Section 4.1	$\theta_{i,j}^{(k)} = -\omega_{i,j}^{(k)} / \omega_{i,i}^{(k)}$	$\sum_{j \neq i} \sum_{g \in \mathcal{G}_{i,j}} \lambda_{i,j}^{[g]} \left\lVert \theta_{i,j}^{[g]} \right\rVert_2$
	Bayesian approach	MRF (Peterson et al., 2015)—Section 4.3	$\omega_{i,j}^{(k)}$	$-\log \left[\prod_{k=1}^{K} \lvert \Omega^{(k)} \rvert^{(b-2)/2} \exp\left\{ -2^{-1} \mathrm{tr}\left(\Omega^{(k)} \mathbf{D} \right) \right\} \right]$

[a] n_k denotes the number of samples of group k, $\widehat{t}_{i,j}$ is the precision matrix estimated by pooling all samples across groups, $\widehat{s}_{i,j}$ is the precision matrix estimated by an individual group, and $r > 0$.
[b] $z_{i,j}$, $w_{i,j}$ are binary variables for $i \neq j$ drawn independently from a Bernoulli distribution and v_0, v_1 are two constants such that $v_1 > v_0 > 0$.

of nodes V, but the underlying connection patterns $E^{(k)}$ may be different due to the heterogeneity between groups. The data for the kth group can be represented as an $n_k \times p$ matrix $\mathbf{X}^{(k)} = \left(\mathbf{x}_1^{(k)}, ..., \mathbf{x}_{n_k}^{(k)}\right)^\top$, where the rows $\mathbf{x}_i^{(k)} = \left(x_{i,1}^{(k)}, ..., x_{i,p}^{(k)}\right)$, $i = 1, ..., n_k$, are p-dimensional vectors of observations. Assuming that the data in each group are distributed according to a p-variate Gaussian distribution, $\mathbf{x}_i^{(k)} \tilde{\mathcal{N}}\left(\mu^{(k)}, \left\{\mathbf{\Omega}^{(k)}\right\}^{-1}\right)$, $i = 1, ..., n_k$, where $\mu^{(k)} \in \mathbb{R}^p$ is the mean, which we assume without loss of generality to be $\mathbf{0}$, and $\mathbf{\Omega}^{(k)} \in \mathbb{R}^{p \times p}$ is the precision matrix. Given observations $\mathbf{X} = \left\{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(K)}\right\}$, we can estimate $\mathbf{\Omega} = \left\{\mathbf{\Omega}^{(1)}, ..., \mathbf{\Omega}^{(K)}\right\}$ by maximizing the penalized joint log-likelihood for K groups:

$$\widehat{\mathbf{\Omega}} = \operatorname{argmax}\ \ell(\mathbf{\Omega}) - \mathbf{P}(\mathbf{\Omega}), \quad \ell(\mathbf{\Omega}) := \sum_{k=1}^{K} n_k \left[\log\left\{\det\left(\mathbf{\Omega}^{(k)}\right)\right\} - \operatorname{tr}\left(\widehat{\mathbf{\Sigma}}^{(k)} \mathbf{\Omega}^{(k)}\right)\right], \tag{3}$$

where $\widehat{\mathbf{\Sigma}}^{(k)} = {n_k}^{-1}\left(X^{(k)}\right)^\top X^{(k)}$, $k = 1, ..., K$, are the sample covariance matrices. Directly solving (3) without the penalty $\mathbf{P}(\mathbf{\Omega})$ gives the maximum log-likelihood estimate of $\mathbf{\Omega}$. However, the solution is equivalent to solving the maximum log-likelihood estimate of each group individually and fails to utilize the shared “information” across different groups. We hence explore different approaches that use the penalty function $\mathbf{P}(\mathbf{\Omega})$ to incorporate the group structure and focus on the structural assumptions behind the penalties. The comparison of different methods introduced in the text can be found in Table 1. Specifically, we consider coarse-grained vs. fine-grained structural assumptions. For coarse-grained structures, all pairs of edge strengths are penalized/regularized in the same way, i.e., invariant to the group identity. In contrast, fine-grained structure uses regularization/priors between edge strengths that vary across groups, for example, using prespecified weights for pairs of groups.

The rest of the manuscript is organized as follows. In Section 3, we introduce methods that employ coarse-grained structural constraints. Methods that employ fine-grained structural constraints are discussed in Section 4. Sections 5 and 6 illustrate two practical examples. Section 5 covers differential graphs, which are special cases of the joint estimation paradigm with two groups. Joint estimation of time-series data is discussed in Section 6. Finally, we close the review with open problems in Section 7.

3 | JOINT GRAPHICAL MODELS USING COARSE-GRAINED STRUCTURE

We outline a variety of approaches for joint graphical model estimation that use prior knowledge of coarse-grained structures across groups. As noted, for coarse-grained structures, all pairs of edge strengths are penalized/regularized in the same way, that is, invariant to the group identity. The illustration of coarse-grained structure is shown in Figure 3. In contrast, fine-grained structure uses regularization/priors between edge strengths that vary across groups, for example, using prespecified weights for pairs of groups. For comparison, the illustration of fine-grained structure is shown in Figure 4. The performance of the coarse-grained estimation procedure is improved using regularization that captures the common structure across the K groups—enabling the use of shared information across groups. We will discuss two directions in detail: hierarchical regularizers and analogous Bayesian priors.

3.1 | Joint graphical models with hierarchical structure

Guo et al. (2011) studied joint estimation of related precision matrices, where the precision matrices are assumed to be related through a hierarchical structure. Specifically, each entry in the precision matrix is the multiplication of a common component across K groups and an individual component: $\omega_{i,j}^{(k)} = \theta_{i,j}\gamma_{i,j}^{(k)}$ for $i \neq j$ and $\omega_{i,i}^{(k)} = \gamma_{i,i}^{(k)}$, where $\theta_{i,j}$ is the shared component and $\gamma_{i,j}^{(k)}$ is the group-specific component. Thus, this approach enforces a common background structure. To encourage sparsity, an ℓ_1-norm penalty term is also included as a regularizer, resulting in the following objective termed Joint Graphical Lasso (JGL):

$$\widehat{\mathbf{\Theta}}, \left\{\widehat{\mathbf{\Gamma}}^{(k)}\right\}_{k=1}^{K} = \operatorname{argmax}\ \ell(\mathbf{\Omega}) - \lambda_1 \sum_{i \neq j} \theta_{i,j} - \lambda_2 \sum_{k=1}^{K} \sum_{i \neq j} \left|\gamma_{i,j}^{(k)}\right|,$$

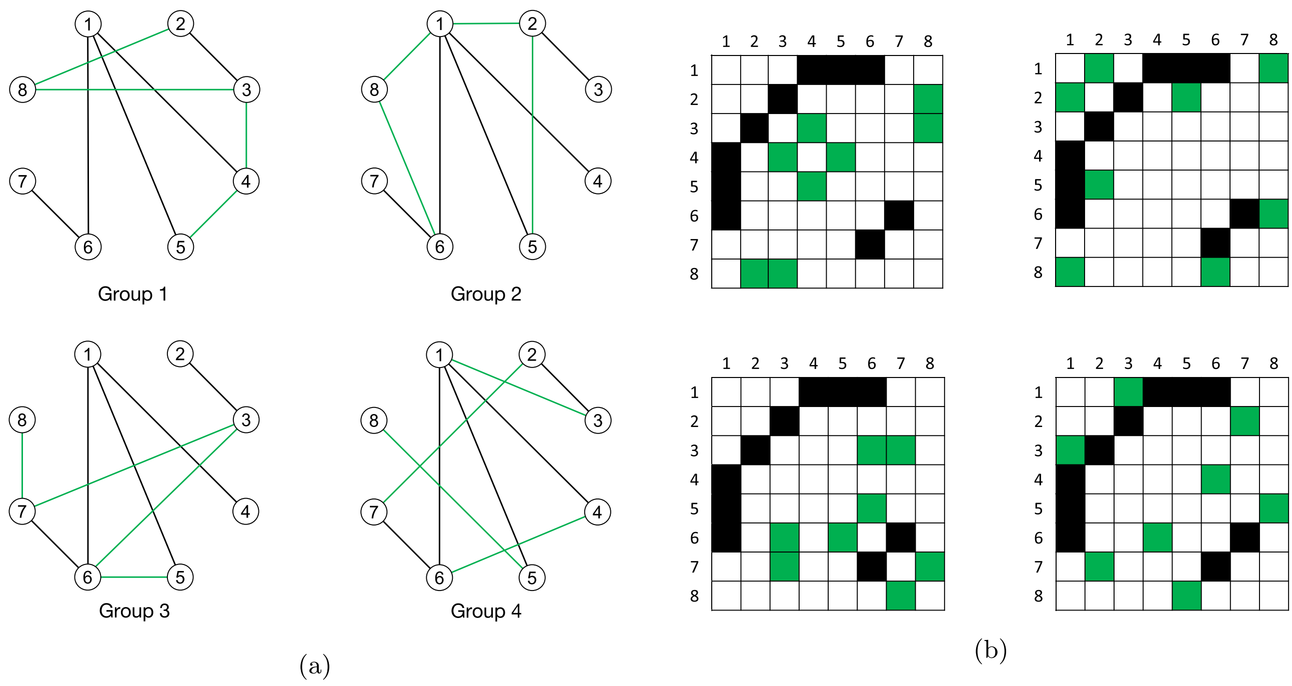

FIGURE 3 Graphical models with shared coarse-grained structure across groups. (a) The black lines denote the common edges, while green lines denote individual edges. Each graph has four unique individual edges that are not present in other graphs and four edges that are present in all graphs. (b) The corresponding adjacency matrices for each graph shown in (a)

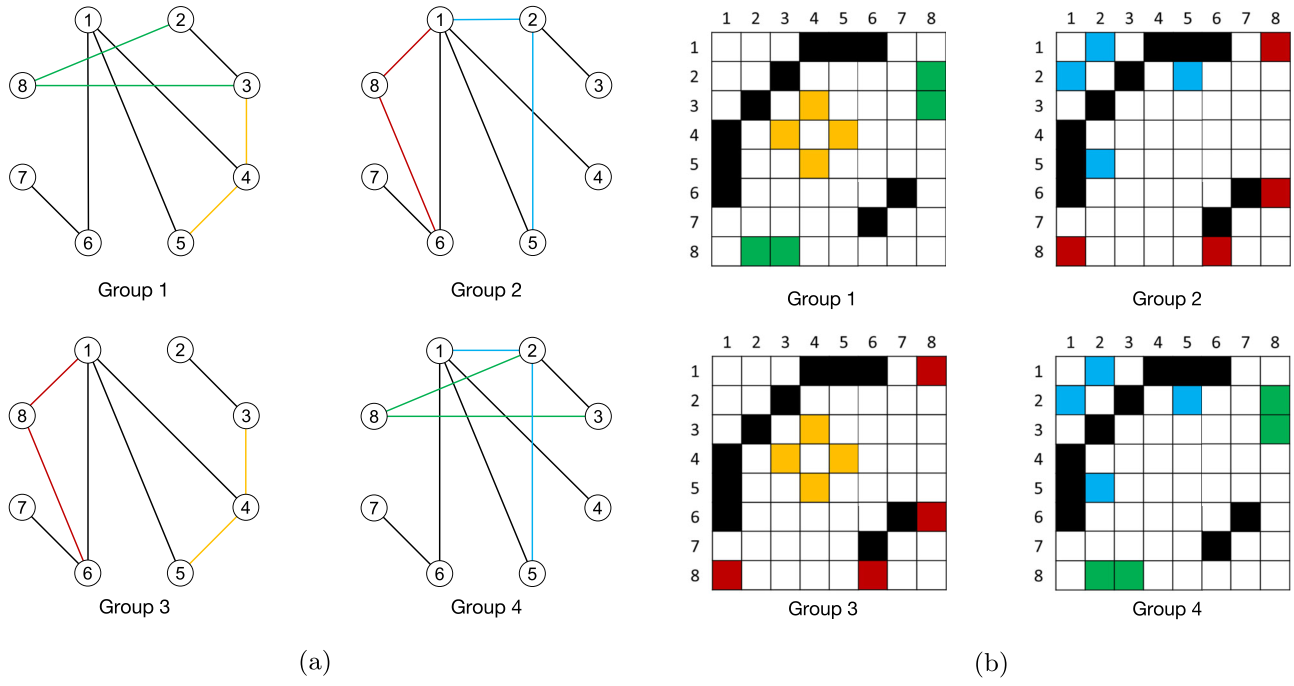

FIGURE 4 Graphical models with shared fine-grained structure across groups. (a) The black lines denote the common edges. Green colored lines represent the common structure of Group 1 and Group 4, yellow lines of Group 1 and Group 3, red lines of Group 2 and Group 3, and blue lines of Group 2 and Group 4. (b) The corresponding adjacency matrices for each graph shown in (a), using the same colors for groups of shared edges

where λ_1 and λ_2 are hyperparameters that control the scale of the penalty. Note that even when the common component $\theta_{i,j}$ is nonzero, an individual entry $\omega_{i,j}^{(k)}$ can still be set to zero by the ℓ_1 penalty, which denotes a missing edge in the associated graph. It is worth pointing out that this method is non-convex and hence only convergence to local minima is guaranteed. Danaher et al. (2014) introduced a similar method where the associated penalty functions are convex, which we discuss in Section 3.2.

Shan et al. (2020) proposed a Joint tWo-Level Graphical Lasso (JWLGL), which is a more expressive model that constructs two-level structures on both the set of common components and individual components. The algorithm further clusters the set of nodes V into M classes and imposes class specific structure: let m and m' be the classes to which nodes i and j belong, respectively. If $m \neq m'$, we have:

$$\omega_{i,j}^{(k)} = \theta_{m,m'}^{(k)} \gamma_{i,j}^{m,m\prime,(k)}; \quad \theta_{m,m'}^{(k)} = \alpha_{m,m'} \beta_{m,m'}^{(k)}.$$

If $m = m'$, we have $\theta_{m,m}^{(k)} = 1$ and

$$\omega_{i,j}^{(k)} = \theta_{m,m}^{(k)} \gamma_{i,j}^{m,m,(k)}; \quad \gamma_{i,j}^{m,m,(k)} = \begin{cases} \iota_{i,j}^{m,m} \rho_{i,j}^{m,m,(k)}, & i \neq j; \\ 1, & i = j. \end{cases}$$

Without loss of generality, we assume $\alpha_{m,m'} \geq 0$ and $\iota_{i,j}^{m,m} \geq 0$ for $i \neq j$ and $m \neq m'$. Here, $\alpha_{m,m'}$ and $\iota_{i,j}^{m,m}$ denote the common components shared across K groups, while $\beta_{m,m'}^{(k)}$ and $\rho_{i,j}^{m,m,(k)}$ denote the individual components that vary across groups.

Lee and Liu (2015) proposed a Joint Estimator of Multiple Precision matrices (JEMP) under an assumption that precision matrices decompose into the sum of two components: $\omega_{i,j}^{(k)} = \theta_{i,j} + \gamma_{i,j}^{(k)}$. In contrast to the maximum likelihood, the estimation procedure of JEMP is motivated by the CLIME estimator (Cai et al., 2011), which estimates a single precision matrix by solving the following optimization problem:

$$\widehat{\boldsymbol{\Omega}}^{(k)} = \operatorname{argmin} \left\| \boldsymbol{\Omega}^{(k)} \right\|_1 \quad \text{subject to} \quad \left| \widehat{\boldsymbol{\Sigma}}^{(k)} \boldsymbol{\Omega}^{(k)} - I \right|_\infty \leq \xi, \tag{4}$$

where ξ is a tuning parameter. The CLIME estimator finds a sparse $\widehat{\boldsymbol{\Omega}}^{(k)}$ while ensuring that $\widehat{\boldsymbol{\Sigma}}^{(k)} \widehat{\boldsymbol{\Omega}}^{(k)}$ is close to an identity matrix. JEMP can be seen as a generalization of CLIME to a multi-group setting as it solves the following optimization problem:

$$\begin{aligned} &\widehat{\boldsymbol{\Theta}}, \left\{ \widehat{\boldsymbol{\Gamma}}^{(k)} \right\}_{k=1}^{K} = \operatorname{argmin} \|\boldsymbol{\Theta}\|_1 + \nu \sum_{k=1}^{K} \left\| \boldsymbol{\Gamma}^{(k)} \right\|_1, \\ &\text{subject to} \left| \frac{1}{K} \sum_{k=1}^{K} \left\{ \widehat{\boldsymbol{\Sigma}}^{(k)} \left(\boldsymbol{\Theta} + \boldsymbol{\Gamma}^{(k)} \right) - I \right\} \right|_\infty \leq \lambda_1, \quad \left| \widehat{\boldsymbol{\Sigma}}^{(k)} \left(\boldsymbol{\Theta} + \boldsymbol{\Gamma}^{(k)} \right) - I \right|_\infty \leq \lambda_2, \quad \sum_{k=1}^{K} \boldsymbol{\Gamma}^{(k)} = 0, \end{aligned}$$

where $\boldsymbol{\Theta}$ denotes the common structure, i.e., the mean of the precision matrices $K^{-1} \sum_{k=1}^{K} \boldsymbol{\Omega}^{(k)}$, and $\boldsymbol{\Gamma}^{(k)}$ denotes the individual residual components $\boldsymbol{\Omega}^{(k)} - \boldsymbol{\Theta}$. In the above optimization problem, the first constraint regularizes the average difference and the second constraint regularizes the individual difference. Thus, the first constraint imposes a common structure across groups. The prespecified weight ν controls the degree of uniqueness of each group, while λ_1 and λ_2 are hyperparameters that measure group average and individual estimation quality, respectively.

3.2 | Regularization approaches for modeling joint structure

Regularization-based approaches (Bilgrau et al., 2020; Danaher et al., 2014; Hao et al., 2018; Shan & Kim, 2018) do not assume the form of the common structure and individual structure, but instead impose similarity constraints across groups. For example, the Fused Graphical Lasso (FGL) and the Group Graphical Lasso (GGL) (Danaher et al., 2014) add convex penalty terms to the log-likelihood function to learn a common structure:

$$\widehat{\boldsymbol{\Omega}} = \operatorname{argmax} \ell(\boldsymbol{\Omega}) - P(\boldsymbol{\Omega});$$

$$P_{\mathrm{FGL}}(\mathbf{\Omega}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{i,j}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\omega_{i,j}^{(k)} - \omega_{i,j}^{(k')}|; \tag{5}$$

$$P_{\mathrm{GGL}}(\mathbf{\Omega}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{i,j}^{(k)}| + \lambda_2 \sum_{i \neq j} \left\{ \sum_{k=1}^{K} \left(\omega_{i,j}^{(k)}\right)^2 \right\}^{\frac{1}{2}}. \tag{6}$$

The first penalty term in both P_{FGL} and P_{GGL} encourages model sparsity. The second term in P_{FGL} encourages groups to have shared edge values, while the P_{GGL} penalty tends to be less restrictive and only encourages a shared sparsity pattern. In addition, an R-package "JGL" is provided that implements both *FGL* and *GGL* (Danaher et al., 2014).

Hao et al. (2018) proposed simultaneous clustering and estimation (SCAN) procedure that addresses the case when the heterogeneous data are missing group labels, for example, when the groups are latent or unknown. SCAN partitions the unlabeled data into K clusters and simultaneously imposes a homogeneous structure across groups using the GGL penalty. Given n unlabeled observations x_i, $i = 1,\ldots,n$, with the density function $l(x,\mu,\mathbf{\Omega}) = \sum_{k=1}^{K} \pi_k l_k\left(x;\mu^{(k)},\mathbf{\Omega}^{(k)}\right)$, where π_k is the probability that x_i belongs to the kth group and $l_k\left(x;\mu^{(k)},\mathbf{\Omega}^{(k)}\right) = \log\left\{\det\left(\mathbf{\Omega}^{(k)}\right)\right\} - \mathrm{tr}\left\{\left(x-\mu^{(k)}\right)\left(x-\mu^{(k)}\right)^{\top}\mathbf{\Omega}^{(k)}\right\}$. SCAN solves the following optimization problem:

$$\{\hat{\pi}_k\}_{k=1}^{K}, \hat{\mu}, \hat{\mathbf{\Omega}} = \mathrm{argmax} \frac{1}{n} \sum_{i=1}^{n} \log(l(x_i,\mu,\mathbf{\Omega})) - P_{\mathrm{SCAN}}(\mu,\mathbf{\Omega});$$

$$P_{\mathrm{SCAN}}(\mu,\mathbf{\Omega}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{i,j}^{(k)}| + \lambda_2 \sum_{i \neq j} \left\{ \sum_{k=1}^{K} \left(\omega_{i,j}^{(k)}\right)^2 \right\}^{\frac{1}{2}} + \lambda_3 \sum_{k=1}^{K} \sum_{i=1}^{p} |\mu_i^{(k)}|.$$

Note that the first two terms of $P_{\mathrm{SCAN}}(\mu,\mathbf{\Omega})$ correspond to the GGL penalty function and the third term is the ℓ_1-norm penalty, used for encouraging the sparsity of the mean vectors. Such regularization is common in the context of high-dimensional data, where many variables contain limited information about the clustering structure. Hence, placing a sparse penalty function realizes selection of informative variables (Pan & Shen, 2007; Sun et al., 2012).

There are several other methods that use the regularization approach to estimate the joint structure. Shan and Kim (2018) proposed the Joint Adaptive Graphical Lasso (JAGL) procedure that introduces a weighted ℓ_1 penalty term to tackle problems with unbalanced data. Bilgrau et al. (2020) proposed Targeted Fused Ridge Estimator (TFRE) that uses an additional ℓ_2 penalty term that incorporates target matrices as prior information to stabilize the estimation process. In addition, an R-package "rags2ridges" provides an implementation of TFRE (Bilgrau et al., 2020).

3.3 | Bayesian methods for joint Gaussian graphical model estimation

We briefly overview Bayesian methods for joint Gaussian graphical model estimation. Bayesian formulations of graphical models use priors to encourage desired properties for model selection. For example, the spike-and-slab prior is commonly used in practice to encourage sparsity in precision matrices. In general, the probabilistic counterpart to the penalty function follows the relation $p(\mathbf{\Omega}) \propto \exp(-P(\mathbf{\Omega}))$, where $p(\mathbf{\Omega})$ is the prior distribution of the precision matrix and $P(\mathbf{\Omega})$ denotes the penalty function. By the Bayes' rule, the posterior distribution is proportional to the product of the likelihood and prior distribution. Therefore, finding a maximum a posteriori probability (MAP) estimate is equivalent to obtaining the estimate by maximizing the log-likelihood (MLE) with an additional penalty function. Several works (Li et al., 2019; Lin et al., 2017; Mitra et al., 2016; Peterson et al., 2015; Tan et al., 2017; Yajima et al., 2014) have addressed Bayesian graphical model estimation by designing priors that incorporate structural information. In the Gaussian graphical model case, a Wishart prior (Atay-Kayis & Massam, 2005; Lenkoski & Dobra, 2011; Mohammadi & Wit, 2015; Roverato, 2002) is often placed on the precision matrix. This prior is a conjugate prior for the Gaussian likelihood, that is, for a Gaussian likelihood, the posterior distribution remains Wishart. Furthermore, the normalizing constant of the posterior distribution has an explicit form when the graph

G is decomposable, that is, when the index set V of a graph can be partitioned into three disjoint nonempty sets $V = A \cup S \cup B$ and (i) S is a clique, (ii) S separates A and B, (iii) $A \cup S$ and $S \cup B$ form decomposable subgraphs.

Li et al. (2019) proposed the joint spike-and-slab graphical lasso prior, designed to encourage global sparse structure. In a related work, Tan et al. (2017) placed a multiplicative prior on the adjacency matrices, where the distribution of each edge depends on the multiplication of the values of two end nodes. This prior not only encourages sparsity, but also allows users to specify the degree of connections. We discuss local common structure methods (Lin et al., 2017; Peterson et al., 2015) in Section 4.3 and the differential graph methods (Mitra et al., 2016; Yajima et al., 2014) in Section 5.2. We will not go into details about the Bayesian formulation of graphical models, but instead give a high-level overview of various approaches.

Consider a single group setting with $\mathbf{\Omega} = \left\{\mathbf{\Omega}^{(1)}\right\}$. From a Bayesian perspective, the lasso regularizer can be viewed as a Laplace prior (Marlin & Murphy, 2009; Wang, 2012) and is formulated as:

$$p(\mathbf{\Omega} \mid \lambda) \propto \prod_{i \neq j} \frac{\lambda}{2} \exp\left(-\lambda |\omega_{i,j}|\right) \prod_{i=1}^{p} \left\{ \frac{\lambda}{2} \exp\left(-\frac{\lambda}{2}\omega_{i,i}\right) 1_{(\omega_{i,i} > 0)} \right\} 1(\mathbf{\Omega} \succ 0), \tag{7}$$

where $1(\mathbf{\Omega} \succ 0)$ restricts the precision matrix to be positive definite. The term $1_{(\omega_{i,i} > 0)}$ ensures that the diagonal entries are non-negative and hence preserves the positive definiteness of $\mathbf{\Omega}$. When taking the logarithm $\log p(\mathbf{\Omega}|\lambda)$, the first product is equal to the lasso regularizer. Therefore, when computing the MAP estimate, the logarithm of the Laplace prior along with the log-likelihood is the penalized MLE estimator with lasso penalty function. In the multiple group case, to promote the group similarity between the precision matrices, Li et al. (2019) converted the GGL and FGL penalties to structural priors.

In Bayesian inference, other than computing the MAP estimator, we are also interested in the posterior mean, mode, and samples. In this case, the shrinkage priors are not enough to produce sparse posterior samples (or mean and mode) because the posterior does not concentrate on sparse parameters. Therefore, additional thresholding is required to obtain sparsity. As an alternative, one may use the spike-and-slab prior (Mitchell & Beauchamp, 1988) to promote the sparsity pattern in the posterior. Consider a single group $\mathbf{\Omega} = \left\{\mathbf{\Omega}^{(1)}\right\}$, the spike-and-slab prior is a hierarchical mixture prior formulated as:

$$p(\mathbf{z}|\lambda) = \prod_{i \neq j} \mathrm{Ber}\left(z_{i,j}|\lambda\right); \tag{8}$$

$$p(\mathbf{\Omega}|\mathbf{z}) = \prod_{i \neq j} \left(1 - z_{i,j}\right)\delta\left(\omega_{i,j}\right) + z_{i,j}\mathcal{N}\left(\omega_{i,j}|0, \sigma^2\right),$$

where $\delta(\cdot)$ denotes the delta function. If $z_{i,j} = 0$, we restrict the variable to be zero. One may also replace the delta function with a normal distribution with small variance, which approximates the delta function. In the multiple group case, a set of latent indicators following the spike-and-slab distribution adaptively control the value of the FGL (resp., GGL) penalty (Li et al., 2019), namely the Doubly Spike-and-Slab Joint Graphical Lasso (DSS-JGL). Consider two constants $v_1 > v_0 > 0$ and $z_{i,j}$, $w_{i,j}$ are binary variables for $i \neq j$. We assume that each $z_{i,j}$ and $w_{i,j}$ are drawn independently from a Bernoulli distribution with a specific parameter. The DSS-JGL prior is represented as:

$$-\log p(\mathbf{\Omega} \mid \mathbf{z}, \mathbf{w}) \propto \lambda_1 \sum_{k=1}^{K} \sum_{i=1}^{p} |\omega_{i,i}^{(k)}| + \lambda_2 \sum_{k=1}^{K} \sum_{i \neq j} \frac{|\omega_{i,j}^{(k)}|}{v_{z_{i,j}}} + \lambda_3 \sum_{k < k'} \sum_{i \neq j} \frac{1}{v_{(w_{i,j} z_{i,j})}} |\omega_{i,j}^{(k)} - \omega_{i,j}^{(k')}|, \tag{9}$$

where the third term can also be replaced by the group lasso penalty function, similar to (6). We can choose v_0 to be small, so that when $z_{i,j} = 0$ for $i \neq j$, the second term in (9) will be large, forcing the posterior to be zero. Similar behavior also follows for the joint regularization term in (9) when either $z_{i,j}$ or $w_{i,j}$ is zero. Additionally, an R-package "SSJGL" provides an implementation of DSS-JGL (Li et al., 2019).

Although the Bayesian approaches introduced above provide expressive structures for joint estimation of multiple graphical models, theoretical guarantees that characterize convergence rates are lacking in general. Gan et al. (2019) provided guarantees on the structure recovery and the convergence rate in ℓ_∞ norm. Specifically, they proposed

Bayesian Joint Estimation of Multiple Graphical Models (BJEMGM) that extends the spike-and-slab prior to multiple graphs, but in a different setting compared with Li et al. (2019). Let $z_{i,j}$ be i.i.d. samples drawn from Ber(λ) with $0 \leq \lambda \leq 1$. The prior on $\boldsymbol{\omega}_{i,j} = \left\{\omega_{i,j}^{(1)},\ldots,\omega_{i,j}^{(K)}\right\}$ is defined as

$$-\log p\left(\boldsymbol{\omega}_{i,j} \mid z_{i,j},\lambda_1,\lambda_2\right) \propto -\log\left\{\prod_{k=1}^{K} z_{i,j}\frac{\lambda_1}{2}\exp\left(-\lambda_1|\omega_{i,j}^{(k)}|\right)+\prod_{k=1}^{K}\left(1-z_{i,j}\right)\frac{\lambda_2}{2}\exp\left(-\lambda_2|\omega_{i,j}^{(k)}|\right)\right\}. \tag{10}$$

The prior on the diagonal entries $\omega_{i,i}^{(k)}$, $i = 1, \ldots, p$, $k = 1, \ldots, K$, is the same as the second term of (7) with parameter λ replaced by λ_3. Marginalizing over $\mathbf{z}$, the log of the prior distribution is expressed as

$$-\log p(\boldsymbol{\Omega} \mid \lambda,\lambda_1,\lambda_2,\lambda_3) = \sum_{i=1}^{p}\sum_{k=1}^{K}\lambda_3\omega_{i,i}^{(k)} + \sum_{i<j}\log\left(\prod_{k=1}^{K}\frac{\lambda}{2\lambda_1}\exp\left(-|\omega_{i,j}^{(k)}|\backslash\lambda_1\right)+\prod_{k=1}^{K}\frac{1-\lambda}{2\lambda_2}\exp(-|\omega_{i,j}^{(k)}|\backslash\lambda_2)\right).$$

From the modeling perspective, the prior in (9) additionally enforces the similarity of inverse covariance values $\omega_{i,j}^{(k)}$ for $k = 1, \ldots, K$ and $i \neq j$, while (10) only constructed a shared latent Bernoulli variable $z_{i,j}$ across K groups that controls the sparsity of $\omega_{i,j}^{(k)}$ for $k = 1, \ldots, K$.

4 | JOINT GRAPHICAL MODELS USING FINE-GRAINED STRUCTURE

When estimating coarse-grained joint graphical models, all pairs of edge strengths are penalized equally; the relationship between $\omega_{i,j}^{(k)}$ and $\omega_{i,j}^{(k\prime)}$ and the relationship between $\omega_{i,j}^{(k)}$ and $\omega_{i,j}^{(k\prime\prime)}$, $k \neq k'$, $k' \neq k''$ are assumed to be equal. However, in many real world settings, some subsets of groups share a local structure that does not appear across all groups. Figure 4 visualizes the adjacency matrices related to edge sets E^k,$k = 1, \ldots, 4$, of graphical models that are not globally similar, but with subsets of groups that share a local structure. The most common approach in modeling such graphical models is to incorporate this prior knowledge of the relation between subgroups into the estimators (Ma & Michailidis, 2016; Saegusa & Shojaie, 2016). We briefly outline some approaches for joint graphical model estimation with fine-grained shared structure.

4.1 | Entry-wise structural information

Given the relation information $\mathcal{G} = \cup_{1 \leq i < j \leq p}\mathcal{G}_{i,j}$, where $\mathcal{G}_{i,j}$ is a set that encodes the group relations of node i and node j, Figure 4a illustrates an example of entry-wise structural information. Consider the pair of nodes $(i,j) = (3,4)$ in Figure 4a: both graphs of Group 1 and Group 2 have edges (yellow lines) connecting these two nodes, while Group 3 and Group 4 do not have an edge in between. Then the corresponding relation information $\mathcal{G}_{3,4} = \mathcal{G}_{4,3}$ is {{1,3},{2,4}}. Similarly, for the pair of nodes $(i,j) = (2,8)$, the graphs of Group 1 and Group 4 have an edge (green line) connecting (2,8), while graphs of Group 2 and Group 3 do not have edge between node (2,8). Then, the corresponding relation information $\mathcal{G}_{2,8} = \mathcal{G}_{8,2}$ is {{1,4},{2,3}}. The relation information of graphs in Figure 4 is $\mathcal{G}_{i,j} = \{\{1,3\},\{2,4\}\}$ for (i,j) ∈{(3,4),(4,3),(4,5),(5,4),(1,2),(2,1),(2,5),(5,2)}; $\mathcal{G}_{i,j} = \{\{1,4\},\{2,3\}\}$ for (i,j) ∈{(2,8),(8,2),(3,8),(8,3),(1,8),(8,1),(6,8),(8,6)}; and $\mathcal{G}_{i,j} = \{\{1,2,3,4\}\}$ for the rest of the node pairs.

Ma and Michailidis (2016) proposed a joint structural estimation method (JSEM) to estimate edge sets $\left\{\widehat{E}^{(k)}\right\}_{k=1}^{K}$ by modifying the neighborhood selection algorithm (Meinshausen & Bühlmann, 2006) to incorporate structural information. Once the edge sets are estimated, each precision matrix is estimated by maximizing the group specific likelihood constrained to have zeros indexed by $\left\{\widehat{E}^{(k)}\right\}_{k=1}^{K}$. We will briefly introduce the neighborhood selection method and then show how JSEM extends it to multiple graphical models.

The neighborhood selection algorithm estimates the conditional independence structure, which is encoded in the pattern of zeros of the precision matrix under a Gaussian model, by solving a collection of regression problems. See chapter 12 in Maathuis et al. (2018) and references therein. Suppose the p-dimensional random variable $\boldsymbol{x}$ follows a normal distribution $\mathcal{N}\left(\mu,(\boldsymbol{\Omega})^{-1}\right)$ with an associated graph $G = (V,E)$. In this case, we can express x_i, $i \in V$, as a linear function of other nodes:

$$x_i = \sum_{j \in V \setminus \{i\}} \theta_{ij} x_j + \varepsilon_j,$$

where ε_j is independent of x_j, $j \in V \setminus \{i\}$, if and only if $\theta_{ij} = -\omega_{i,j}/\omega_{i,i}$. Therefore, the optimal prediction of x_i given the remaining variables can be formulated as the following optimization problem:

$$\left\{\widehat{\theta}_{i,j}\right\}_{j \in V \setminus \{i\}} = \operatorname{argmin} \mathbb{E}\left(x_i - \sum_{j \in V \setminus \{i\}} \theta_{i,j} x_j\right)^2. \tag{11}$$

Since the distribution of $\boldsymbol{x}$ is not known, the expectation term in (11) cannot be computed. Let $\boldsymbol{X}$ be a $n \times p$ matrix that collects n i.i.d. observations of $\boldsymbol{x}$. The ℓ_1 penalized empirical optimization objective is given as:

$$\widehat{\boldsymbol{\Theta}}_i = \underset{\boldsymbol{\Theta}_i:\ \theta_{i,i}=0}{\operatorname{argmin}} \frac{1}{n} \|X_i - X\boldsymbol{\Theta}_i\|_2^2 + \lambda \|\boldsymbol{\Theta}_i\|_1, \tag{12}$$

where $\boldsymbol{\Theta}_i$ is a p-dimensional vector $\boldsymbol{\Theta}_i = \left(\theta_{i,1}, \ldots, \theta_{i,p}\right)^\top$, $\theta_{i,i} = 0$, and X_i is the i-th column of the matrix $\boldsymbol{X}$. To estimate multiple models, JSEM extends the neighborhood selection approach in (12) by solving the following optimization problem:

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}_i^{(k)}:\theta_{i,i}^{(k)}=0,\ k=1,\ldots,K}{\operatorname{argmin}} \sum_{k=1}^{K} \frac{1}{n_k} \left\| X_i^{(k)} - X^{(k)} \boldsymbol{\Theta}_i^{(k)} \right\|_2^2 + 2P_{\text{JSEM}}\left(\left\{\boldsymbol{\Theta}_i^{(k)}\right\}_{k=1}^{K}\right); \tag{13}$$

$$P_{\text{JSEM}}\left(\left\{\boldsymbol{\Theta}_i^{(k)}\right\}_{k=1}^{K}\right) = \sum_{j \neq i} \sum_{g \in \mathcal{G}_{i,j}} \lambda_{i,j}^{[g]} \left\|\theta_{i,j}^{[g]}\right\|_2,$$

where $\boldsymbol{\Theta} = \left\{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(K)}\right\}$, the penalty term incorporates the relation information $\mathcal{G}$, and $\lambda_{i,j}^{[g]}$ is the group-specific tuning parameter. The penalty function P_{JSEM} resembles the GGL penalty in (6), except that the norm is now placed on a subset of groups provided by the relation information $\mathcal{G}$. Minimizing (13) results in the following estimated edge sets

$$\widehat{E}^{(k)} = \left\{(i,j) : 1 \le i < j \le p, \widehat{\theta}_{i,j}^{(k)} \neq 0 \text{ or } \widehat{\theta}_{j,i}^{(k)} \neq 0\right\}, \quad k = 1, \ldots, K.$$

Given $\widehat{E}^{(k)}$, we define $\mathcal{S}^+_{\widehat{E}^{(k)}} = \left\{\boldsymbol{\Omega} : \boldsymbol{\Omega} \succ 0,\ \omega_{i,j} = 0,\ \forall (i,j) \notin \widehat{E}^{(k)} \text{ and } i \neq j\right\}$. The precision matrix set $\boldsymbol{\Omega}$ is finally estimated by maximizing the log-likelihood with constraints that enforce the sparsity pattern:

$$\widehat{\boldsymbol{\Omega}}^{(k)} = \underset{\boldsymbol{\Omega}^{(k)} \in \mathcal{S}^+_{\widehat{E}^{(k)}}}{\operatorname{argmax}} \log\left\{\det\left(\boldsymbol{\Omega}^{(k)}\right)\right\} - \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^{(k)} \boldsymbol{\Omega}^{(k)}\right), \quad k = 1, \ldots, K. \tag{14}$$

Note that we can apply JSEM only when element-wise structural relation information is given. However, when not all element-wise structural relation information is provided, one can still use the GGL penalty on subsets of groups for which prior information is available.

4.2 | Group-wise structural information

As obtaining entry-wise structural information is sometimes challenging, another approach is to use the relationship between groups, described by weights, in estimation. For example, suppose that $K = 3$ and we have the following relationships between groups: Group 1 and Group 2 share similarity, Group 1 and Group 3 also share similarity, but Group 2 and Group 3 are unlikely to be similar. To this end, Saegusa and Shojaie (2016) proposed LAplacian Shrinkage for Inverse Covariance matrices from Heterogeneous populations (LASICH), that uses a weighted graph $G_g = (\Gamma, E, W)$ to record the relations between groups. The node set Γ denotes K groups, the edge set E captures the relations between groups, and the weight set $W: E \to \mathbb{R}_+$ represents the similarities between groups. Then, a Laplacian penalty function is placed on the objective function to impose group similarity. The optimization problem is formulated as follows:

$$\widehat{\mathbf{\Omega}} = \operatorname{argmax}\ \ell(\mathbf{\Omega}) - P_{\text{LASICH}}(\mathbf{\Omega}),$$
$$\text{subject to}\ \mathbf{\Omega}^{(k)} = \left(\mathbf{\Omega}^{(k)}\right)^{\top},\ \mathbf{\Omega}^{(k)} \in \mathcal{S} \succ 0,\quad k = 1, \ldots, K;$$
$$P_{\text{LASICH}}(\mathbf{\Omega}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{i,j}^{(k)}| + \lambda_1 \lambda_2 \sum_{i \neq j} \left\{ \sum_{k,k\prime}^{K} W_{k,k\prime} \left(\omega_{i,j}^{(k)} + \omega_{i,j}^{(k\prime)}\right)^2 \right\}^{\frac{1}{2}},$$

where the first penalty term in $P_{\text{LASICH}}(\mathbf{\Omega})$ promotes the sparsity pattern and the second term encourages the similarities within subsets of groups. In the case where the weight set is not available, Saegusa and Shojaie (2016) further proposed a two-stage algorithm, called Hierarchical Clustering LAISCH (HC-LAISCH), that first uses hierarchical clustering to learn relations between groups and then applies LAISCH. Furthermore, under that Gaussian assumption, the estimates from HC-LAISCH and LAISCH share the same consistency properties. Although the approaches in Ma and Michailidis (2016), and Saegusa and Shojaie (2016) require prior information on the group relations, or the prior information is obtained by another algorithm, they provide a more flexible structure than the global structure methodologies, such as GGL (Danaher et al., 2014; Guo et al., 2011). In particular, the global structure approach can be viewed as a special case of a local structure approach with homogeneous structural information.

While the work introduced earlier required prior information about the group relations, which may not be available for most cases, Graphical Elastic Net Iterative Shrinkage Thresholding Algorithm (GEN-ISTA) (Price et al., 2021) jointly estimates graphs and group membership via k-means clustering. That is, GEN-ISTA further clusters groups $k = 1, \ldots, K$ into Q classes. Let D_q, $q = 1, \ldots, Q$, be the sets that contain group memberships. The objective function for GEN-ISTA is

$$\underset{\mathbf{\Omega}, D}{\operatorname{argmin}} \ell(\mathbf{\Omega}) - \mathbf{P}_{\text{GEN-ISTA}}(\mathbf{\Omega}, D); \tag{15}$$

$$\mathbf{P}_{\text{GEN-ISTA}}(\mathbf{\Omega}, D) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{i,j}^{(k)}| + \lambda_2 \sum_{q=1}^{Q} \frac{1}{|D_q|} \sum_{k,k' \in D_q} \left\| \mathbf{\Omega}^{(k)} - \mathbf{\Omega}^{(k)\prime} \right\|_F^2.$$

It is easy to see that when we fix $\mathbf{\Omega}$ and optimize $\boldsymbol{D}$, then (15) becomes a k-means clustering objective. In contrast, when we fix $\boldsymbol{D}$ and optimize $\mathbf{\Omega}$ alone, the problem (15) reduces to a joint graphical model with a variant of FGL (Danaher et al., 2014). A linear rate of convergence can be shown for the algorithm that minimizes (15) by alternating minimization over $\boldsymbol{D}$ and $\mathbf{\Omega}$ (Price et al., 2021).

4.3 | Bayesian approach

We introduce two Bayesian methods that construct priors to constrain the similarities within subsets of groups. Both approaches do not require prior information to build group relations. This property is particularly desirable because, in most cases, we may not have the structural information for K groups of data.

Peterson et al. (2015) proposed the Markov Random Field (MRF) prior for the graphs $G^{(k)} = (V, E^{(k)}), k = 1, \ldots, K$ to encourage the selection of edge indices in related graphs. In addition, the selection of edge indices is controlled by two variables: a random matrix $\mathbf{\Theta} \in \mathbb{R}^{K \times K}$, whose k,k'th entry measures the degree of similarity between group k and k' and an edge-specific coefficient vector $\boldsymbol{\nu}$ reflecting the probability of the corresponding edge being selected. Let $\mathbf{e}_{i,j} \in \mathbb{R}^K$,

$1 \le i < j \le p$, be a binary vector indicating the existence of an edge between node i and j among K groups. The prior for $\mathbf{e}_{i,j}$ is expressed as

$$p\left(\mathbf{e}_{i,j} \mid v_{i,j}, \boldsymbol{\Theta}\right) \propto \exp\left(v_{i,j} 1^T \mathbf{e}_{i,j} + \mathbf{e}_{i,j}^T \boldsymbol{\Theta} \mathbf{e}_{i,j}\right), \quad 1 \le i < j \le p,$$

where $v_{i,j}$ controls the probability that an edge between the ith and jth nodes is selected. Specifically, $\boldsymbol{v}$ controls the sparsity of graphs. The joint prior for the graphs $\mathbf{G} = \left\{G^{(1)}, \dots G^{(K)}\right\}$ is

$$p(\mathbf{G} \mid \mathbf{v}, \boldsymbol{\Theta}) = \prod_{i<j} p\left(\mathbf{e}_{i,j} \mid v_{i,j}, \boldsymbol{\Theta}\right).$$

The Beta prior is placed on the elements of $\boldsymbol{v}$. Specifically, we have that $v_{i,j} \sim \text{Beta}(1,4)$, which results in a sparse graph. Meanwhile, the spike-and-slab prior is placed on the group similarity matrix Θ, conditioned on the latent random variables $z = \left(z_{i,j}\right)_{i<j}$, to allow discrimination between zero similarity and positive similarity, where $z_{i,j}$, $1 \le i < j \le K$, is a binary random variable denoting the relation between groups i and j. If $z_{i,j} = 1$, the two groups are related, otherwise they are not. Then, the prior on $\boldsymbol{\Theta}$ is defined as

$$p(\boldsymbol{\Theta} \mid z) = \prod_{i<j} p\left(\theta_{i,j} \mid z_{i,j}\right);$$
$$p\left(\theta_{i,j} \mid z_{i,j}\right) = \left(1 - z_{i,j}\right)\delta + z_{i,j} \frac{\beta}{\Gamma(\alpha)} \theta_{i,j}^{\alpha-1} \exp\left(-\beta\theta_{i,j}\right),$$

where $\Gamma(\cdot)$ is the Gamma function and α,β are hyperparameters. Noting that the distribution of $\boldsymbol{z}$ determines the relatedness between groups, Peterson et al. (2015) constructed a Bernoulli prior on $\boldsymbol{z}$:

$$p(z \mid w) = \prod p\left(z_{i,j} \mid w\right);$$
$$p\left(z_{i,j} \mid w\right) = w^{z_{i,j}} (1-w)^{1-z_{i,j}},$$

where $w \in [0,1]$ is a hyperparameter. Shaddox et al. (2020) recently proposed an alternative prior on $\boldsymbol{z}$ in the setting where data types are different. Finally, we apply the G-Wishart prior distribution to the inverse covariance matrices conditional on the graph structures $G^{(k)} = (V, E^{(k)}), k = 1, \dots, K$:

$$p(\boldsymbol{\Omega} \mid G, b, \mathbf{D}) \propto \prod_{k=1}^{K} \left|\Omega^{(k)}\right|^{(b-2)\backslash 2} \exp\left\{-2^{-1} \text{tr}\left(\Omega^{(k)} \mathbf{D}\right)\right\}, \quad \Omega^{(k)} \in P_{G^{(k)}},$$

where $\mathbf{D}$ is a preselected positive definite matrix and $b > 2$ is a constant. The set $P_{G^{(k)}}$ contains all positive definite matrices that have the edge structure $E^{(k)}$. Shaddox et al. (2018) proposed a similar framework as Peterson et al. (2015), but adopted a continuous shrinkage prior, instead of the G-Wishart prior used in Peterson et al. (2015), resulting in a computationally more efficient procedure.

Lin et al. (2017) applied the MRF prior to jointly estimate multiple graphical models but in slightly different setting—they adopted a Bayesian version of the neighborhood regression (Meinshausen & Bühlmann, 2006), see (11), and proposed a hierarchical prior on the set of regression coefficients $\boldsymbol{\Theta} = \left\{\boldsymbol{\Theta}^{(1)}, \dots, \boldsymbol{\Theta}^{(K)}\right\}$. To encourage the sparsity pattern, a spike-and-slab prior, see (8), is placed on $\boldsymbol{\Theta}$ conditioning on binary latent factors $z = \left\{z_{i,j}\right\}$ with $z = \left(z_{i,j}^{(1)}, \dots, z_{i,j}^{(K)}\right) \in \{0,1\}^K$, $i < j$. Lin et al. (2017) studied the setting where the group index is a tuple (s,t) with $s \in \mathcal{S} \subseteq \mathbb{N}$ being the location and $t \in \mathcal{N} \subseteq \mathbb{N}$ being the time. Intuitively, groups with the same time index t should have similar graph structures, while groups with the same location s and small pairwise time difference, that is, $|t - t'| = 1$ should have similar graph structures. Let $A_s = \{(s,t,s',t'): s \neq s', t = t'\}$ and $B_t = \{(s,t,s',t'): s = s', |t - t'| = 1\}$. Let $\lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ be a set of hyperparameters. The indicator function $1_a(x)$ outputs 1 when $x = a$, and 0 otherwise. An MRF prior is placed on $\boldsymbol{z}$ to represent the pairwise interactions between groups:

$$p(\mathbf{z}_{i,j}|\lambda) \propto \exp\left[\lambda_1 \sum_{s\in\mathcal{S},t\in\mathcal{T}} 1_1\left(z_{i,j}^{(s,t)}\right) + \lambda_2 \sum_{A_s}\left\{1_0\left(z_{i,j}^{(s,t)}\right)1_0\left(z_{i,j}^{(s\prime,t\prime)}\right) + 1_1\left(z_{i,j}^{(s,t)}\right)1_1\left(z_{i,j}^{(s\prime,t\prime)}\right)\right\} + \lambda_3 \sum_{B_t}\left\{1_0\left(z_{i,j}^{(s,t)}\right)1_0\left(z_{i,j}^{(s\prime,t\prime)}\right) + 1_1\left(z_{i,j}^{(s,t)}\right)1_1\left(z_{i,j}^{(s\prime,t\prime)}\right)\right\}\right].$$

Under this prior distribution, for any tuple (s,t,s',t') in A_s or B_t, $z_{i,j}^{(s,t)}$ and $z_{i,j}^{(s\prime,t\prime)}$ have a higher probability to have the same sign. In addition, by varying the values of $\lambda_1,\lambda_2,\lambda_3$ one can weigh the importance of individual components, spatially similar components, and temporally similar components, respectively. Lin et al. (2017) provided a MATLAB implementation of the joint temporal and spatial estimation.

Another example of describing the non-uniform relationships between groups is to build a hierarchical diagram, such as a tree graph. In the Bayesian formulation, we can implement this by constructing a hierarchical prior in the factor form. Oates and Mukherjee (2014) proposed a structure learning trees (SLTs) prior that encodes the hierarchical information among groups. Although the SLTs prior is not originally designed for estimating Gaussian graphical models, the prior can be applied to regularize the structure of the inverse covariance matrices.

5 | ESTIMATING DIFFERENTIAL GRAPHICAL MODELS

In contrast to joint estimation, several applications in biomedical research, such as analyzing the gene expression differences in normal cells and cancer cells or differences between the test group and control group, consider the case where $K=2$. Different from the methods in Sections 3 and 4, we will be focusing on finding the "differences" rather than finding the "similarities." In the high dimensional setting, we assume that the difference of two graphs is sparse. Although the differences between two graphs can be naively estimated by using a joint estimation method first and then finding the difference, procedures that directly estimate the difference are statistically more efficient (Shojaie, 2021).

In this section, we briefly overview two approaches that estimate graph differences, the direct estimation method and the regularization based approach. For a detailed introduction, see Shojaie (2021) for a recent review.

5.1 | Direct estimation

The direct approach estimates the difference $\mathbf{\Delta}=\mathbf{\Omega}^{(1)}-\mathbf{\Omega}^{(2)}$ without explicitly estimating individual precision matrices $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$. This approach potentially fits a broader class of precision matrices as the individual precision matrices, $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$, need not be sparse, but only the difference $\mathbf{\Delta}$ is assumed sparse. In addition, jointly estimating $\mathbf{\Omega}^{(1)}$ and $\mathbf{\Omega}^{(2)}$ can be challenging when the sparse assumption is violated. Zhao et al. (2014) directly estimated the difference $\mathbf{\Delta}$ by solving a constrained minimization problem, noting that, by definition, we have $\mathbf{\Sigma}^{(1)}\mathbf{\Delta}\mathbf{\Sigma}^{(2)}-\mathbf{\Sigma}^{(1)}+\mathbf{\Sigma}^{(2)}=\mathbf{0}$. Consequently, estimating the differential graph $\mathbf{\Delta}=\mathbf{\Omega}^{(1)}-\mathbf{\Omega}^{(2)}$ can be achieved by minimizing the following objective:

$$\widehat{\mathbf{\Delta}}=\operatorname{argmin}\|\mathbf{\Delta}\|_1, \quad \text{subject to}\left|\widehat{\mathbf{\Sigma}}^{(1)}\mathbf{\Delta}\widehat{\mathbf{\Sigma}}^{(2)}-\widehat{\mathbf{\Sigma}}^{(1)}+\widehat{\mathbf{\Sigma}}^{(2)}\right|_\infty \le \lambda_1,$$

which is an extension of the CLIME (Cai et al., 2011) method. Xu and Gu (2016); Yuan et al. (2017) utilized the symmetry property $\mathbf{\Sigma}^{(1)}\mathbf{\Delta}\mathbf{\Sigma}^{(2)}=\mathbf{\Sigma}^{(2)}\mathbf{\Delta}\mathbf{\Sigma}^{(1)}$ and hence $2^{-1}\left(\mathbf{\Sigma}^{(1)}\mathbf{\Delta}\mathbf{\Sigma}^{(2)}+\mathbf{\Sigma}^{(2)}\mathbf{\Delta}\mathbf{\Sigma}^{(1)}\right)-\mathbf{\Sigma}^{(1)}+\mathbf{\Sigma}^{(2)}=\mathbf{0}$. They defined the objective function as

$$\widehat{\mathbf{\Delta}}=\operatorname{argmin}\widehat{\ell}(\mathbf{\Delta})+\lambda_1\|\mathbf{\Delta}\|_1; \tag{16}$$

$$\widehat{\ell}(\mathbf{\Delta})=\frac{1}{2}\operatorname{tr}\left(\mathbf{\Delta}\widehat{\mathbf{\Sigma}}^{(1)}\mathbf{\Delta}\widehat{\mathbf{\Sigma}}^{(2)}\right)-\operatorname{tr}\left\{\mathbf{\Delta}\left(\widehat{\mathbf{\Sigma}}^{(1)}-\widehat{\mathbf{\Sigma}}^{(2)}\right)\right\},$$

where the Hessian of the objective with respect to $\mathbf{\Delta}$ is $\left(\widehat{\mathbf{\Sigma}}^{(1)}\otimes\widehat{\mathbf{\Sigma}}^{(2)}+\widehat{\mathbf{\Sigma}}^{(2)}\otimes\widehat{\mathbf{\Sigma}}^{(1)}\right)\backslash 2$, which is positive semi-definite. Therefore, $\widehat{\ell}(\mathbf{\Delta})+\lambda_1\|\mathbf{\Delta}\|_1$ is a convex function with respect to $\mathbf{\Delta}$, hence a unique minimizer exists.

Direct estimation of differential graphs can be extended to other applications as well. Wang et al. (2021) proposed a procedure to estimate the differences of two autoregressive models by leveraging the connection between $\mathbf{\Delta}$ and the difference of a pair transition matrices. Wang et al. (2021) developed an efficient two-stage estimation procedure by first optimizing (16) and then using $\widehat{\boldsymbol{\Delta}}$ to solve a regularized least-squared problem in the second stage. Other recent work extends the direct estimation approach to more expressive structured differential graphs. Na et al. (2021) constructed a latent structure estimator where the underlying difference can be formulated as the sum of a low-rank and sparse matrix—a framework first discussed by Chandrasekaran et al. (2012). Zhao et al. (2019) extended the direct estimation approach to estimating the differential graph of functional data.

5.2 | Regularization based approach

The node-based learning framework (Mohan et al., 2012, 2014) assumes that most parts of the graph are shared, and the difference is generated by a node perturbation. When a node is perturbed, the edges connecting this node to others change across K groups. In addition to maximizing the degree of the overlapping structure between groups, the task is to detect perturbed nodes. An intuitive way to look for the perturbed node is to look at the difference of two graphs $\boldsymbol{\Omega}^{(1)}-\boldsymbol{\Omega}^{(2)}$. When the jth node is being perturbed, the corresponding jth row and jth column of $\boldsymbol{\Omega}^{(1)}-\boldsymbol{\Omega}^{(2)}$ will have non-zeros, constructing a unique symmetric row-column group. Given that there are several nodes being perturbed, $\boldsymbol{\Omega}^{(1)}-\boldsymbol{\Omega}^{(2)}$ will be the union of the row-column groups, each stemming from a perturbed node. Using this concept, the row-column overlap norm (RCON) (Mohan et al., 2012, 2014) is designed to encourage sparsity in the union of the row-column groups:

$$P_{\mathrm{RCON}}(\boldsymbol{\Omega}) = \lambda_1 \sum_{k=1}^{2} \sum_{i,j} |\omega_{i,j}^{(k)}| + \lambda_2 G_q\left(\boldsymbol{\Omega}^{(1)} - \boldsymbol{\Omega}^{(2)}\right);$$

$$G_q(\mathbf{A}) = \min_{\mathbf{V}:\mathbf{A}=\mathbf{V}+\mathbf{V}^\top} f(\mathbf{V}), \;\; f(\mathbf{V}) = \sum_{j=1}^{p} \|\mathbf{v}_j\|_q,$$

where $\mathbf{v}_j$ is the jth column of $\mathbf{V}$. It is easy to see that when $q=1$, the RCON penalty is equivalent to the FGL penalty in (5). This penalty function simultaneously imposes sparse structure on both the individuals, $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$, and the difference $\boldsymbol{\Omega}^{(1)}-\boldsymbol{\Omega}^{(2)}$. As mentioned earlier, this method may not work well under the setting that $\boldsymbol{\Omega}^{(1)}$ and $\boldsymbol{\Omega}^{(2)}$ are not sparse. Additionally, Mohan et al. (2014) provided code for estimating differential graphs.

To infer the relative differences between two graphs in a Bayesian formulation, it is intuitive to place a prior on the differences of two graphs $\theta_{i,j}=E_{i,j}^{(1)}-E_{i,j}^{(1)}$, for every $i<j$. Since the difference $\theta_{i,j}$ is binary, either 0 (no difference) or 1 (difference), (Mitra et al., 2016) placed a Bernoulli prior distribution $\mathrm{Ber}(\pi)$ on $\theta_{i,j}$, $i<j$ where π follows a Beta distribution, specifying the tendency of being different on two graphs.

6 | JOINT ESTIMATION FROM TIME SERIES DATA

Time-varying graphical models (Kolar et al., 2010; Zhou et al., 2010b; Zhu & Koyejo, 2018) can be seen as extensions of joint graphical models with groups organized along the time index. The samples are assumed to be generated as $\mathbf{x}_i^{(t)} \sim \mathcal{N}\left(\mu^{(t)}, \left(\boldsymbol{\Omega}^{(t)}\right)^{-1}\right)$, $i=1, \ldots, n_t$, where $t=1, \ldots, T$ is the time index. Under such a model, the estimation of time-varying precision matrices and corresponding dynamic networks is challenging as data scarcity is a serious issue: in many problems, we only observe a single sample at each time point. Therefore, to make the estimation possible, structural assumptions are imposed on how the underlying precision matrices and dynamic networks change over time. Such assumptions control the model complexity and allow for the development of estimation procedures. Examples of structural assumptions on temporal dynamics include piecewise constant and smoothly changing precision matrices, as well as combinations of both. Piece-wise constant structure captures a discrete temporal evolution from one stage to another. For example, the gene regulatory network in a fruit fly can undergo structural changes as the fruit fly develops from an embryo to an adult state. Smooth temporal structure can be used to model the dynamic functional connectivity of brain networks that exhibit smooth temporal evolution from one brain state to another (Shine et al., 2016). The temporal dynamics of crime rates are often modeled as a combination of smooth dynamics and sudden jumps, where the jumps capture sudden serious crime events. In this section, we will discuss how to apply the FGL penalty and its variants to build a piecewise constant structure. We also introduce a joint estimation framework of multiple autoregressive models to model smooth temporal data.

6.1 | Regularized estimation

The FGL penalty has been widely used in time-varying graphical models to model piecewise constant dynamics (Hallac et al., 2017; Kolar et al., 2010; Kolar & Xing, 2012; Monti et al., 2014). For instance, Smooth Incremental Graphical Lasso Estimation (SINGLE) (Monti et al., 2014) applies the FGL framework to enforce the similarity between consecutive precision matrices:

$$\mathbf{P}_{\text{SINGLE}}(\mathbf{\Omega}) = \lambda_1 \sum_{t=1}^{T} \sum_{i \neq j} |\, \omega_{i,j}^{(t)} \,| + \lambda_2 \sum_{t=2}^{T} \sum_{i \neq j} |\, \omega_{i,j}^{(t)} - \omega_{i,j}^{(t-1)} \,| . \tag{17}$$

The first term encourages the sparsity of each graph and the second term regularizes the "jumps" across time. On the other hand, Group-Fused Graphical Lasso (GFGL) (Gibberd & Nelson, 2017) introduces the Frobenius norm as an alternative to encourage neighboring similarity:

$$\mathbf{P}_{\text{GFGL}}(\mathbf{\Omega}) = \lambda_1 \sum_{t=1}^{T} \sum_{i \neq j} |\, \omega_{i,j}^{(t)} \,| + \lambda_2 \sum_{t=2}^{T} \left\| \mathbf{\Omega}_{-ii}^{(t)} - \mathbf{\Omega}_{-ii}^{(t-1)} \right\|_F,$$

where $\mathbf{\Omega}_{-ii}^{(t)}$ denotes the precision matrix $\mathbf{\Omega}^{(t)}$ with the diagonal part removed. One may wonder what are the differences in the structure assumptions between the Frobenius norm and the ℓ_1-norm in (17). The ℓ_1-norm regularizes individual changes, while the Frobenius norm assumes global changes, implying that several edges within a graph will change simultaneously.

While the methods introduced in the last paragraph encourage the similarity of two neighboring graphs, the graph that is one-step ahead and that of one-step behind, another idea is to enforce the similarities within multiple steps ahead and behind. This can be done by creating a moving window index set (Yang & Peng, 2020) $\mathcal{N}_w(t)$ for each time point $t = 1, \ldots, T$. Consider a window of length $2w$. At every time point t, we look at data w-steps ahead and w-steps behind and hence the index set is $\mathcal{N}_w(t) = \{i = 1, \ldots, T : |t - i| \leq 2w\}$. Note that the index set $\mathcal{N}_w(t)$ also includes t itself. Then, we apply the GGL penalty to the components in the index set.

6.2 | Kernel smoothing graphical models

Another way to construct smoothly varying graphs is by using an autoregressive structure. This model assumes that each data point is a linear combination of previous data points with additional independent noise. Consider the lag-1 case, where $\mathbf{x}^{(t)}$ is a linear transform of $\mathbf{x}^{(t-1)}$ with independent noise $\varepsilon^{(t)} \tilde{\mathcal{N}}\left(\mathbf{0}, \mathbf{G}^{(t)}\right)$:

$$\mathbf{x}^{(t)} = \mathbf{A}\mathbf{x}^{(t-1)} + \varepsilon^{(t)}, \quad t = 1, \ldots, T,$$

where $\mathbf{A} \in \mathbb{R}^{p \times p}$ is the transition matrix. Consequently, the covariance matrix is smoothly varying along t if $\mathbf{G}^{(t)}$ is a smooth function of t:

$$\mathbf{\Sigma}^{(t)} = \mathbf{A}\mathbf{\Sigma}^{(t-1)}\mathbf{A}^T + \mathbf{G}^{(t)}, \quad t = 1, \ldots, T. \tag{18}$$

Motivated by this structure, Zhou et al. (2010) proposed a kernel based method to estimate a smooth time-varying covariance structure. First, a weighted sum of the sample covariance matrices $\widehat{\mathbf{\Sigma}}^{(1)}, \ldots, \widehat{\mathbf{\Sigma}}^{(T)}$ is computed as

$$\widehat{S}^{(t)} = \frac{\sum_{s=1}^{T} w(s,t)\widehat{\boldsymbol{\Sigma}}^{(s)}}{\sum_{s'=1}^{T} w(s',t)},$$

where the weights are constructed by a symmetric nonnegative kernel function $K(|s-t|/h)$. This ensures that the estimated covariance is smoothly varying over time. Subsequently, the precision matrix is estimated using the following objective:

$$\widehat{\boldsymbol{\Omega}}^{(t)} = \operatorname{argmax}\ n\left[\frac{1}{2}\log\left\{\det\left(\boldsymbol{\Omega}^{(t)}\right)\right\} - \frac{1}{2}\operatorname{tr}\left(\widehat{\mathbf{S}}^{(t)}\boldsymbol{\Omega}^{(t)}\right)\right] - \lambda_n \sum_{i \neq j} |\, \omega_{i,j}^{(t)} \,|, \quad t = 1, \ldots, T.$$

The kernel smoothing method can also be extended to model two-way continuous changes. For instance, the ages of subjects from the fMRI dataset vary across an interval, and one can parametrize the transition matrices as $A(u)$ with u taking values in a closed subset of the real line. This model is smooth in two aspects: across the temporal domain and labels (groups). Hence, we have the following autoregressive model:

$$\begin{aligned}
&x_{i,t} = A(u_i)x_{i,t-1} + \varepsilon_{i,t}, \quad i = 1, \ldots, n, \ t = 2, \ldots, T. \\
&\boldsymbol{\Sigma}(u) = A(u)\boldsymbol{\Sigma}(u)A(u)^{\top} + \sigma^2 I.
\end{aligned}$$

The Kernel-Smoothing Estimator (KSE) (Qiu et al., 2016) first uses a kernel based estimator for the covariance matrix and then uses the CLIME (Cai et al., 2011) method introduced in Section 3.1 to recover precision matrices. Consider a set of n data $Y = \{y_1, \ldots, y_n\}$, where $y_i = \{y_{i,1}, \ldots, y_{i,T}\} \in \mathbb{R}^{p \times T}$ and with label $u_i \in [0,1]$. The estimated covariance model of the label $u_0 \in [0,1]$ is formulated as follows:

$$\begin{aligned}
&\widehat{S}(u_0) = \sum_{i=1}^{n} w_i(u_0,h)\widehat{\Sigma}_i; \\
&w_i(u_0,h) := \frac{c(u_0)}{nh} K\left(\frac{u_i - u_0}{h}\right); \\
&c(u_0) = \begin{cases} 2I, & u_0 \in \{0,1\}, \\ 1I, & u_0 \in (0,1), \end{cases}; \\
&\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}y_{i,t}^{\top},
\end{aligned}$$

where w_i is the kernel-based weight with a predefined scale h, $K(\cdot)$ is the kernel, $c(u_0)$ determines the boundary value, and $\widehat{\boldsymbol{\Sigma}}_i$ is the sampled covariance of the time-series data. After obtaining $\widehat{S}(u_0)$, the precision matrix $\widehat{\boldsymbol{\Omega}}(u_0)$ is obtained using CLIME in (4):

$$\begin{aligned}
&\widehat{\boldsymbol{\Omega}}(u_0) = \operatorname{argmin}\|\boldsymbol{\Omega}(u_0)\|_1, \\
&\text{subject to}\left|\widehat{S}(u_0)\boldsymbol{\Omega}(u_0) - I\right|_{\infty} \leq \lambda_1.
\end{aligned}$$

Under this framework, the kernel trick is used to capture the assumption that the covariance matrices are smoothly varying across labels. In addition, the Euclidean distance of two labels reflects the similarity of the two groups, capturing the dependence structure. The kernel-based method can be applied to general joint estimation, where the sampled covariance of time-series data is replaced by the sample covariance of data with the same labels.

7 | OPEN PROBLEMS

Existing and emerging biological data and applications will require novel approaches to joint graphical models. We discuss some of these emerging applications briefly. Joint estimation of functional connectivity networks across multiple subjects allows scaling of the effective sample size and computation of estimates that are more robust to outliers. The joint estimators of brain connectivity networks could be applied to task-based fMRI scans to study group dynamic functional connectivity patterns (Andersen et al., 2018; Calhoun et al., 2014; Gonzalez-Castillo & Bandettini, 2018). While this manuscript is focused on joint estimation with the same set of nodes, one potential direction is to extend it to multiple sources, that is, multimodal data. Recent technologies (Abreu et al., 2018; Huster et al., 2012) have demonstrated the availability of conducting concurrent measurements of EEG and fMRI signals, allowing the estimation of multiple sources possible in the future. While EEG has a higher temporal resolution and fMRI features a higher spatial resolution, we believe that joint estimation with multiple sources could compensate for the limits of the measurement techniques and provide better estimation results. Some recent work (Li & Li, 2021; Lock et al., 2013) has developed methodologies to integrate data from different modalities, however, joint estimation of graphical models from multimodal data is still an open problem.

Approaches for the estimation of the joint graphical models presented in this survey largely rely on penalized estimation, where the penalty biases the estimates toward the assumed structure. Quantifying statistical uncertainty about the model parameters, that is, performing hypothesis tests and constructing confidence intervals, is challenging when penalized estimators are used due to the induced bias and model selection that is implicitly performed. There has been recent work on statistical inference for low-dimensional parameters in graphical models (Barber & Kolar, 2018; Janková & van de Geer, 2015; Janková & van de Geer, 2017; Ren et al., 2015; Wang & Kolar, 2016; Yu et al., 2016; Yu et al., 2020) based on the ℓ_1-penalized estimator in the first stage. However, these approaches were developed only in the setting where parameters of one graph are being inferred. In contrast, work on statistical inference for joint graphical models is much more sparse. Xia et al. (2015); Belilovsky et al. (2016); Liu (2017); Kim et al. (2021) developed techniques for statistical inference in differential graphical models, while Wang and Kolar (2014); Lu et al. (2018); Wang et al. (2020) focused on graphical models for time series data. Wang and Shojaie (2021) developed a hierarchical testing procedure for joint inferences of multiple graphs on Hawkes processes, albeit in non-Gaussian settings. Developing the corresponding inferential techniques for estimators obtained using coarse-grained and fine-grained penalties is an interesting area open for future research.

8 | CONCLUSION

This manuscript has introduced joint Gaussian graphical model estimation methods for joint data with shared structure across multiple groups. In particular, we have considered several examples of extending classical statistical inference methods to joint estimation settings, including the MLE based estimator, neighborhood regression, and the CLIME estimator. We have discussed several methods that exploit coarse-grained structures using a global regularization method that encourages a shared coarse-grained structure across all groups. In contrast, the fine-grained structural regularization methods further partition the groups into subgroups per node, encouraging local shared regularity. With two groups, differential graphs are often a highly effective approach. We have also discussed the applications of joint estimation techniques to the estimation of graphical models from time-series data.

8.1 | Further reading

Recent developments of joint statistical inference are primarily focused on Gaussian graphical models. Other types of graphical models, including discrete graphical models (Drton, 2009; Drton & Richardson, 2008), semiparametric/nonparametric graphical models (Liu et al., 2012; Sun et al., 2015), and latent graphical models (Chandrasekaran et al., 2012), have been well studied for single graph estimation. While such models have broad applications, joint estimation in these models is less studied.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Katherine Tsai: Conceptualization (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Oluwasanmi Koyejo:** Conceptualization (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Mladen Kolar:** Conceptualization (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal).

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Katherine Tsai https://orcid.org/0000-0002-7243-5260
Oluwasanmi Koyejo https://orcid.org/0000-0002-4023-419X
Mladen Kolar https://orcid.org/0000-0001-7353-3404

RELATED WIREs ARTICLE

Differential network analysis: A statistical perspective

REFERENCES

Abreu, R., Leal, A., & Figueiredo, P. (2018). Eeg-informed fmri: A review of data analysis methods. *Frontiers in Human Neuroscience*, *12*, 29.

Andersen, M., Winther, O., Hansen, L. K., Poldrack, R., & Koyejo, O. (2018). Bayesian structure learning for dynamic brain connectivity. In Proceedings of the twenty-first international conference on artificial intelligence and statistics, pp. 1436–1446.

Atay-Kayis, A., & Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, *92*(2), 317–335.

Barber, R. F., & Kolar, M. (2018). Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. *The Annals of Statistics*, *46*(6), 3422–3450.

Belilovsky, E., Varoquaux, G., & Blaschko, M. B. (2016). Testing for differences in Gaussian graphical models: Applications to brain connectivity. In Advances in neural information processing systems, volume 29, pp. 595–603.

Bilgrau, A. E., Peeters, C. F., Eriksen, P. S., Bøgsted, M., & van Wieringen, W. N. (2020). Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *Journal of Machine Learning Research*, *21*(26), 1–52.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.

Cai, T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607.

Calhoun, V. D., Miller, R., Pearlson, G., & Adali, T. (2014). The chronnectome: Time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, *84*(2), 262–274.

Chan, T. E., Stumpf, M. P., & Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, *5*(3), 251–267.

Chandrasekaran, V., Parrilo, P. A., & Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, *40*(4), 1935–1967.

Chiquet, J., Grandvalet, Y., & Ambroise, C. (2011). Inferring multiple graphical structures. *Statistics and Computing*, *21*(4), 537–553.

Chun, H., Zhang, X., & Zhao, H. (2015). Gene regulation network inference with joint sparse Gaussian graphical models. *Journal of Computational and Graphical Statistics*, *24*(4), 954–974.

Colclough, G. L., Woolrich, M. W., Harrison, S. J., López, P. A. R., Valdes-Sosa, P. A., & Smith, S. M. (2018). Multi-subject hierarchical inverse covariance modelling improves estimation of functional brain networks. *NeuroImage*, *178*, 370–384.

Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 373–397.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, *90*(1), 196–212.

Dondelinger, F., Lèbre, S., & Husmeier, D. (2013). Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine Learning*, *90*(2), 191–230.

Drton, M. (2009). Discrete chain graph models. *Bernoulli*, *15*(3), 736–753.

Drton, M., & Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, *4*(1), 365–393.

Drton, M., & Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, *91*(3), 591–602.

Drton, M., & Richardson, T. S. (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 287–309.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376.

Foti, N. J., & Fox, E. B. (2019). Statistical model-based approaches for functional connectivity analysis of neuroimaging data. *Current Opinion in Neurobiology*, *55*, 48–54.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, *7*, 601–620.

Gan, L., Yang, X., Nariestty, N. N., & Liang, F. (2019). Bayesian joint estimation of multiple graphical models. In Proceedings of the conference on neural information processing systems.

Gibberd, A. J., & Nelson, J. D. (2017). Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, *26*(3), 623–634.

Gonzalez-Castillo, J., & Bandettini, P. A. (2018). Task-based dynamic functional connectivity: Recent findings and open questions. *NeuroImage*, *180*, 526–533.

Greenewald, K., Park, S., Zhou, S., & Giessing, A. (2017). Time-dependent spatially varying graphical models, with application to brain fmri data analysis. In Advances in neural information processing systems, volume 30, pp. 5832–5840.

Guo, J., Levina, E., Michailidis, G., & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, *98*(1), 1–15.

Hallac, D., Park, Y., Boyd, S., & Leskovec, J. (2017). Network inference via the time-varying graphical lasso. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 205–213.

Hao, B., Sun, W. W., Liu, Y., & Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, *18*, 217–211.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., & Poldrack, R. (2013). Big & quic: Sparse inverse covariance estimation for a million variables. In Advances in neural information processing systems, volume 26, pp. 3165–3173.

Huster, R. J., Debener, S., Eichele, T., & Herrmann, C. S. (2012). Methods for simultaneous eeg-fmri: An introductory review. *Journal of Neuroscience*, *32*(18), 6053–6060.

Janková, J., & van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, *9*(1), 1205–1229.

Janková, J., & van de Geer, S. A. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, *26*(1), 143–162.

Kim, B., Liu, S., & Kolar, M. (2021). Two-sample inference for high-dimensional markov networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(5), 939–962.

Kling, T., Johansson, P., Sanchez, J., Marinescu, V. D., Jörnsten, R., & Nelander, S. (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*, *43*(15), e98.

Kolar, M., Song, L., Ahmed, A., & Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, *4*, 94–123.

Kolar, M., & Xing, E. P. (2012). Estimating networks with jumps. *Electronic Journal of Statistics*, *6*, 2069–2106.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

Lee, W., & Liu, Y. (2015). Joint estimation of multiple precision matrices with common structures. *The Journal of Machine Learning Research*, *16*(1), 1035–1062.

Lenkoski, A., & Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the g-wishart prior. *Journal of Computational and Graphical Statistics*, *20*(1), 140–157.

Li, Q., & Li, L. (2021). Integrative factor regression and its inference for multimodal data analysis. *Journal of the American Statistical Association*, 1–15. doi:10.1080/01621459.2021.1914635

Li, Z., Mccormick, T., & Clark, S. (2019). Bayesian joint spike-and-slab graphical lasso. In Proceedings of the 36th international conference on machine learning, volume 97 of proceedings of machine learning research, pp. 3877–3885.

Lin, Z., Wang, T., Yang, C., & Zhao, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics*, *73*(3), 769–779.

Liu, H., Han, F., Yuan, M., Lafferty, J., & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, *40*(4), 2293–2326.

Liu, W. (2017). Structural similarity and difference testing on multiple sparse Gaussian graphical models. *The Annals of Statistics*, *45*(6), 2680–2707.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, *7*(1), 523–542.

Lu, J., Kolar, M., & Liu, H. (2018). Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research*, *18*(203), 1–78.

Lukemire, J., Kundu, S., Pagnoni, G., & Guo, Y. (2021). Bayesian joint modeling of multiple brain functional networks. *Journal of the American Statistical Association*, *116*(534), 518–530.

Lurie, D. J., Kessler, D., Bassett, D. S., Betzel, R. F., Breakspear, M., Kheilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A., McIntosh, A. R., Poldrack, R. A., Shine, J. M., Thompson, W. H., Bielczyk, N. Z., Douw, L., Kraft, D., Miller, R. L., Muthuraman, M., Pasquini, L., … Calhoun, V. D. (2020). Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Network Neuroscience*, *4*(1), 30–69.

Ma, J., & Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *The Journal of Machine Learning Research*, *17*(1), 5777–5824.

Maathuis, M., Drton, M., Lauritzen, S., & Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Manning, J. R., Zhu, X., Willke, T. L., Ranganath, R., Stachenfeld, K., Hasson, U., Blei, D. M., & Norman, K. A. (2018). A probabilistic approach to discovering dynamic full-brain functional connectivity patterns. *NeuroImage*, *180*, 243–252.

Marlin, B. M., & Murphy, K. P. (2009). Sparse Gaussian graphical models with unknown block structure. In Proceedings of the 26th annual international conference on machine learning, pp. 705–712.

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

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Mitra, R., Müller, P., & Ji, Y. (2016). Bayesian graphical models for differential pathways. *Bayesian Analysis*, *11*(1), 99–124.

Mohammadi, A., & Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, *10*(1), 109–138.

Mohan, K., Chung, M., Han, S., Witten, D., Lee, S.-I., & Fazel, M. (2012). Structured learning of Gaussian graphical models. In Advances in neural information processing systems, volume 25, pp. 620–628.

Mohan, K., London, P., Fazel, M., Witten, D., & Lee, S.-I. (2014). Node-based learning of multiple Gaussian graphical models. *The Journal of Machine Learning Research*, *15*(1), 445–488.

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., & Montana, G. (2014). Estimating time-varying brain connectivity networks from functional mri time series. *NeuroImage*, *103*, 427–443.

Na, S., Kolar, M., & Koyejo, O. (2021). Estimating differential latent variable graphical models with applications to brain connectivity. *Biometrika*, *108*(2), 425–442.

Oates, C., & Mukherjee, S. (2014). Joint structure learning of multiple non-exchangeable networks. In Proceedings of the seventeenth international conference on artificial intelligence and statistics, volume 33, pp. 687–695.

Pan, W., & Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, *8*(41), 1145–1164.

Peterson, C., Stingo, F. C., & Vannucci, M. (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, *110*(509), 159–174.

Pierson, E., Consortium, G., Koller, D., Battle, A., & Mostafavi, S. (2015). Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Computational Biology*, *11*(5), e1004220.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Price, B. S., Molstad, A. J., & Sherwood, B. (2021). Estimating multiple precision matrices with cluster fusion regularization. *Journal of Computational and Graphical Statistics*, *30*, 1–12.

Qiu, H., Han, F., Liu, H., & Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(2), 487–504.

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, *43*(3), 991–1026.

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*(3), 391–411.

Saegusa, T., & Shojaie, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electronic journal of statistics*, *10*(1), 1341–1392.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 32.

Schwab, S., Harbord, R., Zerbi, V., Elliott, L., Afyouni, S., Smith, J. Q., Woolrich, M. W., Smith, S. M., & Nichols, T. E. (2018). Directed functional connectivity using dynamic graphical models. *NeuroImage*, *175*, 340–353.

Shaddox, E., Peterson, C. B., Stingo, F. C., Hanania, N. A., Cruickshank-Quinn, C., Kechris, K., Bowler, R., & Vannucci, M. (2020). Bayesian inference of networks across multiple sample groups and data types. *Biostatistics*, *21*(3), 561–576.

Shaddox, E., Stingo, F. C., Peterson, C. B., Jacobson, S., Cruickshank-Quinn, C., Kechris, K., Bowler, R., & Vannucci, M. (2018). A Bayesian approach for learning gene networks underlying disease severity in copd. *Statistics in Biosciences*, *10*(1), 59–85.

Shan, L., & Kim, I. (2018). Joint estimation of multiple Gaussian graphical models across unbalanced classes. *Computational Statistics & Data Analysis*, *121*, 89–103.

Shan, L., Qiao, Z., Cheng, L., & Kim, I. (2020). Joint estimation of the two-level Gaussian graphical models across multiple classes. *Journal of Computational and Graphical Statistics*, *29*(3), 562–579.

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., Moodie, C. A., & Poldrack, R. A. (2016). The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron*, *92*(2), 544–554.

Shojaie, A. (2021). Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(2), e1508.

Skripnikov, A., & Michailidis, G. (2019). Regularized joint estimation of related vector autoregressive models. *Computational Statistics & Data Analysis*, *139*, 164–177.

Sun, S., Kolar, M., & Xu, J. (2015). Learning structured densities via infinite dimensional exponential families. In Advances in neural information processing systems, volume 28, pp. 2287–2295.

Sun, W., Wang, J., & Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, *6*, 148–167.

Tan, L. S., Jasra, A., De Iorio, M., & Ebbels, T. M. (2017). Bayesian inference for multiple Gaussian graphical models with application to metabolic association networks. *The Annals of Applied Statistics*, *11*(4), 2222–2251.

Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., & Thirion, B. (2010). Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling. In International conference on medical image computing and computer-assisted intervention, pp. 200–208.

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

Wang, J., & Kolar, M. (2014). Inference for sparse conditional precision matrices. *arXiv Preprint*, *1412*, 7638.

Wang, J., & Kolar, M. (2016). Inference for high-dimensional exponential family graphical models. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics, volume 51 of proceedings of machine learning research* (pp. 1042–1050). PMLR.

Wang, X., Kolar, M., & Shojaie, A. (2020). Statistical inference for networks of high-dimensional point processes. *arXiv Preprint*.

Wang, X., & Shojaie, A. (2021). Joint estimation and inference for multi-experiment networks of high-dimensional point processes. *arXiv Preprint*.

Wang, Y., Ma, J., & Shojaie, A. (2021). Direct estimation of differential granger causality between two high-dimensional time series. *arXiv Preprint*.

Xia, Y., Cai, T., & Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, *102*(2), 247–266.

Xu, P., & Gu, Q. (2016). Semiparametric differential graph models. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 1064–1072). Curran Associates, Inc.

Yajima, M., Telesca, D., Ji, Y., & Müller, P. (2014). Detecting differential patterns of interaction in molecular pathways. *Biostatistics*, *16*(2), 240–251.

Yang, J., & Peng, J. (2020). Estimating time-varying graphical models. *Journal of Computational and Graphical Statistics*, *29*(1), 191–202.

Yu, M., Gupta, V., & Kolar, M. (2016). Statistical inference for pairwise graphical models using score matching. In D. D. Lee, M. Sugiyama, U. von Luxburg, & R. Garnett (Eds.), *Advances in neural information processing systems 29*. Curran Associates, Inc.

Yu, M., Gupta, V., & Kolar, M. (2020). Simultaneous inference for pairwise graphical models with generalized score matching. *Journal of Machine Learning Research*, *21*(91), 1–51.

Yuan, H., Xi, R., Chen, C., & Deng, M. (2017). Differential network analysis via lasso penalized d-trace loss. *Biometrika*, *104*(4), 755–770.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*(1), 19–35.

Zhao, B., Wang, Y. S., & Kolar, M. (2019). Direct estimation of differential functional graphical models. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 2575–2585). Curran Associates, Inc.

Zhao, S. D., Cai, T. T., & Li, H. (2014). Direct estimation of differential networks. *Biometrika*, *101*(2), 253–268.

Zhou, S., Lafferty, J., & Wasserman, L. (2010). Time varying undirected graphs. *Machine Learning*, *80*(2), 295–319.

Zhu, Y., & Koyejo, O. (2018). *Clustered fused graphical lasso* (pp. 487–496). *UAI*.

How to cite this article: Tsai, K., Koyejo, O., & Kolar, M. (2022). Joint Gaussian graphical model estimation: A survey. *WIREs Computational Statistics*, e1582. https://doi.org/10.1002/wics.1582