

# Estimation of Graphical Models: An Overview of Selected Topics

Li-Pang Chen 

Department of Statistics, National Chengchi University, Taipei, Taiwan

Email: [lchen723@nccu.edu.tw](mailto:lchen723@nccu.edu.tw)

## Summary

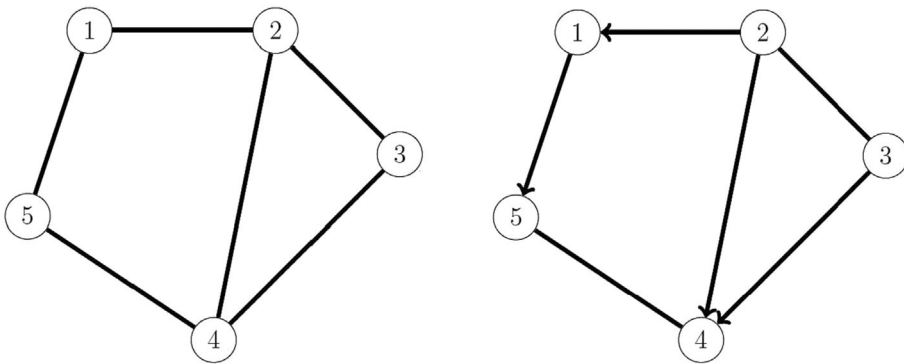
Graphical modelling is an important branch of statistics that has been successfully applied in biology, social science, causal inference and so on. Graphical models illuminate connections between many variables and can even describe complex data structures or noisy data. Graphical models have been combined with supervised learning techniques such as regression modelling and classification analysis with multi-class responses. This paper first reviews some fundamental graphical modelling concepts, focusing on estimation methods and computational algorithms. Several advanced topics are then considered, delving into complex graphical structures and noisy data. Applications in regression and classification are considered throughout.

**Key words:** computational algorithm; complex and noisy data; conditional inference; graphical LASSO; graphical models; multivariate linear models; network structure; optimisation; pairwise dependence; supervised learning.

## 1 Introduction

In the era of Big Data, high-dimensional data become available and make data structures complicated. One of the important features induced by high-dimensional data is *dependence* among variables, which frequently appears in many research topics such as genetic data (e.g. He *et al.*, 2019; Kumar *et al.*, 2020), social science (e.g. Gough *et al.*, 2018), spatial analysis (e.g. Besag, 1974; Okabe, 2017) and so on. To explore the complex dependence structure of high-dimensional variables, graphical models have proven to be useful tools.

Mathematically, let  $V$  be the set of vertices and let  $E \subset V \times V$  denote the set of edges. A *graph* is usually expressed as  $G = (V, E)$ . Figure 1 illustrates the concept of a graph, which is constructed by vertices  $V = \{1, 2, 3, 4, 5\}$  and edges  $E = \{(1, 2), (1, 5), (2, 3), (2, 4), (3, 4), (4, 5)\}$ . In the framework of graphical models, *undirected graphs* and *directed graphs* are two important branches. The main difference between the two types of graph is that the undirected graph only considers the relationship/pairwise dependence between any two vertices by using edges, while an arrow is added to edges in the directed graph. Unlike the undirected graph, as displayed in Figure 1, the directed graph emphasises that the ordering of the variables is taken into account and the relationship between any two vertices is not reversible (e.g.  $i \rightarrow j$  does not imply  $j \rightarrow i$ ). For practical applications, the undirected graph is usually applied to the study of network structure in the biological data, while the directed graph is frequently applied in causal inference (Edwards, 2000, chapter 8; Maathuis *et al.*, 2019, part



**Figure 1.** The graphical structures. The left graph is undirected; the right graph is directed.

IV). In this article, our discussion focuses on the undirected graph; the detailed descriptions of the directed graph can be found in Edwards (2000, chapter 7) and Scutari & Denis (2014).

In statistical perspectives, based on the undirected graph, vertices represent random variables and edges reflect the dependence structure of random variables. As a result, a crucial problem is to identify the dependence structure of high-dimensional random variables. In the framework of graphical model analysis, a big picture of fundamental concepts has been available in many monographs, such as Hastie *et al.* (2015, chapter 9), Hastie *et al.* (2008, chapter 17), Wainwright (2019, chapter 11), and Maathuis *et al.* (2019, chapters 9 and 12). In addition, some early references also comprehensively summarised developments of many types of graphical models and their applications. For example, Koller & Friedman (2009) primarily discussed some probabilistic graphical models as well as their inferential methods and optimisation strategies, including Bayesian network, Gaussian network models and temporal models. Jordan (2004) summarised some algorithmic ideas to deal with large-scale data analysis and presented some applications to error-control coding and language processing. A textbook edited by Jordan (1999) collected several research results for Bayesian network and other structures, such as directed network with hidden variables or latent variable models. For applications of graphical models in bioinformatics, Sinoquet & Mourad (2014) summarised several topics of graphical model analysis to gene expression, genetic association studies and causality.

In recent years, complex structures and noisy data have been frequently explored in the estimation of graphical models, which are regarded as extensions of early settings with restricting assumptions removed. Several interesting topics have been included in the handbook edited by Maathuis *et al.* (2019); however, some important materials are still not well summarised. Due to the limited scope in a paper-length treatment, we focus on some important topics in the developments of graphical models, including

- 1 quantile graphical models;
- 2 non-parametric graphical models;
- 3 multiple graphical models;
- 4 multi-dimensional graphical models;
- 5 error-prone graphical models;
- 6 latent variable graphical models;
- 7 time series graphical models.

Specifically, the first 4 topics belong to complex model structures, and the last 3 topics contain challenges to noisy data.

In addition to detecting network structures of high-dimensional data that are regarded as ‘unsupervised learning’, graphical models have also been applied to supervised learning in recent years, such as high-dimensional regression models, classification and prediction, and lifetime data analysis. There are some impressive results, but to the best of knowledge, they have not been comprehensively summarised yet.

While many methods have been developed to deal with estimation of graphical models as well as their various extensions, a few of articles systematically summarise the existing approaches and the estimation of complex settings, especially for those published in recent years. Hence, in this paper, we aim to present some contemporary and important frameworks of graphical models that have not been comprehensively summarised in existing references (e.g. Jordan, 1999, 2004; Koller & Friedman, 2009; Maathuis *et al.*, 2019; Sinoquet & Mourad, 2014), including some commonly used models and estimation procedures. We first review probabilistic models based on different types of distributions to high-dimensional data, including exponential family distributed graphical models and mixed graphical models. In addition, we summarise and compare some famous strategies for the estimation of network structures, such as the graphical LASSO and conditional inference methods as well as their variants. Those comprehensive discussions were not fully explored in the past monograph (e.g. Koller & Friedman, 2009). After that, we select several important topics and introduce advanced approaches whose ideas are extended from existing estimation methods. Furthermore, we discuss applications of network structures to regression models and classification problems. Different from existing monographs, this paper provides methodological perspectives based on computational algorithms and optimisation approaches. In addition, we also add more research results that have not been included in existing monographs, so that the discussion becomes more comprehensive.

The remainder is organised as follows. In Section 2, we introduce some well-known graphical models and review some popular estimation methods in graphical model theory. In Section 3, we further introduce some selected topics of complex structures and noisy data and outline key strategies in existing methods to address those problems. In Section 4, we present some applications of graphical models in building regression models and classification. To demonstrate applications, we adopt several estimation methods to analyse two real datasets and compare the performance of existing methods in Section 5. Finally, we give a summary of this paper and discuss some future research directions in Section 6.

## 2 Basic Theory of Graphical Models

In this section, we introduce some well-known graphical models and review some fundamental estimation methods to derive graphical structures. In addition, we also briefly outline relevant developments and advanced approaches to estimate the associated parameters in graphical models. To easily understand estimations methods and strategies, we also summarise methods discussed in this section in Table 1.

Before presenting the main discussion, we first define some unified notation that will be used in the remaining of this paper. For a  $p \times p$  matrix  $A = [a_{st}]$  for  $s, t = 1, \dots, p$ , define

$\|A\|_F = \sqrt{\sum_{s=1}^p \sum_{t=1}^p a_{st}^2}$  and  $\|A\|_{\max} = \max_{s, t=1, \dots, p} |a_{st}|$  as the Frobenius norm and the maximum norm, respectively.  $A \succ 0$  indicates that  $A$  is a positive definite matrix. For a  $p$ -dimensional vector

$a = (a_1, \dots, a_p)^\top$ , let  $\|a\|_1 = \sum_{s=1}^p |a_s|$  and  $\|a\|_2 = \sqrt{\sum_{s=1}^p a_s^2}$  represent  $L_1$  and  $L_2$ -norms, respectively. Let  $\|a\|_\infty = \max_{s=1, \dots, p} |a_s|$  denote the infinity norm. Moreover, let  $\langle a, b \rangle$  denote the

Table 1. Summary of estimation methods for graphical models in Section 2. Models summarise the commonly used models in Section 2.1. Estimation methods show the strategies for estimating graphical models. References reflect the citations of methods.

Models	Estimation methods	References
Gaussian graphical models and related precision matrices	(1) GLASSO (2) Adaptive GLASSO (3) P-GLASSO and DP-GLASSO (4) G-ISTA (5) QUIC (6) G-AMA (7) CLIME (8) Robust estimation for precision matrices	Friedman <i>et al.</i> (2008), Yuan & Lin (2007) Zhou <i>et al.</i> (2009) Mazumder & Hastie (2012a) Guillot <i>et al.</i> (2012) Hsieh <i>et al.</i> (2014) Dalal & Rajaratnam (2017) Cai <i>et al.</i> (2011) Chun <i>et al.</i> (2018), Avella-Medina <i>et al.</i> (2018)
Exponential family distributed graphical models	(1) Conditional inference (2) SPACE (3) CONCORD	Ravikumar <i>et al.</i> (2010), Hastie <i>et al.</i> (2015), Yang <i>et al.</i> (2015) Peng <i>et al.</i> (2009) Khare <i>et al.</i> (2015)
Mixture graphical models	(1) Conditional inference (2) Group LASSO (3) Stable edge-specific penalty selection (4) Exponentially distributed vertices (5) Involvement of latent variables	Lee & Hastie (2015) Cheng <i>et al.</i> (2017) Sedgewick <i>et al.</i> (2016) Chen <i>et al.</i> (2015), Yang <i>et al.</i> (2014) Fan <i>et al.</i> (2017)

inner product of two vectors  $a$  and  $b$ . Given a  $p$ -dimensional vector  $a$ , let  $a_{\setminus\{s\}} = (a_1, \dots, a_{s-1}, a_{s+1}, \dots, a_p)^\top$  denote a  $(p - 1)$ -dimensional vector of  $a$  with the  $s$ -th component removed. Finally, let  $\mathbb{E}(\cdot)$  denote the expectation.

## 2.1 Some Well-Known Graphical Models

Let  $X = (X_1, \dots, X_p)^\top$  denote a  $p$ -dimensional random vector with each component  $X_j$  being a random variable with a distribution for  $j = 1, \dots, p$ . Let  $x = (x_1, \dots, x_p)^\top$  denote the realisation values of  $X$ . In graphical models, we let  $X_j$  denote a vertex, and an edge is used to link two vertices  $X_j$  and  $X_k$  for  $j \neq k$ . Let  $n$  denote the sample size in datasets. For  $i = 1, \dots, n$ , let  $X_{i,s}$  denote the  $s$ -th random variable for the  $i$ -th subject. Let  $X_{i,\bullet}$  denote the  $p$ -dimensional vector for a subject  $i$ ,  $i = 1, \dots, n$ . In the following subsections, we introduce some specific distributions of  $X$  and the corresponding graphical models.

### 2.1.1 The Gaussian Graphical Model

Suppose that the random vector  $X$  follows the Gaussian distribution with mean  $\mu$  and positive definite covariance matrix  $\Sigma$ , that is,  $X \sim N(\mu, \Sigma)$ , then its density function can be written as

$$\mathbb{P}_{\mu, \Sigma}(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}. \quad (1)$$

Let  $\gamma = \Sigma^{-1}\mu$  and  $\Theta = \Sigma^{-1}$ , then (1) can be re-parameterised as (e.g. Hastie *et al.*, 2015, p. 246)

$$\mathbb{P}_{\gamma, \Theta}(x) = \exp \left\{ \sum_{s \in V} \gamma_s x_s - \frac{1}{2} \sum_{(s, t) \in E} \theta_{st} x_s x_t + \frac{1}{2} \log \det \left( \frac{\Theta}{2\pi} \right) \right\}, \quad (2)$$

where  $\gamma_s$  is the  $s$ -th component in the vector  $\gamma$  and a  $p \times p$  matrix  $\Theta = [\theta_{st}]$  is often called the *precision matrix*. The model (2) is called the Gaussian graphical model (GGM).

### 2.1.2 The Ising Model

When the random variable is binary, that is,  $X_s \in \{-1, 1\}$  for every  $s = 1, \dots, p$ , then such graphical model is called the *Ising model* that is formulated as

$$\mathbb{P}_{\theta, \Theta}(x) = \exp \left( \sum_{r \in V} \theta_r x_r + \sum_{(s, t) \in E} \theta_{st} x_s x_t - \mathfrak{A}(\Theta) \right) \quad (3)$$

with  $\theta = (\theta_1, \dots, \theta_p)^\top$ , where  $\theta_r$  and  $\theta_{st}$  are parameters associated with  $X_r$  and  $X_s X_t$ , respectively, and  $\mathfrak{A}(\Theta)$  is called *normalising constant*, which makes (3) be integrated as one. The Ising model was first proposed in Ising's PhD thesis (Ising, 1925) and was applied in statistical mechanics (e.g. Huang, 1987, chapter 14). The other typical application of the Ising model is the social network study. According to the descriptions in Hastie *et al.* (2015), an example is the voting behaviour of politicians. By assuming that politician  $r$  provides either a 'yes' vote ( $X_r = +1$ ) or a 'no' vote ( $X_r = -1$ ),  $\theta_r > 0$  (or  $\theta_r < 0$ ) in the model (3) indicates that politician  $r$  is likely to vote 'yes' (or 'no'), and  $\theta_{st} > 0$  can be interpreted as two politician  $s$  and  $t$  are more likely to share the same vote (i.e. both yes or both no) than to disagree while  $\theta_{st} < 0$  gives the opposite interpretation.

### 2.1.3 Exponential Family Type Graphical Models

In biological studies, RNA sequencing (RNA-seq) is known as count data and the existence of network structure is ubiquitous (e.g. Grimes *et al.*, 2019). Because the RNA-seq data are usually non-normal, using Gaussian graphical models is not suitable to detect network structure of the RNA-seq data. Due to this concern, instead of specifying binary or Gaussian distributions, it is natural to consider the exponential family distribution. To see this, we follow the framework in Yang *et al.* (2015) and consider

$$\mathbb{P}_{\beta, \Theta}(x) = \exp \left\{ \sum_{r \in V} \beta_r \mathfrak{B}(x_r) + \sum_{(s, t) \in E} \theta_{st} \mathfrak{B}(x_s) \mathfrak{B}(x_t) + \sum_{r \in V} \mathfrak{C}(x_r) - \mathfrak{A}(\beta, \Theta) \right\}, \quad (4)$$

where  $\beta = (\beta_1, \dots, \beta_p)^\top$  is the  $p$ -dimensional parameter vector and  $\mathfrak{B}(\cdot)$  and  $\mathfrak{C}(\cdot)$  are given functions. The function  $\mathfrak{A}(\beta, \Theta)$  is normalising constant, or called the *log-partition function*, which makes (4) be integrated as 1. The specific form of  $\mathfrak{A}(\beta, \Theta)$  is given by

$$\mathfrak{A}(\beta, \Theta) = \log \int \exp \left\{ \sum_{r \in V} \beta_r \mathfrak{B}(x_r) + \sum_{(s, t) \in E} \theta_{st} \mathfrak{B}(x_s) \mathfrak{B}(x_t) + \sum_{r \in V} \mathfrak{C}(x_r) \right\} \mu(dx), \quad (5)$$

where  $\mu(dx)$  is the probability measure of  $X$ .

The graphical model (4) gives a broad class of models and essentially covers any distributions in the exponential family. For example, if  $\mathfrak{B}(x) = \frac{x}{\sigma}$  and  $\mathfrak{C}(x) = -\frac{x^2}{2\sigma^2}$  with known  $\sigma > 0$ , then (4) reduces to (2). If  $\mathfrak{B}(x) = x$  and  $\mathfrak{C}(x) = 0$  with  $x \in \{-1, 1\}$ , then (4) reduces to (3). Furthermore, taking  $\mathfrak{B}(x) = -x$  and  $\mathfrak{C}(x) = 0$  with  $x \in [0, \infty)$  yields the exponential graphical model

$$\mathbb{P}_{\beta, \Theta}(x) = \exp \left( - \sum_{s=1}^p \beta_s x_s + \sum_{s=1}^p \sum_{t=1}^p \theta_{st} x_s x_t + \mathfrak{A}(\beta, \Theta) \right),$$

provided that  $\theta_s > 0$  and  $\theta_{st} \geq 0$  for all  $s, t \in V$  to ensure a valid model as well as  $\mathfrak{A}(\beta, \Theta) < \infty$  (e.g. Yang *et al.*, 2015, Section 2.5). In addition, replacing  $\mathfrak{B}(x)$  and  $\mathfrak{C}(x)$  in (4), respectively, by  $x$  and  $-\log(x!)$  gives the Poisson graphical model

$$\mathbb{P}_{\beta, \Theta}(x) = \exp \left[ \sum_{s=1}^p \{ \beta_s x_s - \log(x_s!) \} + \sum_{s=1}^p \sum_{t=1}^p \theta_{st} x_s x_t + \mathfrak{A}(\beta, \Theta) \right],$$

provided that  $\theta_{st} \leq 0$  for all  $s, t \in V$  to ensure a valid model and  $\mathfrak{A}(\beta, \Theta) < \infty$ , for example, Yang *et al.* (2015, Section 2.4).

#### 2.1.4 Mixed Graphical Models

In addition to the development of graphical models based on single distributions, a more general situation is datasets containing at least two distributions in variables. To explore such phenomenon and to characterise the dependence structure, *mixed graphical models* are considered. In such models, we can explore the *homogeneity* (dependence structure of the ‘same’ distribution of variables) and the *heterogeneity* (dependence structure of variables in any two ‘different’ distributions).

In this subsection, we introduce the setting that is extended from (4). Let  $X \triangleq (Y^\top, Z^\top)^\top$ , where  $Y = (Y_1, \dots, Y_{p_y})^\top$  is a  $p_y$ -dimensional random vector in a set  $\mathcal{Y}$  and  $Z = (Z_1, \dots, Z_{p_z})^\top$  is a  $p_z$ -dimensional random vector in a set  $\mathcal{Z}$ . Suppose that  $Y_s$  and  $Z_t$  follow exponential family distributions but the distribution of  $Y_s$  is different from that of  $Z_t$ . Then exponential family mixed graphical models can be characterised as

$$\begin{aligned} \mathbb{P}_{\beta, \Theta}(y, z) \propto \exp \Big\{ & \sum_{r \in V_Y} \beta_r^Y \mathfrak{B}_Y(y_r) + \sum_{r' \in V_Z} \beta_{r'}^Z \mathfrak{B}_Z(z_{r'}) + \sum_{(s, t) \in E_Y} \theta_{st}^{YY} \mathfrak{B}_Y(y_s) \mathfrak{B}_Y(y_t) \\ & + \sum_{(s', t') \in E_Z} \theta_{s't'}^{ZZ} \mathfrak{B}_Z(y_{s'}) \mathfrak{B}_Z(y_{t'}) + \sum_{(s, t') \in E_{YZ}} \theta_{st'}^{YZ} \mathfrak{B}_Y(y_s) \mathfrak{B}_Z(y_{t'}) + \sum_{r \in V_Y} \mathfrak{C}_Y(y_r) + \sum_{r' \in V_Z} \mathfrak{C}_Z(z_{r'}) \Big\}, \end{aligned} \quad (6)$$

where  $V_Y$  and  $V_Z$  are the sets of vertices to  $Y$  and  $Z$ , respectively, such that  $V = V_Y \cup V_Z$ ,  $E_Y$  and  $E_Z$  are sets of edges to vertices in  $V_Y$  and  $V_Z$ , respectively, and  $E_{YZ}$  is the set of heterogeneous edges to vertices in  $V_Y$  and  $V_Z$ . Under the general setting, we can observe that the Gaussian–Ising mixed graphical model (e.g. Lee & Hastie, 2015; Cheng *et al.*, 2017) is a special case of (6)

by specifying  $\mathfrak{B}_Y(y) = \frac{y}{\sigma}$ ,  $\mathfrak{C}_Y(y) = -\frac{y^2}{2\sigma^2}$ ,  $\mathfrak{B}_Z(z) = z$ , and  $\mathfrak{C}_Z(z) = 0$  with  $y \in \mathbb{R}$  and  $z \in \{-1, 1\}$ . Moreover, Yang *et al.* (2014) also discussed different mixed graphical models with mixture of different types of domain and relevant restrictions on the parameter space to ensure (6) is well-defined, such as one finite domain (e.g. Poisson–Ising models) and both infinite domains (e.g. Gaussian–Poisson models).



## 2.2 Estimation Procedures

Based on several graphical models described in Section 2.1, to detect the dependence of variables and estimate the network structure, it is equivalent to study the inference on parameters  $\theta_{st}$  in  $\Theta$ . Specifically, if  $\theta_{st} = 0$ , it implies that two variables,  $X_s$  and  $X_t$ , are *conditionally independent*, given other vertices in  $V \setminus \{s, t\}$ ; otherwise,  $X_s$  and  $X_t$  are called *conditionally dependent*, given other vertices in  $V \setminus \{s, t\}$ . As a result, it suffices to do *variable selection* on  $\theta_{st}$  by retaining informative parameters and shrinking non-informative ones to zero. There are two commonly used categories in analysing  $\theta_{st}$ , and the detailed descriptions are in the following two subsections.

### 2.2.1 Graphical LASSO

The first category is the penalised likelihood function methods that directly estimate the matrix  $\Theta$ . A famous method is called *graphical LASSO* (GLASSO), which mainly focuses on the Gaussian graphical model. Without loss of generality, we let  $\gamma = 0$  in (2) and denote it by  $\mathbb{P}_\Theta(x)$ . Then under the sample with size  $n$ , the log-likelihood function based on  $\mathbb{P}_\Theta(x)$  is given by

$$\mathcal{L}(\Theta) = \sum_{i=1}^n \log \mathbb{P}_\Theta(X_{i,\bullet}) = \log\{\det(\Theta)\} - \text{trace}(\mathbf{S}\Theta), \quad (7)$$

where  $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n X_{i,\bullet} X_{i,\bullet}^\top$ ,  $\text{trace}(\cdot)$  is the sum of diagonal entries for a square matrix. The estimator of  $\Theta$ , denoted by  $\hat{\Theta}$ , is given by

$$\hat{\Theta} = \underset{\Theta \in \mathcal{A}}{\operatorname{argmax}} [\log\{\det(\Theta)\} - \text{trace}(\mathbf{S}\Theta) - \lambda\varphi(\Theta)], \quad (8)$$

where  $\mathcal{A}$  is the parametric space that contains positive definite matrices,  $\lambda$  is a tuning parameter and  $\varphi(\Theta)$  is a penalty function. Specifically, there are several choices of penalty functions, including the LASSO (Tibshirani, 1996) and adaptive LASSO (Zou, 2006) methods. In analysis of the Gaussian graphical model, the LASSO method is frequently implemented (e.g. Friedman *et al.*, 2008; Yuan & Lin, 2007; Hastie *et al.*, 2015), while Zhou *et al.* (2009) also examined the adaptive LASSO method. To solve the optimisation (8), the alternating direction method of multiplier (ADMM, Boyd *et al.*, 2011) is a useful strategy, whose key idea is to decompose the objective function into the sum of many simple convex functions. Specifically, following the idea in Boyd *et al.* (2011), the augmented Lagrangian form of (8) is written as

$$\mathcal{L}(\Theta, \Xi, \varsigma) = \text{trace}(\mathbf{S}\Theta) - \log\{\det(\Theta)\} + \lambda\varphi(\Xi) + \text{trace}\{\varsigma(\Theta - \Xi)\} + \frac{\varepsilon}{2} \|\Theta - \Xi\|_F^2$$

with the penalty parameter  $\varepsilon > 0$  and  $\varsigma$  is the dual variable or Lagrange multiplier. Then with two parameters fixed, the remaining one can be updated. That is, at the  $k$ -th step with  $k = 1, 2, \dots$ ,

$$\Theta^{(k+1)} = \underset{\Theta > 0}{\operatorname{argmin}} \left\{ \text{trace}(\mathbf{S}\Theta) - \log\{\det(\Theta)\} + \text{trace}\{\varsigma^{(k)}(\Theta - \Xi^{(k)})\} + \frac{\varepsilon}{2} \|\Theta - \Xi^{(k)}\|_F^2 \right\},$$

$$\Xi^{(k+1)} = \underset{\Xi \geq 0}{\operatorname{argmin}} \left\{ \lambda\varphi(\Xi) + \text{trace}\{\varsigma^{(k)}(\Theta^{(k+1)} - \Xi)\} + \frac{\varepsilon}{2} \|\Theta^{(k+1)} - \Xi\|_F^2 \right\},$$

and

$$\varsigma^{(k+1)} = \varsigma^{(k)} + \varepsilon \left( \Theta^{(k+1)} + \Xi^{(k+1)} \right).$$

The other strategy to deal with the optimisation problem is to transfer (8) to the simple linear equation. The pioneering idea was proposed by Friedman *et al.* (2008). Specifically, taking partial derivative on the function in (8) gives the *subgradient equation*

$$\Theta^{-1} - \mathbf{S} - \lambda \Psi = 0, \quad (9)$$

where  $\Psi$  is the symmetric matrix with diagonal elements being zero. For off-diagonal elements,  $s \neq t$ ,  $\psi_{st} = \text{sign}(\theta_{st})$  if  $\theta_{st} \neq 0$ , and  $\psi_{st} \in [-1, 1]$  if  $\theta_{st} = 0$ .

Before continuing to discuss the method, we define the *partition of matrix*. Without loss of generality, for any  $p \times p$  matrix  $\mathbf{A}$ , we fix the last component and the rule of partitioning a matrix is given by the following way:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1p} \\ a_{21} & \dots & a_{2p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pp} \end{bmatrix} \rightarrow \begin{bmatrix} a_{11} & \dots & a_{1p} \\ a_{21} & \dots & a_{2p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pp} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_{12} \\ \mathbf{a}_{12}^\top & a_{22} \end{bmatrix},$$

where  $\mathbf{A}_{11}$  is a  $(p-1) \times (p-1)$  submatrix,  $\mathbf{a}_{12}$  is a  $(p-1)$ -dimensional vector and  $a_{22}$  is a scalar.

Let  $\mathbf{W}$  denote the current working version of  $\Theta^{-1}$  such that  $\mathbf{W}\Theta = I_{p \times p}$ , where  $I_{p \times p}$  is a  $p \times p$  identity matrix. As suggested in Friedman *et al.* (2008), the working matrix  $\mathbf{W}$  is usually set as  $\mathbf{W} = \mathbf{S} + \lambda I_{p \times p}$ . According to the partition of matrix, three matrices  $\Theta$ ,  $\mathbf{S}$  and  $\mathbf{W}$  can be expressed as

$$\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^\top & s_{22} \end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^\top & w_{22} \end{bmatrix}. \quad (10)$$

Therefore, combining (9) and (10) gives a new equation

$$\mathbf{W}_{11}\beta - \mathbf{s}_{12} + \lambda\psi_{12} = 0, \quad (11)$$

where  $\beta = -\frac{\theta_{12}}{\theta_{22}}$ . The estimator  $\hat{\beta}$  can then be obtained by solving (11). Once  $\hat{\beta}$  is obtained, we have  $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$  and  $\hat{\theta}_{22} = (w_{22} - \mathbf{w}_{12}^\top \mathbf{W}_{11} \mathbf{w}_{12})^{-1}$ .

According to the definition of partition of matrix, we realise that  $(\hat{\theta}_{12}^\top, \hat{\theta}_{22})^\top = (\hat{\theta}_{1p}, \dots, \hat{\theta}_{p-1,p}, \hat{\theta}_{pp})^\top$  is the estimate of the  $p$ -th column of the matrix  $\Theta$ . As a result, repeating the same procedure by fixing the  $r$ -th row and column gives the final estimator  $\hat{\Theta}$ .

However, as pointed out by Mazumder & Hastie (2012a), the objective function in (8) is not monotone. In addition, there is a crucial concern in  $\mathbf{W}$ . Specifically, the relationship  $\mathbf{W}\Theta = I_{p \times p}$  suggests that  $\theta_{12}$  is entangled in  $\mathbf{W}_{11}$ , which is treated as a constant because  $\mathbf{W}_{11}$  is a fixed submatrix of the working matrix  $\mathbf{W}$ . Moreover,  $\mathbf{W}\Theta = I_{p \times p}$  shows that  $\mathbf{W}$  changes when  $\theta_{12}$  is updated, but the entire GLASSO algorithm only updates  $\mathbf{w}_{12}$  and  $\mathbf{w}_{21}$ .

To address those issues and remedy shortcomings, Mazumder & Hastie (2012a) proposed a 'corrected' version. By (11) and some manipulations, a new estimating equation is derived:



$$\Theta_{11}^{-1} \check{\beta} - \mathbf{s}_{12} + \lambda \psi_{12} = 0 \quad (12)$$

with  $\check{\beta} = \theta_{12} w_{22}$ . To deal with (12), we first specify  $\Theta_{11}^{-1} = \mathbf{W}_{11} - \frac{1}{w_{22}} \mathbf{w}_{12} \mathbf{w}_{21}$ . After that, the estimator  $\check{\beta}$  can be obtained by solving (12), or equivalently,

$$\check{\beta} = \underset{\check{\beta}}{\operatorname{argmax}} \left\{ \check{\beta}^\top \Theta_{11}^{-1} \check{\beta} - \check{\beta}^\top \mathbf{s}_{12} + \lambda \|\check{\beta}\|_1 \right\}. \quad (13)$$

Then  $\theta_{12}$  and  $\theta_{22}$  can be updated as  $\hat{\theta}_{12} = \frac{1}{w_{22}} \check{\beta}$  and  $\hat{\theta}_{22} = \frac{1}{w_{22}} + \hat{\theta}_{21} \Theta_{11}^{-1} \hat{\theta}_{12}$ , respectively. Finally, by  $\mathbf{W}\Theta = I_p \times p$ ,  $\mathbf{W}$  can be expressed as  $\Theta^{-1}$  and, thus, can be updated by known  $\Theta_{11}^{-1}$  and two updated values  $\hat{\theta}_{12}$  and  $\hat{\theta}_{22}$ . This ‘corrected’ approach is called P-GLASSO.

The second approach discussed in Mazumder & Hastie (2012a) is called DP-GLASSO, whose idea is first to specify  $\beta^* \triangleq \Theta_{11}^{-1} \check{\beta} - \mathbf{s}_{12}$  and then transfer the estimating equation (12) to the box-constrained quadratic programming (QP) that is given by (e.g. De Angelis *et al.*, 1997)

$$\begin{aligned} \min_{\beta^* \in \mathbb{R}^{p-1}} & \left\{ \frac{1}{2} (\mathbf{s}_{12} + \beta^*)^\top \Theta_{11} (\mathbf{s}_{12} + \beta^*) \right\} \\ \text{s.t.} & \quad \|\beta^*\|_\infty \leq \lambda. \end{aligned}$$

When the estimator of  $\beta^*$ , denoted as  $\hat{\beta}^*$ , is obtained,  $\theta_{12}$  can be updated as  $\hat{\theta}_{12} = -\frac{1}{w_{22}} \Theta_{11} (\mathbf{s}_{12} + \hat{\beta}^*)$  with  $w_{22} = s_{22} + \lambda$ , and  $\theta_{22}$  is updated as  $\hat{\theta}_{22} = \frac{1}{w_{22}} \left\{ 1 - (\mathbf{s}_{12} + \hat{\beta}^*)^\top \hat{\theta}_{12} \right\}$ .

In addition to P-GLASSO and DP-GLASSO, the other strategies were proposed to improve the computation. For example, Witten *et al.* (2011) and Mazumder & Hastie (2012b) coincidentally proposed the exact thresholding of the covariance graph. The idea is outlined as follows: suppose that the precision matrix  $\Theta$  can be expressed as block diagonal with blocks  $\mathcal{C}_1, \dots, \mathcal{C}_K$  if and only if  $|s_{ij}| < \lambda$  for all  $i \in \mathcal{C}_k$  and  $j \in \mathcal{C}_{k'}$  and  $k \neq k'$ , where  $\mathcal{C}_1, \dots, \mathcal{C}_K$  represent a partition of  $p$  vertices with  $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$  for  $k \neq k'$  and  $\mathcal{C}_1 \cup \dots \cup \mathcal{C}_K = \{1, \dots, p\}$ ,  $s_{ij}$  denotes the entry  $(i, j)$  in  $\mathbf{S}$ . Under this representation, we have  $\Theta = \operatorname{diag}(\Theta_1, \dots, \Theta_K)$ , and (8) can be employed to deal with each block matrix  $\Theta_k$  for  $k = 1, \dots, K$ .

On the other hand, there is a scenario that the  $i$ -th vertex can be fully unconnected from all other vertices if  $s_{ij} \leq \lambda$  for all  $j \neq i$ . Suppose that there are  $q$  fully unconnected vertices and  $(p - q)$  vertices are possibly connected, then  $\Theta$  is expressed as

$$\Theta = \operatorname{diag} \left( \frac{1}{s_{11} + \lambda}, \dots, \frac{1}{s_{qq} + \lambda}, \Theta_{q+1} \right),$$

where  $\Theta_{q+1}$  is a  $(p - q) \times (p - q)$  matrix containing variables that are not fully unconnected, and it can be estimated by (8); while the first  $q$  scalars,  $\frac{1}{s_{jj} + \lambda}$  for  $j = 1, \dots, q$ , are determined by (8) with  $\mathbf{S}$  replaced by  $s_{jj}$  and  $\Theta$  treated as ‘scalar’ parameters.

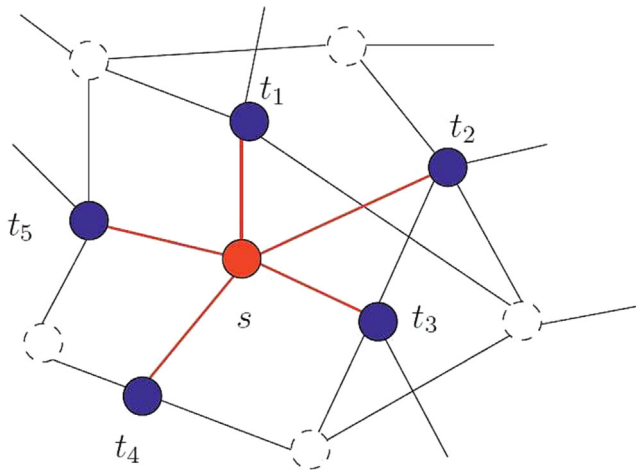
While the GLASSO method can be viewed as the pioneer work in estimating graphical structures and dealing with (8), some methods have also been developed to improve the accuracy of the estimator as well as the convergence rate. For example, Guillot *et al.* (2012) proposed the graphical iterative shrinkage thresholding algorithm (G-ISTA) and a closed form linear convergence rate was established. Hsieh *et al.* (2014) proposed a second-order proximal point algorithm (QUIC), which is shown to converge superlinearly (or quadratically) around the optimum. Dalal & Rajaratnam (2017) proposed the G-AMA method, which aims to transfer (8) to its dual problem and adopts a proximal gradient algorithm to derive the estimator of  $\Theta$ . Compared those three methods with the GLASSO method, G-ISTA exists bounds on the optimal solution to yield global convergence and G-AMA is shown to have linear convergence, while no overall complexity bounds have been established for QUIC and convergence rates seem not to be well established for the GLASSO. Furthermore, instead of imposing the  $L_1$ -norm in the regularisation methods (GLASSO, G-ISTA and QUIC), Won *et al.* (2013) considered to maximise (7) by imposing condition number of the precision matrix as the constraint. In addition to estimating the precision matrix, this approach is able to interpret the regularisation path based on the geometric perspective and then obtain the optimal value for the regularisation parameter as well as investigate the behaviour of the selected regularisation parameter.

In principal, the estimation of graphical models based on the Gaussian distribution is essentially regarded as the estimation of the precision matrix. In addition to the graphical LASSO method, several advanced methods have also been proposed to estimate the precision matrix. To name a few, Cai *et al.* (2011) proposed the constrained  $\ell_1$ -minimisation for inverse matrix estimation (CLIME) to estimate the precision matrix, which could be sparse or non-sparse, and improve the GLASSO method. Their approach provides the rate of convergence between the estimator and the true sparse precision matrix. Ravikumar *et al.* (2011) considered the general setting that the number of vertices in the graph, the number of edges and the maximum vertex degree are allowed to grow as a function of the sample size, and proposed the  $\ell_1$ -regularised log-determined method to estimate the precision matrix. Moreover, the analysis of controlling convergent rate was also examined. Avella-Medina *et al.* (2018) proposed robust matrix estimators, whose performance is guaranteed for a much richer class of distributions, and these estimators achieve the same minimax convergence rates as do existing methods under a sub-Gaussianity assumption. Chun *et al.* (2018) developed the estimation of a sparse scaled precision matrix via weighted median regression with regularisation. Their approach provides robust estimate in the presence of outliers and is consistent under various distributional assumptions including multivariate  $t$ - or contaminated Gaussian distributions.

### 2.2.2 Conditional Inference

Even though the GLASSO method is useful and has efficient computation, it is restricted in the Gaussian graphical model and is not flexible to deal with other models based on different distributions. Alternatively, the other method, called the *conditional inference* (C.I.), is able to handle graphical models based on different distributions.

The conditional inference was first proposed by Meinshausen & Bühlmann (2006), and this method is widely used in the Ising model (Ravikumar *et al.*, 2010), the Gaussian graphical model (Hastie *et al.*, 2015, Section 9.4), and exponential family graphical models (Yang *et al.*, 2015). The key idea of the conditional inference is to build up the penalised likelihood function derived by the conditional distributions of a fixed vertex, given others, because such conditional distributions also belong to the exponential family and have the same distribution as their graphical models (e.g. Yang *et al.*, 2015, p. 3818). Consequently, different from the methods in Section 2.2.1 that estimate  $p^2$  unknown parameters in the precision matrix, the



**Figure 2.** Diagram for the idea of conditional inference (Hastie et al., 2015, p. 254)

conditional inference only needs to deal with  $p - 1$  unknown parameters for each fixed vertex. To see this strategy explicitly, we only take the Gaussian graphical model as an example because the estimation method based on different models is similar.

Without loss of generality, we fix a vertex  $s \in V$ . As shown in Figure 2, the key strategy is first to derive the conditional distribution of  $X_s$  given other variables. After that, the technique of variable selection is implemented to detect non-zero parameters, so that associated variables that are dependent on  $X_s$  can be identified.

Based on the Gaussian distribution, the conditional distribution of  $X_s$  given  $X_{\setminus\{s\}} = (X_1, \dots, X_{s-1}, X_{s+1}, \dots, X_p)^\top$  is still the Gaussian distribution, and the exact form can be expressed as

$$X_s = X_{\setminus\{s\}}^\top \beta^s + \epsilon_s.$$

where  $\beta^s = (\beta_1^s, \dots, \beta_{s-1}^s, \beta_{s+1}^s, \dots, \beta_p^s)^\top$  is a  $(p - 1)$ -dimensional vector of parameters associated with vertex  $s$  and  $\epsilon_s$  is a scalar of noise term. By the penalised least squares estimation, the estimator of  $\beta^s$  is determined by

$$\hat{\beta}^s = \underset{\beta^s \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^n (X_{is} - X_{i,\setminus\{s\}}^\top \beta^s)^2 + \lambda \|\beta^s\|_1 \right\}, \quad (14)$$

where  $X_{i,\setminus\{s\}}$  is a vector  $X_{\setminus\{s\}}$  for subject  $i$ .

Let  $\mathcal{N}(s) = \{t \in V \setminus \{s\} : (s, t) \in E\}$  denote the *neighbourhood set* of  $s \in V$ , which collects variables that are dependent on  $X_s$ . Because  $\hat{\beta}^s$  is determined, then a natural estimator of  $\mathcal{N}(s)$  is given by  $\hat{\mathcal{N}}(s) = \{t \in V \setminus \{s\} : \hat{\beta}_t^s \neq 0\}$ .

In practice, if two variables  $X_s$  and  $X_t$  with  $s \neq t$  are dependent, then  $\beta_t^s$  should be equal to  $\beta_s^t$ . However, in the optimisation (14),  $\hat{\beta}_t^s$  is not necessarily equal to  $\hat{\beta}_s^t$ . To correct it, Meinshausen & Bühlmann (2006) suggested the ‘AND/OR rule’ in the sense that the final estimators  $\hat{\beta}_s^t$  and  $\hat{\beta}_t^s$  are set to be either  $\max\{\hat{\beta}_s^t, \hat{\beta}_t^s\}$  or  $\min\{\hat{\beta}_s^t, \hat{\beta}_t^s\}$  and the estimated edge set is taken as

$$\widehat{E} = \left\{ (s, t) : s \in \widehat{\mathcal{N}}(t) \text{ OR/AND } t \in \widehat{\mathcal{N}}(s) \right\}. \quad (15)$$

While (14) makes variable selection, the LASSO method usually retains too many components with small non-zero estimated regression coefficient. To make a remedy, Zhou *et al.* (2011) suggested the thresholding rule:

$$\widetilde{\beta}_j^s(\lambda, \tau) = \widehat{\beta}_j^s \times \mathbb{I}\left(|\widehat{\beta}_j^s| > \tau\right),$$

where  $\tau > 0$  is a thresholding parameter and  $\widehat{\beta}_j^s$  is the  $j$ -th component of (14). Therefore, with  $\lambda$  and  $\tau$  being determined by cross-validation, the estimated thresholding edge, denoted as  $\widehat{E}(\lambda, \tau)$ , is obtained from (15) with  $\widehat{\beta}_t^s$  and  $\widehat{\beta}_s^t$  replaced by  $\widetilde{\beta}_t^s(\lambda, \tau)$  and  $\widetilde{\beta}_s^t(\lambda, \tau)$ .

Moreover, instead of using the AND/OR rule, Zhou *et al.* (2011) adopted (8) to estimate  $\Theta$ , and the estimator is given by

$$\begin{aligned} \widehat{\Theta} &\triangleq \widehat{\Theta}\{\widehat{E}(\lambda, \tau)\} \\ &= \underset{\Theta \in \mathcal{M}}{\operatorname{argmin}} \left[ \operatorname{trace}(\Theta \check{\mathbf{S}}) - \log\{\det(\Theta)\} \right], \end{aligned}$$

where  $\check{\mathbf{S}} \triangleq \{\operatorname{diag}(\mathbf{S})\}^{-1/2} \mathbf{S} \{\operatorname{diag}(\mathbf{S})\}^{-1/2}$  is the sample correlation matrix and  $\mathcal{M} = \left\{ \Theta \in \mathbb{R}^p \times p : \Theta \succ 0 \text{ and } \theta_{st} = 0 \text{ with } s \neq t \forall (s, t) \notin \widehat{E}(\lambda, \tau) \right\}$ .

While the conditional inference is the famous approach to estimate graphical structures, it has been also extended or modified. For example, Peng *et al.* (2009) proposed the Sparse Partial Correlation Estimation (SPACE) by adopting the correlation coefficient from the precision matrix to replace  $\beta^s$  in (14). According to the findings in Peng *et al.* (2009), the advantages of the SPACE method include the reduction of the number of unknown parameters in the model and faster computation to deal with the case of  $p > n$ . Khare *et al.* (2015) developed the convex correlation selection method and algorithm (CONCORD), which is formulated as the SPACE method with its correlation coefficient term replaced by the square of correlation coefficient. In particular, the CONCORD method ensures the existence of a global minimum and the convergence of the algorithm with a fixed and finite sample size  $n$ , while those properties are not guaranteed for the SPACE method. Finally, while several advanced pseudolikelihood methods (e.g. SPACE and CONCORD) have been established, it is unknown whether those methods are able to deal with arbitrary distributions in the exponential family because only Gaussian distributions were examined in the literature. On the contrary, the conditional inference enables us to detect graphical structure for arbitrarily distributed data in exponential family, as established by Yang *et al.* (2015).

### 2.2.3 Estimation of mixed graphical models

In the literature, Gaussian–Ising graphical models are most widely used settings in the mixed graphical model, and several estimation procedures have been proposed. For example, Lee & Hastie (2015) adopted conditional inference to develop the penalised pseudo-likelihood function. Cheng *et al.* (2017) adopted the group LASSO method to address the pseudo-likelihood function and proposed stable edge-specific penalty selection to choose sparsity parameters.

In this subsection, we focus on the general setting (6) and summarise detailed discussions of the estimation procedure. Similar to other models, the goal is to estimate  $\beta_r^Y, \beta_{r'}^Z, \theta_{st}^Y, \theta_{s't'}^Z$ , and  $\theta_{s't'}^{YZ}$ .

in (6), and it can be addressed by adopting the conditional inference in Section 2.2.2. Specifically, the conditional probability of  $Y_r$  given  $Y_{\setminus\{r\}}$  and  $Z$  based on (6) is formulated as

$$\mathbb{P}_{\theta_r^Y}(y_r | y_{\setminus\{r\}}, z) = \exp \left[ \mathfrak{B}_Y(y_r) \eta(y_{\setminus\{r\}}, z; \theta_r^Y) + \mathfrak{C}_Y(y_r) - \mathfrak{D}_r \left\{ \eta(y_{\setminus\{r\}}, z; \theta_r^Y) \right\} \right], \quad (16)$$

where

$$\eta(y_{\setminus\{r\}}, z; \theta_r^Y) = \beta_r^Y + \sum_{t \in V_Y \setminus \{r\}} \theta_{rt}^{YY} \mathfrak{B}_Y(y_t) + \sum_{t' \in V_Z} \theta_{rt'}^{YZ} \mathfrak{B}_Z(z_{t'}),$$

$\mathfrak{D}_r(\cdot)$  is the normalising constant,  $\theta_r^Y = (\beta_r^Y, \theta_r^{YY \top}, \theta_r^{YZ \top})^\top$ , and  $\theta_r^{YY}$  and  $\theta_r^{YZ}$  are two vectors with components  $\theta_{rt}^Y$  and  $\theta_{rt'}^{YZ}$  for  $t \in V_Y \setminus \{r\}$  and  $t' \in V_Z$ , respectively. Then given samples with size  $n$ , the estimator of  $\theta_r^Y$  is given by

$$\hat{\theta}_r^Y = \underset{\theta_r^Y}{\operatorname{argmin}} \left\{ \ell(\theta_r^Y) + \lambda_Y \|\theta_r^{YY}\|_1 + \lambda_{YZ} \|\theta_r^{YZ}\|_1 \right\}, \quad (17)$$

where  $\ell(\theta_r^Y)$  is the log-likelihood function determined by (16) and  $\lambda_Y$  and  $\lambda_{YZ}$  are tuning parameters that could be different values.

Similarly, let  $\theta_{r'}^Z = (\beta_{r'}^Z, \theta_{r'}^{ZZ \top}, \theta_{r'}^{YZ \top})^\top$ , where  $\theta_{r'}^{ZZ}$  and  $\theta_{r'}^{YZ}$  are two vectors with components  $\theta_{r't'}^Z$  and  $\theta_{r't}^{YZ}$  for  $t' \in V_Z \setminus \{r'\}$  and  $t \in V_Y$ , respectively. The parameter  $\theta_{r'}^Z$  can be estimated by the same strategy as (17), and the estimator is given by

$$\hat{\theta}_{r'}^Z = \underset{\theta_{r'}^Z}{\operatorname{argmin}} \left\{ \ell(\theta_{r'}^Z) + \lambda_Z \|\theta_{r'}^{ZZ}\|_1 + \lambda_{ZY} \|\theta_{r'}^{YZ}\|_1 \right\}, \quad (18)$$

where  $\ell(\theta_{r'}^Z)$  is the log-likelihood function based on the conditional probability

$$\mathbb{P}_{\theta_{r'}^Z}(z_{r'} | z_{\setminus\{r'\}}, y) = \exp \left[ \mathfrak{B}_Z(z_{r'}) \eta(z_{\setminus\{r'\}}, y; \theta_{r'}^Z) + \mathfrak{C}_Z(z_{r'}) - \mathfrak{D}_{r'} \left\{ \eta(z_{\setminus\{r'\}}, y; \theta_{r'}^Z) \right\} \right],$$

where

$$\eta(z_{\setminus\{r'\}}, y; \theta_{r'}^Z) = \beta_{r'}^Z + \sum_{t' \in V_Z \setminus \{r'\}} \theta_{r't'}^{ZZ} \mathfrak{B}_Z(z_{t'}) + \sum_{t \in V_Y} \theta_{rt}^{YZ} \mathfrak{B}_Y(y_t),$$

and  $\mathfrak{D}_{r'}(\cdot)$  being the normalising constant, and  $\lambda_Z$  and  $\lambda_{ZY}$  are tuning parameters.

When (17) is obtained, we are further able to recover the homogeneous neighbourhood  $\mathcal{N}_Y(r) = \{t \in V_Y \setminus \{r\} : \theta_{rt}^Y \neq 0\}$  that contains pairwise connection of  $Y_t$  and the heterogeneous neighbourhood  $\mathcal{N}_{YZ}(r) = \{t' \in V_Z : \theta_{rt'}^{YZ} \neq 0\}$  that includes interactions with vertices  $Z_{t'}$ . Similarly, the analogous strategy can be employed to derive homogeneous and heterogeneous neighbourhood of  $Z_{r'}$  based on the result (18).

In parallel efforts, Chen *et al.* (2015) considered the similar setting, but their approach allows the graph to contain more than two types of vertices, which is different from Yang *et al.* (2014) that the graph contains only two types of vertices. Finally, Fan *et al.* (2017) also explored mixed graphical models with latent variables incorporated; detailed discussions are deferred to Section 3.6.

### 2.3 Some Available R Packages

There are many statistical packages in R software for estimations of graphical models in this section. For the estimation methods in Section 2.2.1, one can adopt the R package `glasso` to implement the GLASSO method. In addition, the R packages `QUIC`,<sup>1</sup> `dpglasso`<sup>2</sup> and `clime` can be used to estimate the precision matrix by implementing the QUIC (Hsieh *et al.*, 2014), DP-GLASSO (Mazumder & Hastie, 2012a) and CLIME (Cai *et al.*, 2011) methods, respectively. Regarding the strategies in Section 2.2.2, one can adopt the R package `XMRP`<sup>3</sup> (Wan *et al.*, 2016) to implement the conditional inference. In addition, two R packages `space`<sup>4</sup> and `gconcord`<sup>5</sup> can be used to demonstrate the SPACE and CONCORD methods, respectively. Moreover, the R packages `gRim` and `mgm` that are respectively discussed by Højsgaard *et al.* (2012, Section 5.8) and Haslbeck & Waldorp (2020) can deal with mixed graphical models.

For numerical performance among those existing methods in R packages, some comparisons have been discussed in the literature. First, Peng *et al.* (2009) compared the SPACE method with the GLASSO method and found that the SPACE method outperforms the GLASSO method because of the improvement of the power of edge detection when false discovery rate (FDR) is controlled at 0.05. Khare *et al.* (2015) compared the CONCORD method with the GLASSO and SPACE methods, and it is found that the CONCORD method has faster computation than the GLASSO and SPACE methods, especially in the ultrahigh-dimensional setting ( $p > n$ ). In addition, numerical experiments show that the CONCORD method has a much better model selection performance, including accurate recovery of the sparsity structure and less variation, compared with the GLASSO method. Finally, while the GLASSO and QUIC methods compute the same estimator, Hsieh *et al.* (2014) numerically showed that the QUIC method outperforms the GLASSO method with more accurate edge detection as well as better true positive and false positive rates.

## 3 Advanced and Complex Network Structures

In this section, we discuss the estimation methods for several advanced and complex network structures that are outlined in Section 1. Two directions are mainly focused: one is different types of structures in models, and the other is complex and noisy data. The features and the corresponding strategies are summarised in Table 2, and the detailed introduction is in the following subsections.

### 3.1 Quantile Graphical Models

In this subsection, we describe quantile graphical models associated with a  $p$ -dimensional random vector  $X$ , which is basically characterised by the conditional distribution of a fixed vertex, given others. Specifically, following the formulation in Ali *et al.* (2016), the  $\alpha$ -conditional quantile of the vertex  $s$  given other vertices is given by

$$\mathcal{Q}_{X_s|X_{\setminus\{s\}}}(\alpha) = \beta_{\alpha,s} + \sum_{t \neq s} f_{\alpha,st}(X_t), \quad (19)$$

where  $\mathcal{Q}_{X_s|X_{\setminus\{s\}}}(\alpha) \triangleq \inf\{x: P(X_s \leq x | X_{\setminus\{s\}}) \geq \alpha\}$  for all  $\alpha \in [0, 1]$ ,  $\beta_{\alpha,s} \in \mathbb{R}$  and  $f_{\alpha,st}(\cdot)$  can be a non-parametric function.

In the spirit of estimation methods of quantile regression, the estimators of  $\beta_{\alpha,s}$  and  $f_{\alpha,st}$  with a fixed vertex  $s = 1, \dots, p$  can be obtained by



Table 2. Summary of complex network structures and their estimations. Topics represent subtitles in Section 3. Key features show the main difference or extension from the setting in Section 2.1, including complex model structures or noisy data. Methods summarise key strategies to handle those features. References reflect the citations of methods.

Topics	Key features	Methods	References
Quantile graphical models	Model the conditional quantile of the vertex $X_s$ given other vertices	(1) CIQGM and PQGM (2) Bayesian approach with a spike and slab prior	Belloni <i>et al.</i> (2019) Guha <i>et al.</i> (2020)
Non-parametric graphical models	Consideration of transformed data (20) with unknown function $f(\cdot)$	(1) Estimate $f(\cdot)$ by the Winsorised estimator and adopt the GLASSO for $\hat{f}(X)$ (2) Spearman's $\rho$ (3) Kendall's $\tau$ (4) Bayesian methods for Poisson graphical models (5) Conditional inference for exponential family graphical models	Liu <i>et al.</i> (2009) Liu <i>et al.</i> (2012) Xue & Zou (2012) Roy & Dunson (2020) Yang <i>et al.</i> (2018)
Multiple graphical models	Heterogeneous data: $K$ different categories share the same variables and have $K$ different precision	(1) Optimise (25) by the local linear approximation (2) Optimise (25) by the fused graphical LASSO and the group graphical LASSO (3) Multi-task learning for a constrained minimisation problem (4) Decompose the random vectors into heterogeneous parts and shared systemic random effect, then derive the estimator (26) (5) Conditional inference and examination of structural similarity (6) FDR and estimations of structural similarity and difference (7) Layered network structures	Guo <i>et al.</i> (2011) Danaher <i>et al.</i> (2014) Lee & Liu (2015) Xie <i>et al.</i> (2016) Ma & Michailidis (2016) Liu (2017) Lin <i>et al.</i> (2016)
Multi-dimensional graphical models	Unlike $p$ -dimensional vector, multi-dimensionality reflects matrix-variate data	(1) Penalised log-likelihood method based on (35) estimates two precision matrices separately (2) Directly estimate $\Sigma \otimes \Psi$ and its inverse (3) $K$ -way tensor with $K > 2$ (4) Canonical correlation for multi-attribute data (5) Partial canonical correlation for multi-attribute data	Leng & Tang (2012) Zhou (2014) He <i>et al.</i> (2014) Katenka & Kolaczyk (2012) Kolar <i>et al.</i> (2014)
Error-prone graphical models	The random variables are collected with measurement error	(1) Bias analysis and corrected GLASSO for Gaussian graphical models (2) Bias analysis and the SIMEX method for exponentially distributed graphical models	Wainwright (2019) Chen & Yi (2022)
Latent variable graphical models	Some variables are latent in the sense that they are unobserved or not accessible	(1) Derive (53) under the normality assumption and solve it by Newton-CG-based proximal point algorithm (2) Solve (53) by ADMM (3) Solve (53) by the decomposable regularisation method	Chandrasekaran <i>et al.</i> (2012) Ma <i>et al.</i> (2013) Meng <i>et al.</i> (2014)
Time series graphical models	Dynamic models, the random variables are dependent on the time	(1) Pioneering study and partial spectral coherence (2) The VAR process (54) (3) A single time-lag of (54) and the constrained convex optimisation method	Dahlhaus (2000) Dahlhaus (2000) Han & Liu (2013)

(Continues)



Table 2 (Continued)

Topics	Key features	Methods	References
		(4) Estimation of (54) based on decomposed structured sparse matrix	Basu <i>et al.</i> (2019)
		(5) Heterogeneous VAR models	Skipnikov & Michailidis (2019)

$$\underset{\substack{\beta_{\alpha, s} \\ f_{\alpha, st} \in \mathcal{F}_{\alpha, st}}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \phi_{\alpha} \left( X_{i, s} - \beta_{\alpha, s} - \sum_{t \neq s} f_{\alpha, st}(X_t) \right) + \sum_{t \neq s} \left\{ \lambda_1 \varphi_1(f_{\alpha, st}) + \lambda_2 \varphi_2(f_{\alpha, st}) \right\}^{\varpi} \right\},$$

where  $\phi_{\alpha}(x) = \max\{\alpha x, (\alpha - 1)x\}$  is the quantile loss,  $\mathcal{F}_{\alpha, st}$  is the space of univariate functions,  $\varpi > 0$  is a fixed exponent,  $\lambda_1$  and  $\lambda_2$  are tuning parameters, and  $\varphi_1$  and  $\varphi_2$  are sparsity and smoothness penalty functions, respectively. With  $f_{\alpha, st}(\cdot)$  expressed by basis expansion model and structural constraints imposed, the alternating direction method of multipliers (ADMM) is adopted to solve the optimisation problem.

In contrast, Belloni *et al.* (2019) and Guha *et al.* (2020) also considered (19) but specified  $f_{\alpha, st}(X_t)$  as  $\theta_{\alpha, st}X_t$ . It reflects that  $X_s$  and  $X_t$  are conditionally independent if and only if  $\theta_{\alpha, st} = 0$  at the  $\alpha$ -th quantile. To estimate the network structure as well as  $\theta_{\alpha, st}$ , Belloni *et al.* (2019) proposed the conditional inference quantile graphical model (CIQGM) and the prediction quantile graphical models (PQGM), where the former method aims to minimise the moment equation based on quantile regression with constraints that ensure sparsity of the parameters, and the latter method aims to estimate  $\theta_{\alpha, st}$  that enables to predict  $X_s$  by using a linear combination of  $X_{\setminus\{s\}}$ , that is, with certain constraints for the parameters, PQGM suggests

$$\hat{\theta}_{\alpha, st} = \underset{\theta_{\alpha, st}}{\operatorname{argmin}} \mathbb{E} \left\{ \phi_{\alpha}^* \left( X_s - X_{\setminus\{s\}}^{\top} \theta_{\alpha, st} \right) \right\},$$

where  $\phi_{\alpha}^*(\cdot)$  is a given loss function. On the contrary, Guha *et al.* (2020) adopted a Bayesian variable selection technique by imposing a spike and slab prior to  $\theta_{\alpha, st}$ . After that, the variational Bayes methodology is used to approximate the posterior distribution and the MCMC method is employed to construct the final graphs.

### 3.2 Non-Parametric Graphical Models

Unlike the estimation methods in Section 2 that detect network structures for random vector  $X$ , in this subsection, we explore non-parametric graphical models by considering unknown and non-linear functions in random vector  $X$  (e.g. Lafferty *et al.*, 2012).

Parallel with the ideas behind sparse additive models for regression, we replace the random vector  $X = (X_1, \dots, X_p)^{\top}$  by the transformed random vector  $f(X) = (f_1(X_1), \dots, f_p(X_p))^{\top}$ , where  $f_1, \dots, f_p$  are unknown, monotone and differentiable functions. Based on this transformation, it is assumed that  $f(X)$  follows multivariate Gaussian

distributions, that is,  $f(X) \sim N(\mu, \Sigma)$ , then  $X$  is said to follow ‘non-paranormal’ distributions, and is denoted as  $X \sim NPN(\mu, \Sigma, f)$ .

Similar to the GLASSO method, the main purpose is to estimate  $\Theta \equiv \Sigma^{-1}$  because the entry  $\theta_{st} = 0$  if and only if  $X_s$  and  $X_t$  are conditionally independent given other vertices. In addition, a challenge is to deal with unknown function  $f(\cdot)$ .

An intuitive idea is to directly estimate  $f_j(\cdot)$  and  $\Theta$ . According to Liu *et al.* (2009),  $f_j(\cdot)$  can be expressed as

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x)), \quad (20)$$

where  $\mu_j$  is the  $j$ -th component of  $\mu$ ,  $F_j(x) = P(X_j \leq x)$  is the cumulative distribution function (CDF) of  $X_j$ ,  $\sigma_j$  is the  $j$ -th diagonal entry of  $\Sigma$ ,  $\Phi(\cdot)$  is the univariate standard Gaussian CDF. In addition, (20) can be empirically estimated by

$$\hat{f}_j(x) = \hat{\mu}_j + \hat{\sigma}_j \Phi^{-1}(\hat{F}_j(x)), \quad (21)$$

where  $\hat{F}_j(x)$  is the Winsorised estimator based on the empirical distribution of  $X_j$ ,  $\hat{\mu}_j$  and  $\hat{\sigma}_j$  are empirical estimates of  $\mu_j$  and  $\sigma_j$ , respectively. After that, (21) can be adopted to define the empirical estimate of the covariance matrix of  $f(X)$

$$S_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \left\{ \hat{f}(X_{i,\cdot}) - \mu(\hat{f}) \right\} \left\{ \hat{f}(X_{i,\cdot}) - \mu(\hat{f}) \right\}^T \quad (22)$$

with  $\mu(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \hat{f}(X_{i,\cdot})$ . Finally, following the idea of GLASSO,  $\Theta$  can be estimated by (8)

with  $\mathbf{S}$  replaced by (22).

Alternatively, instead of a two-stage procedure that estimates  $f_j$  and  $\Theta$  separately, Liu *et al.* (2012) and Xue & Zou (2012) proposed to use Spearman's  $\rho$  and Kendall's  $\tau$  to non-parametrically calculate correlation between random variables  $X_s$  and  $X_t$ . Specifically, the estimated Spearman's  $\rho$  is defined as

$$\hat{\rho}_{st} = \frac{\sum_{i=1}^n (r_s^i - \bar{r}^s)(r_t^i - \bar{r}^t)}{\sqrt{\sum_{i=1}^n (r_s^i - \bar{r}^s)^2 \sum_{i=1}^n (r_t^i - \bar{r}^t)^2}}, \quad (23)$$

where  $r_s^i$  represents the rank of  $X_{i,s}$  among  $n$  samples  $X_{1,s}, \dots, X_{n,s}$  of the  $s$ -th random variable  $X_s$  and  $\bar{r}^s = \frac{1}{n} \sum_{i=1}^n r_s^i$ . In addition, the estimated Kendall's  $\tau$  is given by

$$\hat{\tau}_{st} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}\{(X_{i,s} - X_{j,s})(X_{i,t} - X_{j,t})\}. \quad (24)$$

Let  $\mathbf{S}^*$  denote the resulting matrix whose entry  $(s, t)$  with  $s \neq t$  is based on the transformation of (23) and (24), say  $2\sin\left(\frac{\pi}{6}\hat{\rho}_{st}\right)$  and  $2\sin\left(\frac{\pi}{2}\hat{\tau}_{st}\right)$ , and the diagonal entries in  $\mathbf{S}^*$  are specified as one. Finally, one can adopt the GLASSO method with  $\mathbf{S}$  replaced by  $\mathbf{S}^*$  to estimate the precision matrix  $\Theta$ . The R package *huge* developed by Zhao *et al.* (2012) is implemented to handle non-parametric graphical models.

In addition to Gaussian distributions, another type of data considered by Roy & Dunson (2020) is count data, whose network structure can be characterised by the Poisson

graphical model in Section 2.1.3. Under the non-parametric setting, Roy & Dunson (2020) explored the following model:

$$\mathbb{P}_{\beta, \Theta}(x) \propto \exp \left[ \sum_{s=1}^p \{\beta_s x_s - \log(x_s!)\} + \sum_{s=1}^p \sum_{t=1}^p \theta_{st} f(x_s) f(x_t) \right].$$

By specifying  $f(x) = \{\tan^{-1}(x)\}^\alpha$  for some  $\alpha \in \mathbb{R}^+$ , Roy & Dunson (2020) proposed the Bayesian method by imposing prior distributions to  $\beta_s$  and  $\theta_{st}$  and implemented the Markov chain Monte Carlo (MCMC) sampling scheme to estimate the network structure.

Finally, to explore a general setting, Yang *et al.* (2018) extended exponential family graphical models by imposing unknown base measure function  $f_s$  to the conditional distribution function for  $s \in V$ , yielding

$$\mathbb{P}_{\Theta}(x_s | x_{\setminus s}) = \exp \{x_s \eta_s(x_{\setminus s}) + f_s(x_s) - b_s(\eta_s, f_s)\},$$

where  $\eta_s(x_{\setminus s}) = \sum_{t \neq s} \theta_{st} x_t$  and  $b_s(\eta_s, f_s)$  is the log-partition function. To eliminate nuisance function  $f_s$ , Yang *et al.* (2018) applied the pairwise pseudo likelihood function

$$L_s(\theta_s) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \log [1 + \exp \{-(X_{is} - X_{js}) \theta_s^\top (X_{i, \setminus \{s\}} - X_{j, \setminus \{s\}})\}],$$

where  $\theta_s = (\theta_{1s}, \dots, \theta_{(s-1)s}, \theta_{(s+1)s}, \dots, \theta_{ps})^\top$  is a  $(p-1)$ -dimensional vector of parameters. Then  $\theta_s$  can be estimated by the penalised likelihood function

$$\hat{\theta}_s = \operatorname{argmin}_{\theta_s} \left\{ L_s(\theta_s) + \sum_{t \neq s} \varphi_\lambda(|\theta_{ts}|) \right\},$$

where  $\varphi_\lambda(\cdot)$  can be convex penalty functions (e.g. LASSO) or non-convex penalty functions (e.g. SCAD).

### 3.3 Multiple Graphical Models

In applications, it is possible to collect ‘heterogeneous’ data, which reflect the same variables in several different categories. The key feature of this type of data is possibly different dependence structures among different categories. That is, some edges are common across all categories and other edges are unique to each category. A typical example is gene expression/microarray data, where subjects are classified into subgroups, and every group shares the same gene expressions. The goal is to identify graphical structures for different subgroups (e.g. Danaher *et al.*, 2014; Lee & Liu, 2015).

Suppose that a heterogeneous dataset contains  $p$  variables and  $K$  categories. For the  $k$ -th category with  $k = 1, \dots, K$ , let  $X_i^{(k)}$  denote a  $p$ -dimensional random vector for subject  $i = 1, \dots, n_k$  with sample size  $n_k$ , and it follows the multivariate normal distribution with the covariance matrix  $\Sigma^{(k)}$ . Let  $\Theta^{(k)}$  denote the  $k$ -th precision matrix in the  $k$ -th category, and define the corresponding  $(s, t)$  entry as  $\theta_{st}^{(k)}$ . The main interest is to estimate  $\Theta^{(k)}$  for all  $k = 1, \dots, K$ . To simultaneously estimate  $\Theta^{(k)}$  for all  $k = 1, \dots, K$ , the joint estimation method based on the Gaussian graphical model was developed. Specifically, motivated by (8), the penalised likelihood function based on  $K$  categories is defined as

$$\mathcal{L}(\Theta^{(1)}, \dots, \Theta^{(K)}) = \sum_{k=1}^K \left[ \log \{ \det(\Theta^{(k)}) \} - \operatorname{trace}(\mathbf{S}^{(k)} \Theta^{(k)}) - \varphi(\Theta^{(k)}) \right], \quad (25)$$

where  $\varphi(\Theta^{(k)})$  is the penalty function and  $\mathbf{S}^{(k)}$  is the empirical estimate of the covariance matrix in the  $k$ -th category. Different choices of  $\varphi(\Theta^{(k)})$  and computation of (25) were considered by different literature. For example, Guo *et al.* (2011) re-parameterised the entry  $(s, t)$  in  $\Theta^{(k)}$  by  $\theta_{st}^{(k)} = \vartheta_{st} \zeta_{st}^{(k)}$  and  $\varphi(\Theta^{(k)})$  is specified as

$$\varphi(\Theta^{(k)}) = \lambda_1 \sum_{s \neq t} \vartheta_{st} + \lambda_2 \sum_{k=1}^K \sum_{s \neq t} \left| \zeta_{st}^{(k)} \right|$$

with  $\lambda_1$  and  $\lambda_2$  being two tuning parameters. To compute (25), Guo *et al.* (2011) suggested an iterative approach based on local linear approximation. The other choices considered by Danaher *et al.* (2014) are the fused graphical LASSO

$$\varphi(\Theta^{(k)}) = \lambda_1 \sum_{k=1}^K \sum_{s \neq t} \left| \theta_{st}^{(k)} \right| + \lambda_2 \sum_{k < k'} \sum_{s \neq t} \left| \theta_{st}^{(k)} - \theta_{st}^{(k')} \right|$$

and the group graphical LASSO

$$\varphi(\Theta^{(k)}) = \lambda_1 \sum_{k=1}^K \sum_{s \neq t} \left| \theta_{st}^{(k)} \right| + \lambda_2 \sum_{s \neq t} \left( \sum_{k=1}^K \theta_{st}^{(k)^2} \right)^{1/2}.$$

An alternative direction method of multiple algorithm (ADMM) was implemented to solve (25), and the R package JGL developed by Danaher *et al.* (2014) is used to the implementation.

Another strategy for the joint estimation method is based on the multi-task learning perspective. Specifically, Lee & Liu (2015) proposed to decompose parameters into common structures

$$M = \frac{1}{K} \sum_{k=1}^K \Theta^{(k)}$$

and the unique structure

$$R^{(k)} = \Theta^{(k)} - M \quad \forall k = 1, \dots, K.$$

Then the remaining task is to estimate  $M$  and  $R^{(k)}$  for all  $k = 1, \dots, K$ . A valid strategy is to solve the following constrained minimisation problem:

$$\begin{aligned} & \min \left\{ \|M\|_1 + \lambda_1 \sum_{k=1}^K \|R^{(k)}\|_1 \right\} \\ & \text{s.t. } \left| \frac{1}{K} \sum_{k=1}^K \left\{ \mathbf{S}^{(k)} (M + R^{(k)}) - I_{p \times p} \right\} \right| \leq \eta_1, \\ & \quad \left| \mathbf{S}^{(k)} (M + R^{(k)}) - I_{p \times p} \right| \leq \eta_2, \text{ and} \\ & \quad \sum_{k=1}^K R^{(k)} = 0, \end{aligned}$$

where  $\lambda_1$  is a tuning parameter and  $\eta_1$  and  $\eta_2$  are thresholding values. Let  $\hat{M}$  and  $\hat{R}^{(k)}$  denote the resulting solutions, and thus, we have  $\hat{\Theta}^{(k)} = \hat{M} + \hat{R}^{(k)}$  with the entry  $(s, t)$  being  $\hat{\theta}_{st}^{(k)}$ . To ensure  $\hat{\Theta}^{(k)}$  as a symmetry matrix, Lee & Liu (2015) suggested ‘redefining’ the entry  $(s, t)$  as

$$\tilde{\theta}_{st}^{(k)} = \hat{\theta}_{st}^{(k)} \mathbb{I}\left(\sum_{k=1}^K |\hat{\theta}_{st}^{(k)}| \leq \sum_{k=1}^K |\hat{\theta}_{ts}^{(k)}|\right) + \hat{\theta}_{ts}^{(k)} \mathbb{I}\left(\sum_{k=1}^K |\hat{\theta}_{st}^{(k)}| > \sum_{k=1}^K |\hat{\theta}_{ts}^{(k)}|\right)$$

for all  $k = 1, \dots, K$ .

Inspired by the common and unique structures, Xie *et al.* (2016) proposed to decompose  $X_{i,\bullet}^{(k)}$  by  $X_{i,\bullet}^{(k)} = Y_{i,\bullet}^{(k)} + Z_{i,\bullet}$  for  $i = 1, \dots, n_k$  and  $k = 1, \dots, K$ , where  $Y_{i,\bullet}^{(k)}$  is the random vector corresponding to the  $k$ -th category and  $Z_{i,\bullet}$  is the random vector corresponding to the shared systemic random effect. Assume that  $Y_{i,\bullet}^{(k)}$  is independent of  $Z_{i,\bullet}$  and they follow multivariate normal distributions with mean zero and covariance matrix  $\Sigma_k$  and  $\Sigma_0$ , respectively. To explore the network structures among different categories, it is sufficient to estimate  $\Omega_k \triangleq \Sigma_k^{-1}$  based on the observed data  $X_{i,\bullet}^{(k)}$  for  $k = 1, \dots, K$ .

An intuitive approach to estimate  $\Omega_{\text{Multi}} \triangleq \{\Omega_k\}_{k=1}^K$  is the penalised likelihood estimation. The estimator of  $\Omega_k$  is given by

$$\hat{\Omega}_{\text{Multi}} = \arg \max_{\Omega_{\text{Multi}}} \left( \mathcal{L}(\Omega_{\text{Multi}}) - \lambda_1 \sum_{k=1}^K \|\Omega_k\|_1 - \lambda_2 \|\Omega_0\|_1 \right), \quad (26)$$

where  $\Omega_0 = \Sigma_0^{-1}$ ,

$$\begin{aligned} \mathcal{L}(\Omega_{\text{Multi}}) \propto & \sum_{k=1}^K [\log\{\det(\Omega_k)\} - \text{trace}(\mathbf{S}_{X,kk}\Omega_k)] + \log\{\det(\Omega_0)\} \\ & - \log\{\det(D)\} + \sum_{k,k'=1}^K \text{trace}(\Omega_k \mathbf{S}_{X,kk'} \Omega_{k'} D^{-1}) \end{aligned}$$

with  $D = \sum_{k=0}^K \Omega_k$  and  $\mathbf{S}_{X,kk'} = \frac{1}{n} \sum_{i=1}^n X_{i,\bullet}^{(k)} X_{i,\bullet}^{(k')\top}$ .

The other simpler approach is referred to the one-step method. Its idea is first to estimate  $\Sigma_0$  and  $\Sigma_k$ , respectively, by

$$\hat{\Sigma}_0 = \frac{1}{K(K-1)} \sum_{k \neq k'} \mathbf{S}_{X,kk'}$$

and

$$\hat{\Sigma}_k = \mathbf{S}_{X,kk} - \hat{\Sigma}_0.$$

After that,  $\Omega_k$  for  $k = 0, \dots, K$  can be estimated by adopting (8) with  $\Theta$  and  $\mathbf{S}$  replaced, respectively, by  $\Omega_k$  and  $\hat{\Sigma}_k$  for  $k = 0, \dots, K$ .

Unlike the likelihood-based approaches that directly estimate  $K$  precision matrices, Ma & Michailidis (2016) developed the joint structural estimation method, which basically extends the conditional inference introduced in Section 2.2.2 and examines structural similarity among  $K$  graphs. Specifically, for  $k = 1, \dots, K$  and  $i = 1, \dots, n_k$ , let  $X_{i,s}^{(k)}$  denote the  $s$ -th component in  $X_{i,\bullet}^{(k)}$  and let  $X_{i,\setminus\{s\}}^{(k)}$  denote a  $(p-1)$ -dimensional vector of  $X_{i,\bullet}^{(k)}$  with the  $s$ -th component removed. Then for the  $s$ -th vertex in the  $k$ -th category, we have

$$X_{i,s}^{(k)} = X_{i,\setminus\{s\}}^{(k)\top} \beta_s^{(k)} + \epsilon_s^{(k)}, \quad (27)$$

where  $\beta_s^{(k)} = \left( \beta_{1s}^{(k)}, \dots, \beta_{(s-1)s}^{(k)}, \beta_{(s+1)s}^{(k)}, \dots, \beta_{ps}^{(k)} \right)^\top$  is a  $(p-1)$ -dimensional vector of parameters associated with  $X_{i, \setminus \{s\}}^{(k)}$  and  $\epsilon_s^{(k)}$  is a scalar of noise term. For a given vertex  $s$ , the *joint* least squares function based on  $K$  categories is given by

$$\mathcal{L}(B_s) \triangleq \frac{1}{n} \sum_{k=1}^K \left\{ \sum_{i=1}^{n_k} \left( X_{i,s}^{(k)} - X_{i, \setminus \{s\}}^{(k)\top} \beta_s^{(k)} \right)^2 \right\},$$

where  $B_s \triangleq [\beta_s^{(1)} \dots \beta_s^{(K)}]$  is a  $(p-1) \times K$  matrix with columns indicating the regression coefficients from (27) and rows reflecting the coefficients at pairs  $(s, j)$  for  $j = 1, \dots, s-1, s+1, \dots, p$ . Noting that, for arbitrary two columns in  $B_s$ , it is possible to have the same (non)zero values for some rows. This basically says that the corresponding two categories have the same edges connecting the same vertices. Let  $\mathcal{P}_{sj}$  denote a set containing categories that have the same edges between vertices  $s$  and  $j$ . For  $1 \leq s, j \leq p$  with  $s \neq j$  and a group  $g \in \mathcal{P}_{sj}$ , let  $\beta_{sj}^{[g]} \triangleq \left( \beta_{sj}^{(k)} : k \in g \right)^\top$  denote a vector containing all coefficients from graphs in  $g$ , where  $\beta_{sj}^{(k)}$  represents the  $j$ -th component in  $\beta_s^{(k)}$ . To estimate parameters  $B_s$  with the structural similarity accommodated, Ma & Michailidis (2016) suggested the following group LASSO estimator:

$$\begin{aligned} \widehat{B}_s &\triangleq [\widehat{\beta}_s^{(1)} \dots \widehat{\beta}_s^{(K)}] \\ &= \min_{B_s} \left\{ \mathcal{L}(B_s) + 2 \sum_{j: j \neq s} \sum_{g \in \mathcal{P}_{sj}} \lambda_{sj}^g \left\| \beta_{sj}^{[g]} \right\|_2 \right\} \end{aligned} \quad (28)$$

for  $s = 1, \dots, p$ , where  $\lambda_{sj}^g$  is the associated tuning parameter. By estimates  $\widehat{\beta}_{sj}^{(k)}$  and  $\widehat{\beta}_{js}^{(k)}$  derived in (28), the resulting edge set for the  $k$ -th category is given by

$$\widehat{E}^{(k)} \triangleq \left\{ (s, j) : 1 \leq s, j \leq p, s \neq j, \widehat{\beta}_{sj}^{(k)} \neq 0 \text{ OR/AND } \widehat{\beta}_{js}^{(k)} \neq 0 \right\}$$

for  $k = 1, \dots, K$ .

Motivated by the structural similarity, the other attractive issue is the test of structural similarity and difference. Specifically, Liu (2017) aims to examine a multiple testing problem

$$H_{0st} : D_{st}(\boldsymbol{\kappa}) = 0 \text{ versus } H_{1st} : D_{st}(\boldsymbol{\kappa}) \neq 0 \quad (29)$$

for  $1 \leq s, t \leq p$  with  $s \neq t$ , where

$$D_{st}(\boldsymbol{\kappa}) = \sqrt{\sum_{1 \leq k < l \leq K} \left( \boldsymbol{\kappa}_{st}^{(k)} - \boldsymbol{\kappa}_{st}^{(l)} \right)^2}$$

with  $\boldsymbol{\kappa}_{st}^{(k)} = -\frac{\theta_{st}^{(k)}}{\sqrt{\theta_{ss}^{(k)} \theta_{tt}^{(k)}}}$  being the partial correlation coefficient of vertices  $s$  and  $t$  given other vertices. By (29), rejecting  $H_{0st}$  refers to the differential substructure

$$\mathcal{B}_D \triangleq \{(s, t) : D_{st}(\boldsymbol{\kappa}) \neq 0, 1 \leq s, t \leq p\},$$

and the set of vertex pairs with non-zero partial correlation coefficients in the complement of  $\mathcal{B}_D$ , denoted  $\mathcal{B}_D^c$ , is called similar substructure:

$$\mathcal{B}_S \triangleq \{(s, t) \in \mathcal{B}_D^c : \left( \boldsymbol{\kappa}_{st}^{(1)}, \dots, \boldsymbol{\kappa}_{st}^{(K)} \right) \neq 0\}.$$

To estimate  $\mathcal{B}_D$ , Liu (2017) proposed the false discovery rate (FDR) procedure for the multiple testing (29). Under  $H_{0st}$  in (29), the test statistic under the  $k$ -th category is defined as

$$T_{st}^{(k)} = \sqrt{\frac{1}{\hat{r}_{ss}^{(k)} \hat{r}_{tt}^{(k)}}} T_{st,0}^{(k)}, \quad (30)$$

where  $\hat{r}_{ss}^{(k)}$  and  $T_{st,0}^{(k)}$  are formulated by the residual deriving from (27); detailed formulas can be found in equation (2.3) in Liu (2017). Based on (30), the two-sample test statistic for two categories  $k, l = 1, \dots, K$  is given by

$$T_{st}^{(k,l)} = \frac{T_{st}^{(k)} - T_{st}^{(l)}}{\sqrt{\frac{1}{n_k} \left(1 - \hat{\kappa}_{st}^{(k)}\right)^2 + \frac{1}{n_l} \left(1 - \hat{\kappa}_{st}^{(l)}\right)^2}}, \quad (31)$$

where  $\hat{\kappa}_{st}^{(k)} = T_{st}^{(k)} \mathbb{I} \left\{ \left| T_{st}^{(k)} \right| \geq 2 \sqrt{\frac{\log p}{n_k}} \right\}$  is the estimator of  $\kappa_{st}^{(k)}$  and  $\mathbb{I}(\cdot)$  is an indicator function.

To perform the FDR control procedure, Liu (2017) suggests translating (31) into a  $z$ -value

$$T_{st,D} \triangleq \Phi^{-1} \left( \mathcal{T}(T_{st,*}) \right),$$

where  $T_{st,*} = \|\mathbf{T}_{st}\|_2$  with  $\mathbf{T}_{st} = \left( T_{st}^{(k,l)}, 1 \leq k < l \leq K \right)^\top$ ,  $\Phi(\cdot)$  is the CDF of the standard

normal distribution, and  $\mathcal{T}(t) = P \left( \sqrt{\sum_{i=1}^M \lambda_i Z_i^2} \leq t \right)$  with  $Z_i$  being independent and identically

distributed (i.i.d.)  $N(0, 1)$  random variable and  $\lambda_1, \dots, \lambda_M$  are eigenvalues of the asymptotic covariance matrix of  $\mathbf{T}_{st}$ . As a result, with a suitable critical value  $\hat{t}_D$  defined in equation (2.7) of Liu (2017),  $H_{0st}$  in (29) is rejected if  $T_{st,D} \geq \hat{t}_D$ , and thus, the estimated differential substructure is given by

$$\hat{\mathcal{B}}_D = \{(s, t): T_{st,D} \geq \hat{t}_D, s \neq t\}.$$

Next, the estimation of the similar substructure  $\mathcal{B}_S$  can be transformed to the following multiple testing problem:

$$H'_{0st}: \left( \kappa_{st}^{(1)}, \dots, \kappa_{st}^{(K)} \right) = 0 \text{ versus } H'_{1st}: \left( \kappa_{st}^{(1)}, \dots, \kappa_{st}^{(K)} \right) \neq 0 \quad (32)$$

with  $(s, t) \in \hat{\mathcal{B}}_D^c$ . To address the hypothesis test (32), the partial sum type test statistic is proposed:

$$T_{st,*} = \frac{\sum_{k=1}^K n_k T_{st}^{(k)}}{\sqrt{\sum_{k=1}^K n_k \left\{ 1 - \left( \hat{\kappa}_{st}^{(k)} \right)^2 \right\}^2}} \text{ for } (s, t) \in \hat{\mathcal{B}}_D^c,$$

and the corresponding transformed  $z$ -value is defined as

$$T_{st,S} \triangleq \Phi^{-1} \left( 2\Phi(|T_{st,*}|) - 1 \right).$$



With the suitable critical value  $\widehat{t}_S$  defined in equation (2.11) of Liu (2017), for  $(s, t) \in \widehat{\mathcal{B}}_D^c$ ,  $H'_{0st}$  in (32) is rejected if  $T_{st,S} \geq \widehat{t}_S$ . Therefore, the resulting estimated similar substructure is given by

$$\widehat{\mathcal{B}}_S = \{(s, t) : T_{st,S} \geq \widehat{t}_S, (s, t) \in \widehat{\mathcal{B}}_D^c, s \neq t\}.$$

Finally, we introduce layered network structures. Unlike multiple graphical models described earlier, the main difference is that layered network structures not only possess undirected edges among vertices in each layer but also exhibit a directed acyclic graph structure between the layers. In addition, the number of vertices in each layer can be different from each other. Specifically, for  $k = 1, \dots, K$ , let  $X^{(k)} = (X_1^{(k)}, \dots, X_{p_k}^{(k)})^\top$  denote the  $p_k$ -dimensional random vector in the  $k$ -th layer. Following the discussion in Lin *et al.* (2016), the first layer  $X^{(1)}$  follows a multivariate normal distribution with the covariance matrix  $\Sigma^{(1)}$ ; for the  $k$ -th layer with  $k = 2, \dots, K$ , the  $j$ -th component can be characterised by preceding layers, that is,

$$X_j^{(k)} = \sum_{l=1}^{k-1} \left\{ (D_j^{lk})^\top X^{(l)} \right\} + \epsilon_j^{(k)}$$

for  $j = 1, \dots, p_k$ , where  $\epsilon^{(k)} \triangleq (\epsilon_1^{(k)}, \dots, \epsilon_{p_k}^{(k)})^\top$  follows a multivariate normal distribution with the covariance matrix  $\Sigma^{(k)}$ , and  $D_j^{lk} \in \mathbb{R}^{p_k}$  for  $l = 1, \dots, k-1$  represents directed edges that encode the dependencies across layers. The interest of layered network study is to simultaneously estimate directed edges  $D_{p_k}^{lk}$  for  $1 \leq l < k \leq K$  among all layers as well as precision matrices  $\Theta^{(k)}$  for all layers  $k = 1, \dots, K$ . Let  $\ell(X^{(k)}; D^{lk}, \Theta^{(k)}, 1 \leq l < k \leq K)$  denote the log-likelihood function with  $D^{lk} = [D_1^{lk} \dots D_{p_k}^{lk}]$ . By the Markov factorisation, it can be decomposed as

$$\begin{aligned} & \ell(X^{(k)}; D^{lk}, \Theta^{(k)}, 1 \leq l < k \leq K) \\ &= \ell(X^{(1)}; \Theta^{(1)}) + \sum_{k=2}^K \ell(X^{(k)} | X^{(1)}, \dots, X^{(k-1)}; D^{1k}, \dots, D^{(k-1)k}, \Theta^{(k)}), \end{aligned} \quad (33)$$

which suggests that parameters in each layer can be estimated by maximising the individual likelihood function separately.

Motivated by this, Lin *et al.* (2016) particularly considered  $K = 2$  with normal distributed random vector, which yields (33) to be

$$\ell(X^{(1)}, X^{(2)}; D^{12}, \Theta^{(1)}, \Theta^{(2)}) = \ell(X^{(1)}; \Theta^{(1)}) + \ell(X^{(2)} | X^{(1)}; D^{12}, \Theta^{(2)}),$$

where  $\ell(X^{(1)}; \Theta^{(1)})$  has the same formulation as (7) and  $\Theta^{(1)}$  can be estimated by the GLASSO method, and  $\ell(X^{(2)}; D^{12}, \Theta^{(2)})$  is

$$\begin{aligned} & \ell(X^{(2)} | X^{(1)}; D^{12}, \Theta^{(2)}) \\ & \propto \frac{n}{2} \log \det \Theta^{(2)} - \frac{1}{2} \sum_{j=1}^{p_2} \sum_{i=1}^{p_2} \sigma_{ij}^{(2)} \left( X_i^{(2)} - X^{(1)} D_i^{12} \right)^\top \left( X_j^{(2)} - X^{(1)} D_j^{12} \right), \end{aligned}$$

where  $\sigma_{ij}^{(2)}$  is the entry  $(i, j)$  in  $\Theta^{(2)}$ . To simultaneous estimate sparse parameters  $D_j^{12}$  and  $\Theta^{(2)}$ , a penalised optimisation with two penalty functions is proposed:

$$\min_{D^{12}, \Theta^{(2)}} \left\{ -\ell(X^{(2)} | X^{(1)}; D^{12}, \Theta^{(2)}) + \lambda_D \sum_{j=1}^{p_2} \|D_j^{12}\|_1 + \lambda_\Theta \sum_{i \neq j} |\sigma_{ij}^{(2)}| \right\} \quad (34)$$

with two tuning parameters  $\lambda_D$  and  $\lambda_\Theta$ . The computation of minimisation (34) can be achieved by the alternating search approach, as outlined in Algorithm 1 of Lin *et al.* (2016).

### 3.4 Multi-Dimensional Graphical Models

In usual datasets, each individual has only  $p$ -dimensional vector of variables. However, in some applications, such as options contingent in financial studies or electroencephalography (EEG) in brain imaging studies, a matrix-variate data may be collected for individuals.

Specifically, we denote  $X \in \mathbb{R}^{p \times q}$  as matrix-variate data. Familiar with multivariate Gaussian distributions, the probability density function (pdf) of the matrix-variate normal distribution is

$$\begin{aligned} \mathbb{P}_{M, \Sigma, \Psi}(x) &= (2\pi)^{-\frac{qp}{2}} (\Sigma^{-1})^{q/2} (\Psi^{-1})^{p/2} \\ &\times \text{etr} \left\{ -\frac{1}{2} (x - M) \Psi^{-1} (x - M)^\top \Sigma^{-1} \right\}, \end{aligned} \quad (35)$$

where  $M \in \mathbb{R}^{p \times q}$  is the mean matrix,  $\Sigma \in \mathbb{R}^{p \times p}$  and  $\Psi \in \mathbb{R}^{q \times q}$  are row and column variance matrices,  $\text{etr}(A) \equiv \exp\{\text{trace}(A)\}$  for a matrix  $A$ . We denote (35) as  $X \sim MN_{p \times q}(M, \Sigma, \Psi)$ , or equivalently,  $\text{vec}(X) \sim N_{pq}(\text{vec}(M), \Sigma \otimes \Psi)$ , where  $\otimes$  is the Kronecker product and  $\text{vec}(M)$  represents the vectorisation of  $M$ .

While there are abundant literature in Section 2.2 to estimate graphical structure for  $\text{vec}(X)$ , they cannot be trivially adopted due to the difficulty of estimating a  $p^2 \times q^2$  precision matrix directly and ignorance of all row and column structural information. To remedy these shortcomings and estimate two matrices  $\Theta \triangleq \Sigma^{-1}$  and  $\Gamma \triangleq \Psi^{-1}$ , under the i.i.d. sample  $X_i$  with  $i = 1, \dots, n$ , Leng and Tang (2012) proposed the penalised log-likelihood function

$$(\hat{\Theta}, \hat{\Gamma}) = \underset{\Theta, \Gamma}{\text{argmin}} \{ -\ell(\Theta, \Gamma) + \phi_{\lambda_1}(\Theta) + \phi_{\lambda_2}(\Gamma) \},$$

where

$$\ell(\Theta, \Gamma) = -\frac{nq}{2} \log\{\det(\Theta)\} - \frac{np}{2} \log\{\det(\Gamma)\} + \frac{1}{2} \sum_{i=1}^n \text{trace}(X_i \Gamma X_i^\top \Theta),$$

$\phi_{\lambda_j}(\cdot)$  with  $j = 1, 2$  is the penalty function based on the LASSO or SCAD methods and  $\lambda_j$  for  $j = 1, 2$  is a tuning parameter. Alternatively, another interest is  $\Sigma \otimes \Psi$  and its inverse, which was considered by Zhou (2014). The key idea is to estimate correlation matrices of  $\Sigma$  and  $\Psi$  by adopting a pair of penalised functions, and then combine the estimators of  $\Sigma$  and  $\Psi$  to yield the desired estimator of  $\Sigma \otimes \Psi$ . Specifically,  $\Sigma \otimes \Psi$  is first expressed as

$$\Sigma \otimes \Psi = \{J_1 \varrho(\Sigma) J_1\} \otimes \{J_2 \varrho(\Psi) J_2\} \{\text{trace}(\Sigma) \text{trace}(\Psi)\},$$

where  $\varrho(\Sigma)$  and  $\varrho(\Psi)$  are correlation matrices with components  $\frac{\sigma_{st}}{\sqrt{\sigma_{ss}\sigma_{tt}}}$  and  $\frac{\psi_{st}}{\sqrt{\psi_{ss}\psi_{tt}}}$ , respectively,  $\sigma_{st}$  and  $\psi_{st}$  are the entry  $(s, t)$  of  $\Sigma$  and  $\Psi$  separately, and  $J_1$  and  $J_2$  are two matrices satisfying  $J_1 / \sqrt{\text{trace}(\Psi)} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}})$  and  $J_2 / \sqrt{\text{trace}(\Sigma)} = \text{diag}(\sqrt{\psi_{11}}, \dots, \sqrt{\psi_{qq}})$ . Under sparsity assumptions,  $\varrho(\Sigma)$  and  $\varrho(\Psi)$  can be estimated separately by

$$\widehat{\varrho(\Sigma)} = \underset{\varrho(\Sigma) > 0}{\operatorname{argmin}} \left[ \operatorname{trace}(\Sigma_{\varrho} \{\varrho(\Sigma)\}^{-1}) + \log\{\det(\varrho(\Sigma))\} + \lambda_1 \|\{\varrho(\Sigma)\}^{-1}\|_1 \right]$$

and

$$\widehat{\varrho(\Psi)} = \underset{\varrho(\Psi) > 0}{\operatorname{argmin}} \left[ \operatorname{trace}(\Psi_{\varrho} \{\varrho(\Psi)\}^{-1}) + \log\{\det(\varrho(\Psi))\} + \lambda_2 \|\{\varrho(\Psi)\}^{-1}\|_1 \right],$$

where  $\Sigma_{\varrho}$  and  $\Psi_{\varrho}$  are two sample correlation matrices with the entry  $(s, t)$  being, respectively,

$$\frac{\sum_{i=1}^n \langle X_{i, \text{col}}^{(s)}, X_{i, \text{col}}^{(t)} \rangle}{\sqrt{\sum_{i=1}^n \|X_{i, \text{col}}^{(s)}\|_2^2 \times \sum_{i=1}^n \|X_{i, \text{col}}^{(t)}\|_2^2}} \text{ and } \frac{\sum_{i=1}^n \langle X_{i, \text{row}}^{(s)}, X_{i, \text{row}}^{(t)} \rangle}{\sqrt{\sum_{i=1}^n \|X_{i, \text{row}}^{(s)}\|_2^2 \times \sum_{i=1}^n \|X_{i, \text{row}}^{(t)}\|_2^2}},$$

$X_{i, \text{col}}^{(s)}$  is the  $s$ -th column vector in  $X$  for subject  $i$ , and  $X_{i, \text{row}}^{(s)}$  is the  $s$ -th row vector in  $X$  for subject  $i$ .

In addition,  $J_1$  and  $J_2$  can be estimated, respectively, by

$$\widehat{J}_1 = \operatorname{diag} \left( \sqrt{\sum_{i=1}^n \|X_{i, \text{col}}^{(s)}\|_2^2} : s = 1, \dots, q \right) \text{ and } \widehat{J}_2 = \operatorname{diag} \left( \sqrt{\sum_{i=1}^n \|X_{i, \text{row}}^{(s)}\|_2^2} : s = 1, \dots, p \right).$$

Consequently, we have the estimator of  $\Sigma \otimes \Psi$

$$\widehat{\Sigma \otimes \Psi} = \left\{ \widehat{J}_1 \widehat{\varrho(\Sigma)} \widehat{J}_1 \right\} \otimes \left\{ \widehat{J}_2 \widehat{\varrho(\Psi)} \widehat{J}_2 \right\} \left( \frac{1}{n} \sum_{i=1}^n \|X_{i, \cdot}\|_F^2 \right),$$

and thus,  $\widehat{\Sigma \otimes \Psi}^{-1}$  is the corresponding estimator of the inverse of  $\Sigma \otimes \Psi$ .

Finally, there are some extensions of matrix-variate data. The first setting explored by He *et al.* (2014) is  $K$ -way tensor with  $K > 2$ , which treats the matrix-variate data ( $K = 2$ ) as a special case. Here, we denote  $X$  as a  $K$ -way tensor with dimension  $\{p_1, \dots, p_K\}$ , and its elements are denoted by  $\{X_{(i_1, \dots, i_K)} : i_k = 1, \dots, p_k, k = 1, \dots, K\}$ . Then the tensor normal distribution of  $X$  is denoted as  $X \sim \text{anorm}(0, \Sigma_1 \circ \dots \circ \Sigma_K)$  with the symbol ' $\circ$ ' being the outer product, and the pdf is given by

$$\mathbb{P}_{\Sigma_1, \dots, \Sigma_K}(x) = (2\pi)^{-p/2} \prod_{k=1}^K \{\det(\Sigma_k)\}^{-\frac{p}{2p_k}} \exp \left( -\frac{1}{2} \left\| x \times \Sigma^{-\frac{1}{2}} \right\|^2 \right),$$

where  $p = p_1 + \dots + p_K$ ,  $\Sigma^{-\frac{1}{2}} = \left\{ \Sigma_1^{-\frac{1}{2}}, \dots, \Sigma_K^{-\frac{1}{2}} \right\}$  with covariance matrix  $\Sigma_k$  for the  $k$ -th array and  $\|X\|^2 = \sum_{i_1, \dots, i_K} X_{(i_1, \dots, i_K)}^2$ .

Similar to other cases, the interest is to estimate  $\Theta_k \triangleq \Sigma_k^{-1}$  for  $k = 1, \dots, K$ . The penalised likelihood function is employed, and the estimator of  $\Theta \triangleq \{\Theta_1, \dots, \Theta_K\}$  given by

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left[ -\sum_{k=1}^K \frac{p}{p_k} \log\{\det(\Theta_k)\} + \operatorname{trace}\{\mathbf{S}(\Theta_K \otimes \dots \otimes \Theta_1)\} - \sum_{k=1}^K \lambda_k \sum_{s \neq t} \varphi(\theta_{k, st}) \right], \quad (36)$$

where  $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n \operatorname{vec}(X_{i, \cdot}) \operatorname{vec}(X_{i, \cdot})^\top$  and  $\otimes$  represents the Tucker product. To solve (36), the block coordinate descent algorithm can be adopted to iteratively minimise (36) with respect to  $\Theta_k$  while keeping the other matrices  $\Theta_j$  with  $j \neq k$  fixed at current values.

The second important structure in multi-dimensional graphical model is multi-attribute data, whose main feature is that the vertices reflect vectors instead of ‘scalar’ in conventional graphical models. Mathematically, for  $j = 1, \dots, p$ , let a random vector  $X_j \in \mathbb{R}^{q_j}$  follow a multivariate Gaussian distribution with mean  $\mu_j$  and covariance matrix  $\Sigma_{jj}$ . Then the multi-attribute data is defined as  $X = (X_1^\top, \dots, X_p^\top)^\top$  with mean  $\mu = (\mu_1^\top, \dots, \mu_p^\top)^\top$  and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p1} & \Sigma_{p2} & \dots & \Sigma_{pp} \end{pmatrix},$$

where  $\Sigma_{st} = \text{cov}(X_s, X_t)$ . Because  $X_j$  is a vector, to measure total association strength between multiple vertex attributes  $X_s$  and  $X_t$ , canonical correlation (e.g. Katenka & Kolaczyk, 2012) and partial canonical correlation (e.g. Kolar *et al.*, 2014) can be employed, and their formulations are given, respectively, by

$$\rho(X_s, X_t) = \max_{\substack{u \in \mathbb{R}^{q_s} \\ v \in \mathbb{R}^{q_t}}} \text{corr}(u^\top X_s, v^\top X_t) \quad (37)$$

and

$$\rho(X_s, X_t; X_{\setminus\{s,t\}}) = \max_{\substack{u \in \mathbb{R}^{q_s} \\ v \in \mathbb{R}^{q_t}}} \text{corr}\left\{u^\top (X_s - \widehat{A}X_{\setminus\{s,t\}}), v^\top (X_t - \widehat{B}X_{\setminus\{s,t\}})\right\} \quad (38)$$

with  $X_{\setminus\{s,t\}} = (X_j: j \neq s, t)$  is based on  $X$  with  $X_s$  and  $X_t$  removed,  $\widehat{A} = \underset{A}{\text{argmin}} \mathbb{E}\{\|X_s - AX_{\setminus\{s,t\}}\|_2^2\}$  and  $\widehat{B} = \underset{B}{\text{argmin}} \mathbb{E}\{\|X_t - BX_{\setminus\{s,t\}}\|_2^2\}$ . Different from (37), (38) enables to measure  $X_s$  and  $X_t$  with the effect of  $X_{\setminus\{s,t\}}$  removed, and (38) equals zero if and only if vectors  $X_s$  and  $X_t$  are conditionally independent. Moreover, it can be further shown that

$$\rho(X_s, X_t; X_{\setminus\{s,t\}}) \neq 0 \text{ if and only if } \max_{\substack{u \in \mathbb{R}^{q_s} \\ v \in \mathbb{R}^{q_t}}} u^\top \Theta_{st} v \neq 0, \quad (39)$$

where  $\Theta_{st}$  is the block entry  $(s, t)$  in  $\Theta \triangleq \Sigma^{-1}$ . It essentially says that (non)zero partial canonical correlation can be reflected by the estimated (non)zero block precision matrix.

Motivated by (39), the first strategy, which is analogue of the work proposed by Katenka & Kolaczyk (2012), is to regress  $X_s$  to other components  $X_{\setminus\{s\}} \triangleq (X_k: k \neq s)$ . That is,

$$\mathbb{E}(X_s | X_{\setminus\{s\}}) = \Sigma_{s, \setminus\{s\}} \Sigma_{\setminus\{s\}, \setminus\{s\}}^{-1} X_{\setminus\{s\}}, \quad (40)$$

where  $\Sigma_{s, \setminus\{s\}}$  is the  $s$ -th row of  $\Sigma$  with the  $s$ -th component removed,  $\Sigma_{\setminus\{s\}, \setminus\{s\}}$  is a submatrix of  $\Sigma$  with the  $s$ -th row and column deleted. Because  $\Theta_{s, \setminus\{s\}} = -\left(\Sigma_{ss} - \Sigma_{s, \setminus\{s\}} \Sigma_{\setminus\{s\}, \setminus\{s\}}^{-1} \Sigma_{\setminus\{s\}, s}\right)^{-1} \Sigma_{s, \setminus\{s\}} \Sigma_{\setminus\{s\}, \setminus\{s\}}^{-1}$ , it indicates that a zero block matrix  $\Theta_{st}$  can be identified from the regression coefficient in (40).

The other approach proposed by Kolar *et al.* (2014) is to adopt the penalised log-likelihood function

$$\widehat{\Theta} = \underset{\Theta > 0}{\text{argmin}} \left[ \text{trace}(\mathbf{S}\Theta) - \log \left\{ \det(\Theta) + \lambda \sum_{s, t} \|\Theta_{st}\|_F \right\} \right] \quad (41)$$

with  $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n X_i \cdot X_i^\top$ . By tedious derivations, the closed form of  $\widehat{\Theta}_{st}$  in (41) as well as the

estimator of  $\Sigma_{st}$  in  $\Sigma$  can be obtained. Applying an inexact block coordinate descent procedure with iteration until convergence yields the final results.

### 3.5 Error-Prone Graphical Models

Sometimes, we are unable to collect data precisely due to the measurement based on inaccurate devices. As a result, measurement error usually exists in the datasets, which reflects that the observed data are not necessarily equal to the underlying unobserved data. In applications, measurement error frequently appears in the datasets, such as cell signalling data (e.g. Bandara *et al.*, 2009; Yörük *et al.*, 2011) and gene expression data (e.g. Rocke & Durbin, 2001). In the early developments, measurement errors have been considered and established in the developments of regression models, where detailed descriptions can be found in some monographs, such as Carroll *et al.* (2006) and Yi (2017). In this section, we primarily discuss measurement error in graphical models, which is rarely explored in the literature.

Let  $X$  be the truly unobserved random vector defined in Section 2.1 and denote  $X^*$  as the observed random vector that can be regarded as the surrogate version of  $X$ . In the frameworks of measurement error, if both  $X$  and  $X^*$  are continuous, then the classical measurement error model is usually adopted to characterise  $X$  and  $X^*$ :

$$X^* = X + \epsilon, \quad (42)$$

where  $\epsilon$  is the noise term with mean zero and positive definite covariance matrix  $\Sigma_\epsilon$ ; if both  $X$  and  $X^*$  are discrete and contain binary components, then the misclassification model is adopted, which is formulated by

$$\begin{pmatrix} P(X^* = x_{(1)}) \\ \vdots \\ P(X^* = x_{(m)}) \end{pmatrix} = \mathcal{P} \begin{pmatrix} P(X = x_{(1)}) \\ \vdots \\ P(X = x_{(m)}) \end{pmatrix}, \quad (43)$$

where  $x_{(1)}, x_{(2)}, \dots, x_{(m)}$  are vectors of  $m$  possible combinations of binary variables and  $\mathcal{P}$  is the  $m \times m$  (mis)classification matrix with the component  $(k, l)$  being the (mis)classification probability, denoted as  $p_{kl} = P(X^* = x_{(k)} | X = x_{(l)})$  for  $k, l = 1, \dots, m$ . To ease notation, we let  $MC[\mathcal{P}](X)$  denote the misclassification operator indicated by (43) and notationally write (43) as  $X^* = MC[\mathcal{P}](X)$ . Furthermore,  $\mathcal{P}$  is assumed to have the spectral decomposition  $\mathcal{P} = \Lambda \mathcal{D} \Lambda^{-1}$ , where  $\mathcal{D}$  is the diagonal matrix with diagonal elements being the eigenvalues of  $\mathcal{P}$  and  $\Lambda$  is the corresponding matrix of eigenvectors.

In the presence of measurement error, under the Gaussian graphical model, Wainwright (2019, Section 11.4.1) showed that the estimator determined by (8) with  $\mathbf{S}$  replaced by the error-prone covariance matrix  $\frac{1}{n} \sum_{i=1}^n X_{i\cdot}^* X_{i\cdot}^{*\top}$  is inconsistent to the true  $\Theta$ , suggesting that measurement error would incur wrong conclusion. To address this concern and eliminate measurement error effects, a natural estimate of  $\Sigma$  is  $\mathbf{S}^* = \frac{1}{n} \sum_{i=1}^n X_{i\cdot}^* X_{i\cdot}^{*\top} -$

$\Sigma_\epsilon$ , which can be further shown that  $\|\mathbf{S}^* - \Sigma\|_{\max} \leq \sqrt{\frac{\log p}{n}}$  with high probability. Thus, solving (8) with  $\mathbf{S}$  replaced by  $\mathbf{S}^*$  gives the ‘corrected’ GLASSO estimator.

Similar idea can be applied to (14) when measurement error exists. Started by the least squares function in (14), its population-level objective function is

$$L(\beta^s) = \beta^{s\top} \Sigma_{\setminus\{s\}} \beta^s - \beta^{s\top} \Sigma_{s, \setminus\{s\}}, \quad (44)$$

where  $\Sigma_{\setminus\{s\}} = \text{var}(X_{\setminus\{s\}})$  and  $\Sigma_{s, \setminus\{s\}} = \text{cov}(X_s, X_{\setminus\{s\}})$ . Inspired by  $\mathbf{S}^*$ ,  $\Sigma_{\setminus\{s\}}$  and  $\Sigma_{s, \setminus\{s\}}$  can be estimated by

$$\widehat{\Sigma}_{\setminus\{s\}} = \frac{1}{n} \sum_{i=1}^n X_{i, \setminus\{s\}}^* X_{i, \setminus\{s\}}^{*\top} - \Sigma_{\epsilon; \setminus\{s\}} \quad (45)$$

and

$$\widehat{\Sigma}_{s, \setminus\{s\}} = \frac{1}{n} \sum_{i=1}^n X_{i, s}^* X_{i, \setminus\{s\}}^{*\top} - \Sigma_{\epsilon; s}, \quad (46)$$

where  $\Sigma_{\epsilon; \setminus\{s\}}$  is the  $(p-1) \times (p-1)$  submatrix with the  $s$ -th row/column deleted and  $\Sigma_{\epsilon; s}$  is the  $s$ -th column of  $\Sigma_{\epsilon}$ . Therefore, the ‘corrected’ LASSO estimator is given by

$$\widehat{\beta}^s = \underset{\beta^s}{\text{argmin}} \left\{ \widehat{L}(\beta^s) + \lambda \|\beta^s\|_1 \right\},$$

where  $\widehat{L}(\beta^s)$  is (44) with  $\Sigma_{\setminus\{s\}}$  and  $\Sigma_{s, \setminus\{s\}}$  replaced by (45) and (46), respectively.

To explore a general setting and provide a flexible strategy to deal with measurement error effects, Chen & Yi (2022) considered mixed graphical models (6) with  $Y$  and  $Z$  replaced by  $p_C$ -dimensional vector of continuous variables  $X_C$  and  $p_D$ -dimensional vector of discrete variables  $X_D$ , respectively. In the presence of measurement error in continuous and discrete random vector, two measurement error models (42) and (43) can be accommodated to characterise  $X_C$  and  $X_D$  as well as their surrogate versions, respectively. To correct for measurement error effects and recover the underlying true graph, the simulation-based neighbourhood-set likelihood method was proposed, whose key idea is to employ the working data generated based on (42) and (43) to eliminate measurement error effects. The estimation procedure is outlined below:

### Stage 1: Simulation

Suppose that  $X$  can be decomposed as  $X = (X_C^\top, X_D^\top)^\top$ . Let  $X_C^*$  and  $X_D^*$  denote the surrogate version of  $X_C$  and  $X_D$ , respectively. Let  $R$  be a given positive integer and let  $\mathcal{Z} = \{\zeta_0, \zeta_1, \dots, \zeta_M\}$  be a sequence of pre-specified values with  $0 = \zeta_0 < \zeta_1 < \dots < \zeta_M$ , where  $M$  is a positive integer and  $\zeta_M$  is a pre-specified positive number. In applications,  $R$  is set as a value between 100 and 500 and  $\mathcal{Z}$  is taken as a collection of  $M$  points that equally cut the interval  $[0, \zeta_M]$  with  $M$  set as 5 or 10 and  $\zeta_M$  set as 1 or 2. For  $i = 1, \dots, n$  and  $r = 1, \dots, R$ , we generate  $\epsilon_{i,r}$  from  $N(0, \Sigma_{\epsilon})$  and then define

$$W_{C,i}(r, \zeta) = X_{C,i}^* + \sqrt{\zeta} \epsilon_{i,r} \quad (47)$$

for  $\zeta \in \mathcal{Z}$ . For the discrete random vector  $X_D^*$ , we generate

$$W_{D,i}(r, \zeta) = MC[\mathcal{P}^\zeta](X_{D,i}^*) \quad (48)$$

for  $\zeta \in \mathcal{Z}$ , where  $\mathcal{P}^\zeta = \Lambda \mathcal{D}^\zeta \Lambda^{-1}$  with  $\mathcal{D}^\zeta$  representing a diagonal matrix whose diagonal entries are determined by corresponding entries in  $\mathcal{D}$  with power  $\zeta$ . Let

$W_{i,\bullet}(r, \zeta) = \left( W_{C,i}^\top(r, \zeta), W_{D,i}^\top(r, \zeta) \right)^\top$ , and we call  $W_{i,\bullet}(r, \zeta)$  the *working data*

for  $r = 1, \dots, R, \zeta \in \mathcal{Z}$ , and  $i = 1, \dots, n$ . We explain the purpose of adopting the working data. Noting that for a given  $r$ ,

$$W_{C,i}(r, \zeta) | X_{C,i} \sim N(X_{C,i}, (1 + \zeta)\Sigma_\epsilon) \text{ and } W_{D,i}(r, \zeta) = MC[\mathcal{P}^{1+\zeta}](X_{D,i})$$

for  $\zeta \in \mathcal{Z}$ , where the value of  $\zeta$  reflects the degree of mismeasurement in the working data. With  $\zeta = 0$ ,  $W_{C,i}(r, \zeta)$  and  $W_{D,i}(r, \zeta)$  recover our actually collected surrogates  $X_{C,i}^*$  and  $X_{D,i}^*$ . With a positive and increasing  $\zeta$ ,  $W_{C,i}(r, \zeta)$  and  $W_{D,i}(r, \zeta)$  incur an increasing amount of mismeasurement. When  $\zeta = -1$ ,  $W_{C,i}(r, \zeta)$  and  $W_{D,i}(r, \zeta)$  reduce to  $X_{C,i}$  and  $X_{D,i}$ , respectively, the ideal situation without mismeasurement.

#### Stage 2: Estimation

Let  $V_C$  and  $V_D$  denote vertex sets of continuous and discrete random variables, respectively. With two vertices  $s \in V_C$  and  $s' \in V_D$  fixed, replacing the unobserved variable  $X_{i,\bullet}$  in (17) and (18) by the working data  $W_{i,\bullet}(r, \zeta)$  yields two optimisers, denoted as  $\hat{\theta}_C(s; \zeta, r)$  and  $\hat{\theta}_D(s'; \zeta, r)$ , respectively, for  $\zeta \in \mathcal{Z}$ ,  $r = 1, \dots, R$ ,  $s = 1, \dots, p_C$  and  $s' = 1, \dots, p_D$ . Next, for fixed  $s, s'$  and  $\zeta \in \mathcal{Z}$ , we calculate

$$\hat{\theta}_C(s; \zeta) = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_C(s; \zeta, r) \text{ and } \hat{\theta}_D(s'; \zeta) = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_D(s'; \zeta, r). \quad (49)$$

#### Stage 3: Extrapolation

Motivated by the explanation in Stage 1, the goal is to obtain estimators corresponding to the error-free scenario (i.e.  $\zeta = -1$ ). The key strategy in this stage is to employ a regression model based on the patterns obtained from Stage 2 for different degrees of mismeasurement and then obtain the desired estimator by taking  $\zeta = -1$  as the predicted value. Specifically, grouping the estimators obtained from (49), we obtain two sequences  $\mathbb{S}_{C;s} = \left\{ \left( \zeta, \hat{\theta}_C(s; \zeta) \right) : \zeta \in \mathcal{Z} \right\}$  and  $\mathbb{S}_{D;s'} = \left\{ \left( \zeta, \hat{\theta}_D(s'; \zeta) \right) : \zeta \in \mathcal{Z} \right\}$  for  $s \in V_C$  and  $s' \in V_D$ . Then we regress  $\hat{\theta}_C(s; \zeta)$  or  $\hat{\theta}_D(s'; \zeta)$  on  $\zeta$  by fitting models

$$\hat{\theta}_C(s; \zeta) = \mathcal{G}_C(\zeta, \Gamma_C) + \delta_C \text{ and } \hat{\theta}_D(s'; \zeta) = \mathcal{G}_D(\zeta, \Gamma_D) + \delta_D \quad (50)$$

to the sequences  $\mathbb{S}_{C;s}$  and  $\mathbb{S}_{D;s'}$ , where  $\mathcal{G}_C(\cdot, \cdot)$  and  $\mathcal{G}_D(\cdot, \cdot)$  are user-specified regression functions (such as linear or quadratic functions),  $\Gamma_C$  and  $\Gamma_D$  are the associated parameter vectors, and  $\delta_C$  and  $\delta_D$  represent the noise terms. Parameters  $\Gamma_C$  and  $\Gamma_D$  can be estimated by applying the least squares method to the sequences  $\mathbb{S}_{C;s}$  and  $\mathbb{S}_{D;s'}$ ; let  $\hat{\Gamma}_C$  and  $\hat{\Gamma}_D$  denote the resulting estimates of  $\Gamma_C$  and  $\Gamma_D$ , respectively. Next, we extrapolate models (50) by letting  $\zeta = -1$  and calculate the predicted vectors

$$\hat{\theta}_C(s) = \mathcal{G}_C(-1, \hat{\Gamma}_C) \text{ and } \hat{\theta}_D(s') = \mathcal{G}_D(-1, \hat{\Gamma}_D). \quad (51)$$

Furthermore, following the discussion in Section 2.2.3, we can adopt (51) to recover homogeneous and heterogeneous neighbourhood sets of continuous and binary variables.

### 3.6 Latent Variables in Graphical Models

The latent variable is the case that variables are unobserved or not accessible. In the standard setup, suppose  $X \in \mathbb{R}^{p_O + p_H}$  is a Gaussian random vector, which can be decomposed as



$X \equiv (X_O^\top, X_H^\top)^\top$ , where  $X_O$  is the observed variable,  $X_H$  is the hidden/latent variable, and  $O$  and  $H$  are disjoint subset of indices in  $\{1, 2, \dots, p_O + p_H\}$  with  $|O| = p_O$  and  $|H| = p_H$ .

Let  $\Sigma$  denote the covariance matrix of  $X$ , which can be further decomposed to sub-block covariances matrices of  $X_O$ ,  $X_H$ , and  $(X_O, X_H)$ , denoted as  $\Sigma_O$ ,  $\Sigma_H$ ,  $\Sigma_{OH}$  and  $\Sigma_{HO} = \Sigma_{OH}^\top$ , respectively. The main interest is to estimate the observed concentration matrix

$$\Theta_O^* \triangleq \Sigma_O^{-1} = \Theta_O - \Theta_{OH} \Theta_H^{-1} \Theta_{HO}, \quad (52)$$

where  $\Theta_O$ ,  $\Theta_H$ ,  $\Theta_{OH}$  and  $\Theta_{HO}$  are sub-block matrices of  $\Sigma^{-1}$ .

Motivated by (52), rewrite  $\Theta \equiv \Theta_O$  and define

$$\Theta^* \triangleq \Theta - \mathbb{L}$$

with  $\mathbb{L}$  being assumed as low-ranked matrix. Inspired by the GLASSO method in Section 2.2.1, Chandrasekaran *et al.* (2012) proposed to estimate  $\Theta$  and  $\mathbb{L}$  by the following optimisation:

$$\min_{\substack{\Theta - \mathbb{L} \succ 0 \\ \mathbb{L} \geq 0}} [\text{trace}\{\mathbf{S}(\Theta - \mathbb{L})\} - \log\{\det(\Theta - \mathbb{L})\} + \lambda_1 \psi(\Theta) + \lambda_2 \text{trace}(\mathbb{L})], \quad (53)$$

and the Newton-CG-based proximal point algorithm can be employed to solve (53). To efficiently solve the optimisation problem, Ma *et al.* (2013) proposed the first-order ADMM and proximal gradient-based alternating direction methods by re-expressing (53) to a convex minimisation problem with two blocks of variables and two separable functions. Moreover, Meng *et al.* (2014) adopted the decomposable regularisation method to derive error bound for the precision matrix and its estimate.

To relax the normality assumption and parametric setting, Fan *et al.* (2017) further extended latent variables to mixed graphical models. Specifically,  $X$  is defined as  $X = (X_C^\top, X_D^\top)^\top$ , where  $X_C$  is a  $p_C$ -dimensional continuous vector and  $X_D$  is a  $p_D$ -dimensional discrete vector whose components are defined as  $X_{D,j} = \mathbb{I}(Z_j > C_j)$  for all  $j = 1, \dots, p_D$ , where  $C = (C_1, \dots, C_{p_D})^\top$  is a vector of constant and  $Z = (Z_1, \dots, Z_{p_D})^\top$  is a  $p_D$ -dimensional vector satisfying  $(Z, X_C) \sim \text{NPN}(0, \Sigma, f)$  that has been defined in Section 3.2 with  $\mu = 0$ . Thus, based on this structure, we refer  $X$  to follow a latent Gaussian copula model, denoted as  $X \sim \text{LNPN}(0, \Sigma, f, C)$  with  $\Sigma$  being the latent correlation matrix, because the observed binary variables are obtained by dichotomising latent variables  $Z$ . Because of unavailability of  $Z$ , to estimate  $\Sigma$  based on observed data  $X_C$  and  $X_D$ , the Kendall's  $\tau$  (24) is employed, and a suitable transformation gives the estimator of  $\Sigma$ . To further estimate  $\Theta \triangleq \Sigma^{-1}$ , the GLASSO or CLIME methods are applied with  $\Sigma$  replaced by its estimator.

Here, we give a remark to clarify the differences between measurement error in Section 3.5 and latent variables in this section. First, in measurement error models,  $X$  may not be precisely measured, but its observed version  $X^*$  can be collected; our inferences would be based on using measurement  $X^*$  with suitable adjustment to facilitate the possible differences between  $X$  and  $X^*$ . The key difficulties in the framework of measurement error are to develop a proper adjustment to fit each specific model for the response process and the measurement error process, and the likelihood-based methods are not the only approach.

A second noticeable difference lies in the interpretation and nature of the variables. Latent variables are random variables which *can never* be observed; their behaviour is mainly featured by an assumed distribution which cannot be testified. On the other hand, for the problems with measurement error, although the true variable  $X$  may not be observed for *every* subject in the study, it is possible to obtain the true value of  $X$  in situations where validation data are available.

In addition,  $X$  does not have to be always taken as a random variable, and its distribution does not have to be specified when conducting inferences (Chen & Yi, 2021b).

### 3.7 Time Series Graphical Models

In the preceding sections, we have discussed graphical models under complex settings or noisy data, but their common feature is time-independent. An attractive setting is dynamic graphical models, which incorporate the time series structure in high-dimensional data.

Let  $X(v) = (X_1(v), \dots, X_p(v))^T$  with  $v \in \mathbb{Z}$  be a multiple time series, where  $X_s(v)$  for  $s = 1, \dots, p$  are univariate real components.

We start the discussion by introducing the pioneering work of Dahlhaus (2000). Similar to preceding sections, an edge  $(s, t)$  reflects conditional dependence of  $X_s(v)$  and  $X_t(v)$  given other components. Rigorously, define  $Y_{st}(v) = (X_j(v): j \neq s, t)$  as a vector with  $X_s(v)$  and  $X_t(v)$  removed. We first remove the linear effect of  $Y_{st}(v)$  from  $X_s(v)$  and  $X_t(v)$  by minimising

$$\mathbb{E} \left\{ X_s(v) - \mu_s - \sum_{u=-\infty}^{\infty} d_s(v-u) Y_{st}(u) \right\}^2 \text{ and } \mathbb{E} \left\{ X_t(v) - \mu_t - \sum_{u=-\infty}^{\infty} d_t(v-u) Y_{st}(u) \right\}^2$$

with respect to  $\mu_s, \mu_t$ , and filters  $d_s(u)$  and  $d_t(u)$ . Let  $\hat{\mu}_s$  and  $\hat{\mu}_t$  denote the resulting optimisers, and the ‘residuals’ are denoted as

$$\mathcal{E}_s(v) \triangleq \mathcal{E}_{s|\{s,t\}^c}(v) \triangleq X_s(v) - \hat{\mu}_s - \sum_{u=-\infty}^{\infty} \hat{d}_s(v-u) Y_{st}(u)$$

and

$$\mathcal{E}_t(v) \triangleq \mathcal{E}_{t|\{s,t\}^c}(v) \triangleq X_t(v) - \hat{\mu}_t - \sum_{u=-\infty}^{\infty} \hat{d}_t(v-u) Y_{st}(u).$$

Let  $\mathcal{X}_s \triangleq (X_s(v): v \in \mathbb{Z})$  and  $\mathcal{Y}_{st} \triangleq (Y_{st}(v): v \in \mathbb{Z})$ . Then  $\mathcal{X}_s$  is independent of  $\mathcal{X}_t$  given  $\mathcal{Y}_{st}$  if and only if  $\text{cov}\{\mathcal{E}_s(v), \mathcal{E}_t(v+u)\} = 0$  for all  $u \in \mathbb{Z}$ . This approach is called partial correlation graph.

Another characterisation of the edges in the graph can be obtained from the partial spectral coherence. Let the cross-spectrum of  $X_s(v)$  and  $X_t(v)$  be defined as

$$f_{X_s X_t}(z) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} C_{st}(u) \exp(-\mathbf{i}zu),$$

where  $\mathbf{i}$  is the imaginary unit with  $\mathbf{i}^2 = -1$  and  $C_{st}(u) = \text{cov}\{X_s(v+u), X_t(v)\}$  is the covariance function of the process with  $\sum_{u=-\infty}^{\infty} |C_{st}(u)| < \infty$ . Under those definitions, a measure of the

dependence between  $X_s(v)$  and  $X_t(v)$  given  $Y_{st}(v)$  is given by  $f_{X_s X_t | Y_{st}}(z) \triangleq f_{\mathcal{E}_s \mathcal{E}_t}(z)$ . Rescaling it leads to the partial spectral coherence:

$$\mathcal{R}_{X_s X_t | Y_{st}}(z) \triangleq \frac{f_{X_s X_t | Y_{st}}(z)}{\left\{ f_{X_s X_s | Y_{st}}(z) f_{X_t X_t | Y_{st}}(z) \right\}^{1/2}}.$$

We have that  $\mathcal{R}_{X_s X_t | Y_{st}}(\cdot) \neq 0$  if and only if  $(s, t) \in E$ .

A general class of multivariate autoregressive processes is also explored in graphical models. Following the concept in time series analysis, the vector autoregressive (VAR) process is

$$X(v) = \sum_{j=1}^q \Gamma_j X(v-j) + U(v), \quad (54)$$

where  $\Gamma_j$  is the  $p \times p$  matrix for  $j = 1, \dots, q$  and  $U(v) \sim N(0, \Sigma)$ . Let  $\Gamma(z) = I_{p \times p} - \sum_{j=1}^q \Gamma_j z^j$ .

If  $\det\{\Gamma(z)\} \neq 0$  for all  $z \in \mathbb{C}$  with  $|z| \leq 1$ , then the recursion  $\Gamma(z)$  has a stationary solution (e.g. Dahlhaus, 2000; Wilson *et al.*, 2016, p. 31). Moreover, when component  $(k, l)$  in  $\Gamma_j$ , denoted  $\Gamma_j(k, l)$ , for  $j = 1, \dots, q$  is significantly larger than 0, then we say  $X_k$  is Granger-causal for  $X_l$ , which indicates that  $X_l$  can be predicted efficiently if the information in the  $X_k$  process is taken into account (e.g. Lütkepohl, 2005, pp. 42 and 44).

The main interest is to examine the component in  $\Gamma_j$ , denoted as  $\Gamma_{j, st}$  for  $s \neq t$ , as it is regarded as the ‘influence’ from  $X_t(v - j)$  on  $X_s(v)$ . In other words, there is no influence from component  $t$  on  $s$  if the entry  $(s, t)$  in  $\Gamma(\cdot)$  is equal to zero, that is,  $\Gamma_{st}(\cdot) \equiv 0$ . A more detailed justification to explain this phenomenon can be referred to Dahlhaus (2000, Section 4).

To deal with this problem, some methods have been developed. To name a few, Han & Liu (2013) and Basu *et al.* (2019) considered a single time-lag

$$X(v) = \Gamma X(v - 1) + U(v)$$

that is a special case of (54) with  $q = 1$ . Regarding the methodologies, Han & Liu (2013) proposed the constrained convex optimisation problem

$$\begin{aligned} \min_{\Gamma \in \mathbb{R}^{p \times p}} \quad & \sum_{s, t} |\Gamma_{st}| \\ \text{s.t.} \quad & \|S\Gamma - S_1\|_{\max} \leq \lambda_0, \end{aligned}$$

where  $\lambda_0 > 0$  is a tuning parameter,  $S = \frac{1}{T} \sum_{v=1}^T X(v)X(v)^\top$ , and  $S_1 = \frac{1}{T-1} \sum_{v=1}^{T-1} X(v)X(v+1)^\top$ . The other approach proposed by Dahlhaus (2000) is to decompose  $\Gamma$  as  $\Gamma = \Gamma_L + \Gamma_R$ , where  $\Gamma_L$  is a low-rank matrix and  $\Gamma_R$  is a structured sparse matrix. To estimate  $\Gamma$ , it suffices to solve the following minimisation problem:

$$(\hat{\Gamma}_L, \hat{\Gamma}_R) = \underset{\substack{\Gamma_L, \Gamma_R \\ \Gamma_L \in \mathcal{S}}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbb{Y} - \mathbb{X}(\Gamma_L + \Gamma_R)\|_F^2 + \lambda_L \|\Gamma_L\|_* + \lambda_R \|\Gamma_R\|_1 \right\},$$

where  $\mathbb{Y} = (X(T), \dots, X(1))^\top$ ,  $\mathbb{X} = (X(T-1), \dots, X(0))^\top$ ,  $\mathcal{S} = \left\{ \Gamma_L \in \mathbb{R}^{p \times p} : \|\Gamma_L\|_{\max} \leq \frac{\kappa}{p} \right\}$  with  $\kappa$  being the parameter to control the degree of non-identifiability of the matrices allowed in the model class,  $\|\cdot\|_*$  is the nuclear norm that is the sum of the singular values of a matrix, and  $\lambda_L$  and  $\lambda_R$  are tuning parameters.

Furthermore, Skripnikov & Michailidis (2019) explored (54) with consideration of various groups of subjects. A new VAR model is formulated by

$$X^{(k)}(v) = \sum_{j=1}^q \Gamma_j^{(k)} X^{(k)}(v - j) + U^{(k)}(v),$$

for  $k = 1, \dots, K$ , where  $U^{(k)}(v) \sim N(0, \sigma_{(k)}^2 I_{p \times p})$  with a variance  $\sigma_{(k)}^2$ . To characterise shared structure across all  $K$  subjects and account for the presence of heterogeneity from subject-specific effects, we decompose  $\Gamma_j^{(k)}$  as  $\Gamma_j^{(k)} = \Gamma_{j, C}^{(k)} + \Gamma_{j, L}^{(k)}$ , where  $\Gamma_{j, C}^{(k)}$  is the common component of order  $p$  temporal effects for the  $k$ -th subject and  $\Gamma_{j, L}^{(k)}$  is the idiosyncratic component. To estimate  $\Gamma_{j, C}^{(k)}$  and  $\Gamma_{j, L}^{(k)}$ , an intuitive approach is to separately estimate each row of  $\Gamma_{j, C}^{(k)}$  and  $\Gamma_{j, L}^{(k)}$ . Specifically, for a fixed  $k$ , the  $i$ -th component in  $X^{(k)}(v)$  for all  $v$  can be expressed in the following form:

$$\tilde{X}_i^{(k)} = W^{(k)} \left( \Gamma_C^{(k)}[i, \cdot] + \Gamma_L^{(k)}[i, \cdot] \right) + \tilde{U}_i^{(k)},$$

where  $\tilde{X}_i^{(k)} = \left( X_i^{(k)}(0), \dots, X_i^{(k)}(T) \right)^\top$ ,  $\tilde{U}_i^{(k)} = \left( U_i^{(k)}(0), \dots, U_i^{(k)}(T) \right)^\top$ ,  $W^{(k)}$  is a matrix with components  $X_i^{(k)}(v - u)$  for  $u = 1, \dots, v$  and  $v = q, \dots, T$ ,  $\Gamma_C^{(k)}[i, \cdot] = \left( \Gamma_{1,C}^{(k)\top}[i, \cdot], \dots, \Gamma_{p,C}^{(k)\top}[i, \cdot] \right)^\top$  and  $\Gamma_{j,C}^{(k)}[i, \cdot]$  is a vector with component  $\Gamma_{j,C}^{(k)}[i, l]$  that is the entry  $(i, l)$  in  $\Gamma_{j,C}^{(k)}$  for  $l = 1, \dots, p$  and  $j = 1, \dots, p$ , and similar definition to  $\Gamma_L^{(k)}[i, \cdot]$ . Furthermore, let  $\mathbb{W}$  denote a block diagonal matrix with the  $k$ -th block being  $W^{(k)}$ , and define  $\tilde{\mathbb{X}}_i^{(k)} = \left( \tilde{X}_i^{(1)\top}, \dots, \tilde{X}_i^{(K)\top} \right)^\top$ ,  $\mathbb{D} = \text{diag}(\sigma_{(1)}^2, \dots, \sigma_{(1)}^2, \dots, \sigma_{(K)}^2, \dots, \sigma_{(K)}^2)$ ,  $\tilde{\Gamma}_C[i, \cdot] = \left( \Gamma_C^{(1)\top}[i, \cdot], \dots, \Gamma_C^{(K)\top}[i, \cdot] \right)^\top$ ,  $\tilde{\Gamma}_L[i, \cdot] = \left( \Gamma_L^{(1)\top}[i, \cdot], \dots, \Gamma_L^{(K)\top}[i, \cdot] \right)^\top$ . Therefore, to estimate  $\Gamma_C^{(k)}[i, \cdot]$  and  $\Gamma_L^{(k)}[i, \cdot]$  for  $k = 1, \dots, K$ , we can adopt the penalised weighted least squares method

$$\begin{aligned} \min_{\tilde{\Gamma}_C[i, \cdot], \tilde{\Gamma}_L[i, \cdot]} & \left\| \mathbb{D}^{-1/2} \left\{ \tilde{\mathbb{X}}_i - \mathbb{W}(\tilde{\Gamma}_C[i, \cdot] + \tilde{\Gamma}_L[i, \cdot]) \right\} \right\|_2^2 + \lambda_{i,L} \|\tilde{\Gamma}_L[i, \cdot]\|_1 \\ & + \lambda_{i,C} \sum_{l=1}^p \left\| \left( \Gamma_C^{(1)}[i, l], \dots, \Gamma_C^{(K)}[i, l] \right)^\top \right\|_2 \\ & + \lambda^* \sum_{l=1}^p \sum_{k=1}^K \left| \Gamma_C^{(k)}[i, l] \times \Gamma_L^{(k)}[i, l] \right|, \end{aligned}$$

where  $\lambda_{i,L}$ ,  $\lambda_{i,C}$ , and  $\lambda^*$  are tuning parameters. Here, the first penalty is the well-known sparse LASSO penalty that aims to detect non-zero elements for each  $\Gamma_L^{(k)}[i, \cdot]$  with  $k = 1, \dots, K$  and  $i = 1, \dots, p$ . The second penalty function is referred to the group LASSO penalty that either shrinks the  $l$ -th element to zero for all  $K$  vectors or estimates it to be non-zero for all  $K$  vectors. Finally, in the last penalty function, a tuning parameter value  $\lambda^*$  is set high enough, so that the intersection of the supports for  $\Gamma_C^{(k)}[i, \cdot]$  and  $\Gamma_L^{(k)}[i, \cdot]$  is empty. To solve this optimisation problem, a two-stage algorithm performing an alternate convex search method is adopted.

## 4 Network Structures With Regression Models

In Sections 2 and 3, we introduce the estimation methods for graphical models that focus on characterising the (pairwise) dependence structure of variables. When we build regression models, it is crucial to incorporate the network structures since the multivariate covariates and/or responses may not be independent in most situations. Therefore, in this section, we introduce regression models with network structures accommodated. A brief summary is available in Table 3.

### 4.1 Linear Models

In this subsection, we discuss the multivariate linear model with the network structure accommodated in the response. The motivated example of this study comes from the glioblastoma multiforme (GBM) cancer dataset (Lee & Liu, 2012; Wang, 2015). This dataset contains 534 microRNA expression values and 11861 gene expression values. The sample size in this dataset

is 202. The main interest in this dataset is to regress the microRNA expressions on the gene expressions by linear models and explore (a) the relationship between the responses (microRNA) and covariates (gene expression) and (b) the network structure of microRNA based on fitted regression models.

Let  $n$ ,  $m$  and  $p$  denote the numbers of subjects, responses and covariates/parameters, respectively. A multivariate linear model is formulated by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}, \quad (55)$$

where  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top$  is a  $n \times m$  response matrix with  $m$ -dimensional vectors of responses  $\mathbf{y}_i = (y_{i1}, \dots, y_{im})^\top$  for  $i = 1, \dots, n$ ,  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top$  is a  $n \times p$  design matrix with  $p$ -dimensional vectors of covariates  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$  for  $i = 1, \dots, n$ ,  $\mathbf{e} = [\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n]^\top$  is a  $n \times m$  error matrix with  $m$ -dimensional vectors of errors  $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im})^\top$  for  $i = 1, \dots, n$ , and  $\mathbf{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m]^\top$  is a  $p \times m$  parameter matrix with  $p$ -dimensional vectors of parameters  $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{ip})^\top$  for  $i = 1, \dots, m$ .

We further assume that  $\mathbf{X}$  is fixed effect and  $\boldsymbol{\epsilon}_i$  is i.i.d. and follows the Gaussian distribution  $N(0, \Sigma)$ , where  $\Sigma = [\sigma_{st}]_{s,t=1}^m$  is assumed to be positive definite. Let  $\Theta = [\theta_{st}]_{s,t=1}^m \triangleq \Sigma^{-1}$ . The main target is to estimate  $\mathbf{B}$  and  $\Theta$ , where the estimator of  $\mathbf{B}$  gives the similar interpretation in conventional regression models and the estimator of  $\Theta$  reflects the network structure of the response  $\mathbf{Y}$ .

We primarily introduce two methods to deal with this problem. The first method is proposed by Lee & Liu (2012). The key idea of this method is based on the graphical LASSO method. Specifically, based on (55), we have  $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{XB}, \Sigma)$ . Similar to the optimisation in (8), estimators of  $\mathbf{B}$  and  $\Theta$  is given by

$$(\hat{\mathbf{B}}, \hat{\Theta}) = \underset{\mathbf{B}, \Theta}{\operatorname{argmin}} \left[ -n \log \det(\Theta) + \operatorname{trace} \left\{ (\mathbf{Y} - \mathbf{XB})\Theta(\mathbf{Y} - \mathbf{XB})^\top \right\} + \lambda_1 \sum_{j,k} w_{jk} |\beta_{jk}| + \lambda_2 \sum_{s \neq t} v_{st} |\theta_{st}| \right],$$

where  $\lambda_1$  and  $\lambda_2$  are two tuning parameters and  $w_{jk}$  and  $v_{st}$  are weights for the adaptive LASSO. In the computational perspective, to derive  $\hat{\mathbf{B}}$  and  $\hat{\Theta}$  in numerics, Lee & Liu (2012) provide the following computational algorithm:

Step 1: Set the separate LASSO solutions  $\beta_{jk}^{(old)}$  with  $j = 1, \dots, p$  and  $k = 1, \dots, m$ , and  $\Theta^{(old)}$ .

Table 3. Summary of supervised learning methods with network structures accommodated. Topics summarise the commonly used models or data structures in Section 4. Estimation methods show the strategies for constructing models. References reflect the citations of methods.

Topics	Estimation methods	References
Multivariate linear models	(1) The graphical LASSO (2) Conditional inference	Lee & Liu (2012) Wang (2015)
Multi-class classification	(1) Logistic regression with homogeneous or class-dependent graphically structured covariates accommodated (2) SVM with network based surrogate covariate (3) Network based linear/quadratic discriminant analysis	Chen <i>et al.</i> (2019) He <i>et al.</i> (2019) Chen (2022a, 2022c)
Survival analysis	(1) Variable selection and SIMEX methods for modelling the Cox PH model with network based and error-prone covariates	Chen & Yi (2021a)

Step 2: Given  $\Theta^{(old)}$ , the updated value  $\mathbf{B}^{(new)}$  is given by

$$\mathbf{B}^{(new)} = \underset{\mathbf{B}}{\operatorname{argmin}} \left[ \operatorname{trace} \left\{ (\mathbf{Y} - \mathbf{B}\mathbf{X})\Theta^{(old)}(\mathbf{Y} - \mathbf{B}\mathbf{X})^\top \right\} + \lambda_1 \sum_{j,k} w_{jk} |\beta_{jk}| \right].$$

Step 3: Given  $\mathbf{B}^{(new)}$ , the updated value  $\Theta^{(new)}$  is given by

$$\begin{aligned} \Theta^{(new)} = \underset{\Theta}{\operatorname{argmin}} & \left[ \operatorname{trace} \left\{ (\mathbf{Y} - \mathbf{B}^{(new)}\mathbf{X})^\top (\mathbf{Y} - \mathbf{B}^{(new)}\mathbf{X})\Theta \right\} \right. \\ & \left. - \log \det(\Theta) + \frac{\lambda_2}{n} \sum_{s \neq t} v_{st} |\theta_{st}| \right]. \end{aligned}$$

Step 4: Continue Steps 2 and 3 until convergence.

The second method, proposed by Wang (2015), borrows the idea of the conditional inference and extends it to the multivariate linear model. To see this, we fix  $k$  without loss of generality. Let  $\mathbf{y}^k = (y_{1k}, \dots, y_{nk})^\top$  be the  $k$ -th response and let  $\mathbf{Y}^{-k}$  denote the response matrix of  $\mathbf{Y}$  without  $\mathbf{y}^k$ . Define  $\mathbf{B}_{-k}$  as the parameter matrix of  $\mathbf{B}$  with  $\beta_k$  removed. For any matrix  $\mathbf{A}$ , let  $\mathbf{A}_{-k, k}$  be the  $k$ -th column of  $\mathbf{A}$  with  $k$ -th component deleted, and let  $\mathbf{A}_{-k, -k}$  denote a sub-matrix of  $\mathbf{A}$  with  $k$ -th row and column deleted.

By some derivations, the conditional distribution of  $\mathbf{y}^k | (\mathbf{X}, \mathbf{Y}^{-k})$  is given by

$$\mathbf{y}^k | (\mathbf{X}, \mathbf{Y}^{-k}) \sim N(\mathbf{X}\beta_k + (\mathbf{Y}^{-k} - \mathbf{X}\mathbf{B}_{-k})\gamma_k, \tilde{\sigma}_{kk}I_n), \quad (56)$$

where  $\tilde{\sigma}_{kk} = \sigma_{kk} - \Sigma_{-k, k}^\top \Sigma_{-k, -k}^{-1} \Sigma_{-k, k}$  and  $\gamma_k = \Sigma_{-k, -k}^{-1} \Sigma_{-k, k} = \frac{\Theta_{-k, k}}{\theta_{kk}}$ . Note that  $\gamma_k$  is a function of  $\Theta$ , it implies that  $\gamma_k$  can reflect the dependence between  $k$ -th variable and others. Therefore, the estimators of  $\beta_k$  and  $\gamma_k$  can be determined by

$$\begin{aligned} (\tilde{\beta}_k, \tilde{\gamma}_k) = \underset{\beta_k, \gamma_k}{\operatorname{argmin}} & \left\{ \left\| \mathbf{y}^k - \mathbf{X}\beta_k - (\mathbf{Y}^{-k} - \mathbf{X}\mathbf{B}_{-k}^{(0)})\gamma_k \right\|_2^2 \right. \\ & \left. + \zeta_1 \sum_{j=1}^p u_{jk} |\beta_{jk}| + \zeta_2 \sum_{s \neq k} v_{sk} |\gamma_{sk}| \right\}, \end{aligned}$$

where  $\mathbf{B}_{-k}^{(0)}$  is an initial consistent estimate of  $\mathbf{B}_{-k}$ ,  $\zeta_1$  and  $\zeta_2$  are two tuning parameters, and  $u_{jk}$  and  $v_{sk}$  are weights for the adaptive LASSO.

Since  $\Theta$  is a symmetric matrix, it implies that  $\operatorname{sign}(\theta_{sk}) = \operatorname{sign}(\theta_{ks})$  and  $\operatorname{sign}(\gamma_{sk}) = \operatorname{sign}(\gamma_{ks})$ . Similar to the crucial issue in Section 2.2.2, two estimators  $\hat{\gamma}_{sk}$  and  $\hat{\gamma}_{ks}$  may not be equal. Therefore, Wang (2015) suggested setting the final estimators to be zero by the ‘AND rule’  $\tilde{\gamma}_{sk} = 0$  and  $\tilde{\gamma}_{ks} = 0$  or the ‘OR rule’  $\tilde{\gamma}_{sk} = 0$  or  $\tilde{\gamma}_{ks} = 0$ .

Regarding comparisons of numerical results, Wang (2015) examined his proposed method and compared the performance with Lee & Liu (2012). Numerical results showed that the method proposed by Wang (2015) outperforms that developed by Lee & Liu (2012) with smaller biases of estimators of  $\mathbf{B}$  and  $\Theta$ .

In addition to two methods that are based on the GLASSO method and the conditional inference, respectively, some approaches were also proposed to deal with estimations of the precision matrix in the presence of the multivariate responses. For example, as motivated by analysis of genetical genomics data, Cai *et al.* (2013) and Yin & Li (2013) proposed a two-stage estimation

procedure to first identify the relevant covariates that affect the means by a  $\ell_1$  penalisation and then estimate the precision matrix using the estimated regression coefficients in the first stage. Rothman *et al.* (2010) focused on improving estimation of regression coefficients by incorporating the covariance information. Li *et al.* (2012) developed a method that is based on a combination of a kernel-based estimate of the means and a regularised estimate of the precision matrix.

## 4.2 Classification

While several methods have been developed to estimate graphical structures based on multivariate linear models as presented in Section 4.1, little work has been available to address graphical structure in classification, especially for multi-label classes, which is an important problem in supervised learning. The goal of classification is to use the information of covariates to classify subjects to  $\mathcal{I}$  different classes, where  $\mathcal{I} \geq 2$  is the number of labels. As discussed in Chen (2018), it is expected that there exists the (pairwise) dependence structure within covariates and that network structures in each class may be different from each other.

In the existing literature, some machine learning methods with network structures accommodated have been discussed. To name a few, Cai *et al.* (2018) proposed the network linear discriminant analysis that takes network information in predictive variables into consideration. Zhu *et al.* (2009) considered penalised support vector machine (SVM) whose penalty function is based on the set of all pairs of connected covariates. However, those approaches focused on the binary outcome, that is,  $\mathcal{I} = 2$ , which is a special case of multiclassification with  $\mathcal{I} \geq 2$ .

In this section, we introduce some recent works that can handle classification for multi-classes ( $\mathcal{I} > 2$ ). The first method proposed by Chen *et al.* (2019) is the logistic regression with graphically structured covariates accommodated. Specifically, to describe the covariate  $X$  as well as the network structure, the exponential family graphical model (4) is employed, and the corresponding conditional distribution of  $x_s$  given  $x_{\setminus\{s\}}$  for  $s = 1, \dots, p$  is given by

$$\mathbb{P}_{\theta_s}(x_s | x_{\setminus\{s\}}) = \exp \left[ x_s \left\{ \sum_{t \in V \setminus \{s\}} \theta_{st} x_t \right\} + \mathfrak{C}(x_s) - \mathfrak{D} \left\{ \sum_{t \in V \setminus \{s\}} \theta_{st} x_t \right\} \right],$$

where  $\theta_s = (\theta_{s1}, \dots, \theta_{s(s-1)}, \theta_{s(s+1)}, \dots, \theta_{sp})^\top$ ,  $x_{\setminus\{s\}}$  is the  $(p-1)$ -dimensional vector as defined in (14), and  $\mathfrak{D}(\cdot)$  is the normalising constant. Based on sample with size  $n$ , the conditional inference in Section 2.2.2 can be employed, and the estimator of  $\theta_s$  for  $s = 1, \dots, p$  is given by

$$\hat{\theta}_s = \operatorname{argmin}_{\theta_s} \left\{ -\frac{1}{n} \sum_{j=1}^n \log \mathbb{P}_{\theta_s}(x_{js} | x_{j, \setminus\{s\}}) + \lambda \|\theta_s\|_1 \right\}, \quad (57)$$

where  $\lambda$  is a tuning parameter. Thus, the estimated edge set and the estimated graph can be obtained by AND rule.

There are two methods to do classification. The first approach is called *logistic regression with homogeneous graphically structured predictors* (LR-HomoGraph), which considers the case where the subjects in different classes share a common network structure in the predictors. Specifically, let  $\hat{E}$  denote the resulting edge set based on whole data. Then the network based nominal logistic regression is given by



$$p_{ij}(x_j) = \frac{\exp\left(\alpha_{i0} + \sum_{(s,t) \in \widehat{E}} \alpha_{i, st} x_{js} x_{jt}\right)}{1 + \sum_{l=1}^{\mathcal{I}-1} \exp\left(\alpha_{l0} + \sum_{(s,t) \in \widehat{E}} \alpha_{l, st} x_{js} x_{jt}\right)} \quad (58)$$

for  $i = 1, 2, \dots, \mathcal{I} - 1$ , where  $(\alpha_{i0}, (\alpha_{i, st} : (s, t) \in \widehat{E})^\top)^\top$  is the vector of parameters associated with class  $i$  and the constraint  $\sum_{i=1}^{\mathcal{I}} p_{ij}(x) = 1$  is imposed for every  $j = 1, \dots, n$ . When  $(\alpha_{i0}, (\alpha_{i, st} : (s, t) \in \widehat{E})^\top)^\top$  for  $i = 1, \dots, \mathcal{I}$  is estimated by the likelihood function of (58) (e.g. Agresti, 2012, p. 273), (58) can be estimated accordingly, and denote the estimator as  $\widehat{p}_{ij}(x_j)$ . The predicted class of a subject  $j$ , denoted as  $i^*$ , is then determined by the largest value of  $\{\widehat{p}_{1j}(x_j), \dots, \widehat{p}_{\mathcal{I}j}(x_j)\}$ , that is,  $i^* = \operatorname{argmax}_{i=1, \dots, \mathcal{I}} \widehat{p}_{ij}(x_j)$ .

The second method, called the *logistic regression with class-dependent graphically structured covariates* (LR-ClassGraph), stratifies the covariate information by class when characterising the covariate network structures and uses network structures in different classes to classify subjects. Specifically, for every  $i$  and  $j$  with  $i = 1, \dots, \mathcal{I}$  and  $j = 1, \dots, n$ , define a binary and surrogate response variable

$$Y_j^i = \begin{cases} 1, & \text{the } j\text{-th subject is in class } i, \\ 0, & \text{otherwise} \end{cases}$$

and let  $\widehat{E}^i$  denote an estimated set of edges for predictors in class  $i$ . After that, define  $\pi^i(x_j) = P(Y_j^i = 1 | X_j = x_j)$  and consider the class-dependent logistic regression

$$\operatorname{logit}\{\pi^i(x_j)\} = \gamma_0^i + \sum_{(s,t) \in \widehat{E}^i} \gamma_{st}^i x_{js} x_{jt}, \quad (59)$$

where  $(\gamma_0^i, (\gamma_{st}^i : (s, t) \in \widehat{E}^i)^\top)^\top$  is the vector of parameters associated with class  $i$ . Applying the maximum likelihood estimation method based on (59) yields the estimator of  $(\gamma_0^i, (\gamma_{st}^i : (s, t) \in \widehat{E}^i)^\top)^\top$ , and thus, the estimator  $\widehat{\pi}^i(x_j)$  can be obtained from (59). Therefore, the predicted class label for a subject  $j$  is determined by  $i^* = \operatorname{argmax}_{i=1, \dots, \mathcal{I}} \widehat{\pi}^i(x_j)$ .

Regarding machine learning methods, network structures are also accommodated to support vector machine (SVM). For the multi-class response, He *et al.* (2019) considered the exponential family graphical model and employed (57) to estimate the network structure. Different from the approach in Chen *et al.* (2019) that adopt pairwise interactions to reflect network structures, He *et al.* (2019) proposed the network based surrogate covariates to replace the ‘original’ covariates in SVM.

Suppose that  $\widehat{G} = (V, \widehat{E})$  is the estimated graph obtained by (57) based on whole data. To reflect different association structures among the covariates, we divide the estimated graph  $\widehat{G}$  as a sequence of non-overlapped and interconnected subgraphs  $\{\widehat{G}^k : k = 1, \dots, K\}$ , where  $1 \leq K \leq p$  is the number of subgraphs in  $\widehat{G}$ , and  $\widehat{G}^k = (V^k, \widehat{E}^k)$  represents the  $k$ -th subgraph

with  $V^k$  and  $\widehat{E}^k$  being the corresponding vertex and edge subsets, respectively. Moreover,  $\bigcup_{k=1}^K V^k = V$  and two subsets  $V^{k_1}$  and  $V^{k_2}$  are disjoint for  $k_1 \neq k_2$ . When the edge subset  $\widehat{E}^k$  is empty, the corresponding vertex subset  $V^k$  contains a single element.

We now adopt  $\widehat{G}^k$  with  $k = 1, \dots, K$  to define surrogate covariates, and each subgraph reflects a new covariate, yielding a  $K$ -dimensional vector of predictors, denoted as  $X_j^* \triangleq (X_{j,1}^*, \dots, X_{j,K}^*)^\top$ . The first formulation summarises the predictor measurements using the vertex information in the subgraphs and defines  $X_j^*$  as  $X_j^V = (X_{j,1}^V, \dots, X_{j,K}^V)^\top$  with

$$X_{j,k}^V = \frac{1}{|V^k|} \sum_{s \in V^k} X_{j,s} \quad (60)$$

for  $k = 1, \dots, K$ , where  $|V^k|$  is the cardinality of the vertex subset  $V^k$ . If there exists the  $k$ -th vertex that is fully unconnected other vertices, that is,  $V^k = \{k\}$ , then the  $k$ -th surrogate variable is defined as  $X_{j,k}^V = X_{j,k}$ .

The second formulation uses the edge information in the subgraphs and defines  $X_j^*$  as  $X_j^E = (X_{j,1}^E, \dots, X_{j,K}^E)^\top$  with

$$X_{j,k}^E = \frac{1}{|\widehat{E}^k|} \sum_{(s,t) \in \widehat{E}^k} X_{j,s} X_{j,t} \quad (61)$$

for  $k = 1, \dots, K$ , where  $|\widehat{E}^k|$  is the cardinality of the edge subset  $\widehat{E}^k$ . Moreover, noting that when  $\widehat{E}^k$  is empty,  $X_{j,k}^E$  is defined as the predictor  $X_{j,k}$  whose index falls in the corresponding vertex subset  $V^k$ . Finally, when the vector of surrogate covariates is derived, replacing the original covariate  $X$  by the surrogate covariate  $X^*$  in the SVM algorithm enables us to do classification.

The last strategy based on supervised learning is based on discriminant analysis. Let  $f_{ji}(X_{j,\bullet})$  denote the conditional probability density function of the predictor  $X_{j,\bullet}$  given that subject  $j$  comes from the  $i$ -th class for  $i = 1, \dots, \mathcal{I}$  and  $j = 1, \dots, n$ . Let  $\pi_i = P(Y_j = i)$  denote the probability that the  $j$ -th subject is randomly selected from class  $i$ . It is immediate that  $\sum_{i=1}^{\mathcal{I}} \pi_i = 1$ . By some algebra (Hastie *et al.*, 2008, p. 108) and the Bayes theorem, we obtain the posterior probability

$$P(Y_j = i | X_{j,\bullet}) = \frac{f_{ji}(X_{j,\bullet})\pi_i}{\sum_{l=1}^{\mathcal{I}} f_{jl}(X_{j,\bullet})\pi_l} \quad (62)$$

for  $i = 1, \dots, \mathcal{I}$  and  $j = 1, \dots, n$ . For arbitrary two classes  $i$  and  $l$  with  $i \neq l$ , the log-ratio of (62) is defined as

$$\log \left\{ \frac{P(Y_j = i | X_{j,\bullet})}{P(Y_j = l | X_{j,\bullet})} \right\} = \log \left( \frac{f_{ji}(X_{j,\bullet})}{f_{jl}(X_{j,\bullet})} \right) + \log \left( \frac{\pi_i}{\pi_l} \right). \quad (63)$$

If we particularly specify  $f_{ji}(\cdot)$  as the normal distribution  $N(\mu_i, \Sigma_i)$  for class  $i$  and define  $\Theta_i = \Sigma_i^{-1}$  as in Section 2.1.1, then (63) will become

$$\log\left(\frac{\pi_i}{\pi_l}\right) + \log\left(\frac{|\Theta_l|^{-1/2}}{|\Theta_i|^{-1/2}}\right) + \frac{1}{2}\left\{(X_{j\cdot} - \mu_l)^\top \Theta_l(X_{j\cdot} - \mu_l) - (X_{j\cdot} - \mu_i)^\top \Theta_i(X_{j\cdot} - \mu_i)\right\}. \quad (64)$$

Moreover, if  $\Sigma_i$  is equal to a common matrix  $\Sigma$  for all  $i = 1, \dots, \mathcal{I}$ , then we can define  $\Theta \triangleq \Sigma^{-1}$ , and (64) will reduce to

$$\log\left(\frac{\pi_i}{\pi_l}\right) - \frac{1}{2}(\mu_i + \mu_l)^\top \Theta(\mu_i + \mu_l) + X_{j\cdot}^\top \Theta(\mu_i + \mu_l). \quad (65)$$

Following the discussion in Hastie *et al.* (2008), (64) and (65) separately give a quadratic function with respect to  $x$  based on the class  $i$

$$\varphi_i(x) = \log(\pi_i) + \frac{1}{2}\log|\Theta_i| - \frac{1}{2}(x - \mu_i)^\top \Theta_i(x - \mu_i) \quad (66)$$

and a linear function with respect to  $x$  based on the class  $i$

$$\delta_i(x) = \log(\pi_i) - \frac{1}{2}\mu_i^\top \Theta \mu_i + x^\top \Theta \mu_i. \quad (67)$$

If  $\Theta_i$  in (66) and  $\Theta$  in (67) are sparse, then one can adopt the GLASSO method in Section 2.2.1 to estimate them, yielding NetQDA and NetLDA for estimated (66) and (67), respectively (e.g. Chen, 2022c). The implementation can be found by the R package `NetDA` discussed by Chen (2022a). In fact, if  $\Sigma_i$  and  $\Sigma$  are empirically estimated and are directly implemented to (66) and (67), then they are referred to conventional linear/quadratic discriminant analysis (LDA/QDA) (e.g. Hastie *et al.*, 2008). For the comparisons among NetLDA/NetQDA and LDA/QDA, the former methods are able to deal with estimation of sparse  $\Theta$  or  $\Theta_i$ , while the latter ones fail to address conditional independence of two predictors. Hence, in the presence of dependence structure of predictors, it is expected that the NetLDA and NetQDA methods outperform the conventional LDA and QDA methods.

#### 4.3 Joint Modelling for Survival Data

Survival analysis is an important topic in statistical analysis and it has been widely applied in biostatistics, actuarial science and so on. In the framework of survival analysis, the failure time is set as the outcome of main interest. Different from generic linear models where the response is complete, the main challenge is that the survival outcome is usually incomplete due to right-censoring, which is mainly caused by the loss of follow-up of individuals.

A motivated example of this study is the breast cancer data collected by the Netherlands Cancer Institute (NKI) (van de Vijver *et al.*, 2002). Tumours from 295 women with breast cancer were collected from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. Of all those patients, 79 patients died before the study ended, yielding approximately the 73.2% censoring rate. In addition, the dataset also contains 70 genes that are useful for tumour diagnosis. In this study, the main interest is to construct a survival model by treating gene expressions as predictors.

Let  $\tilde{T}$  and  $\tilde{C}$  be the failure time and the censoring time, respectively, and let  $\Delta = \mathbb{I}(\tilde{T} \leq \tilde{C})$  be the censoring indicator. Let  $T = \min\{\tilde{T}, \tilde{C}\}$  denote the ‘observed’ survival time and let  $X = (X_C^\top, X_D^\top)^\top$  be a  $p$ -dimensional random vector of covariates with  $X_C$  and  $X_D$  being

continuous and discrete variables, respectively. As commented by Chen (2018), we allow  $X$  to have network structure that can be characterised by mixed graphical models (6).

In standard survival analysis, the Cox proportional hazards (PH) model is often employed with the hazard function specified as

$$\mathcal{H}(t|X) = \mathcal{H}_0(t) \exp\{g(X; \alpha)\},$$

where  $\mathcal{H}_0(\cdot)$  is the unspecified baseline hazard function, and  $g(X; \alpha)$  is the link function of the linear predictor with the covariate vector  $X$  and the unknown parameter  $\alpha$ . To incorporate the PH model with network structure in covariates, Chen & Yi (2021a) suggested specifying  $g(X; \alpha) = \log\{\mathbb{P}_{\beta, \Theta}(X)\}$ , where  $\mathbb{P}_{\beta, \Theta}(X)$  is given by

$$\mathbb{P}_{\beta, \Theta}(X) = \exp\left\{\sum_{r \in V} \beta_r \mathfrak{B}(X_r) + \sum_{(s, v) \in E} \theta_{sv} \mathfrak{B}(X_s) B(X_v) + \sum_{r \in V} \mathfrak{C}(X_r) - \mathfrak{A}(\beta, \Theta)\right\}.$$

It yields the generalised Cox proportional hazards model:

$$\mathcal{H}(t|X) = \mathcal{H}_0(t) \exp\left\{\sum_{r \in V} \beta_r \mathfrak{B}(X_r) + \sum_{(s, v) \in E} \theta_{sv} \mathfrak{B}(X_s) B(X_v) + \sum_{r \in V} \mathfrak{C}(X_r) - \mathfrak{A}(\beta, \Theta)\right\}, \quad (68)$$

where  $\beta_r$  for  $r \in V$  is the parameter that reflects the main effect associated with the covariate  $X_r$ , and for  $(s, v) \in E$ , the parameter  $\theta_{sv}$  facilitates the association of  $X_s$  and  $X_v$  in the sense that  $\theta_{sv} \neq 0$  shows the *conditional dependence* of  $X_s$  and  $X_v$  given other covariates.

To estimate unknown parameters in (68), the partial likelihood function is frequently employed (e.g. Lawless, 2003). Based on the observed sample  $\{(T_i, X_i, \Delta_i) : i = 1, \dots, n\}$  with  $\mathfrak{B}(x) = x$ , the likelihood function is given by

$$\begin{aligned} \ell(\beta, \Theta) = \sum_{i=1}^n \int & \left[ \left( \sum_{r \in V} X_{i,r} \beta_r + \sum_{(s, v) \in E} X_{i,s} X_{i,v} \theta_{sv} \right) \right. \\ & \left. - \log \left\{ \sum_{j=1}^n \exp \left( \sum_{r \in V} X_{j,r} \beta_r + \sum_{(s, v) \in E} X_{j,s} X_{j,v} \theta_{sv} \right) \mathbb{Y}_j(t) \right\} \right] d\mathbb{N}_i(t), \end{aligned} \quad (69)$$

where  $\mathbb{N}_i(t) = \mathbb{I}(T_i < t, \Delta_i = 1)$  and  $\mathbb{Y}_i(t) = \mathbb{I}(T_i \geq t)$ .

Ideally, the estimators of  $\beta$  and  $\Theta$  can be obtained by maximising (69). However, this approach would fail when  $\beta$  and  $\Theta$  are assumed to be sparse and the covariate is contaminated with measurement error. To simultaneously deal with measurement error, variable selection for  $\beta$  and network detection for  $\Theta$ , Chen & Yi (2021a) proposed a simulation-based three-stage procedure. Specifically, following the strategy in Section 3.5, (47) and (48) are employed to generate the working data  $W_{i,\cdot}(r, \zeta)$  in the first stage. After that, for  $r = 1, \dots, R$  and  $\zeta \in \mathcal{Z}$  as described in Section 3.5, we define the surrogate likelihood function  $\ell_{r,\zeta}(\beta, \Theta)$  that is determined by (69) with  $X_{i,\cdot}$  replaced by the working data  $W_{i,\cdot}(r, \zeta)$ . Then the optimisation problem of the penalised likelihood function with double penalty functions is proposed:

$$(\hat{\beta}_r(\zeta), \hat{\Theta}_r(\zeta)) = \underset{\beta, \Theta}{\operatorname{argmin}} \{ \ell_{r,\zeta}(\beta, \Theta) + \lambda_1 \varphi_1(\beta) + \lambda_2 \varphi_2(\Theta) \},$$

which can be solved by the block-coordinate-descent algorithm. In the last stage, we fit regression models to each of the two sequences  $\left\{(\zeta, \hat{\beta}(\zeta)) : \zeta \in \mathcal{Z}\right\}$  and  $\left\{(\zeta, \hat{\Theta}(\zeta)) : \zeta \in \mathcal{Z}\right\}$  with  $\hat{\beta}(\zeta) = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_r(\zeta)$  and  $\hat{\Theta}(\zeta) = \frac{1}{R} \sum_{r=1}^R \hat{\Theta}_r(\zeta)$ . Finally, the estimators of  $\beta$  and  $\Theta$  are obtained by specifying as the predicted values of fitted models at  $\zeta = -1$ .

## 5 Real Data Applications

### 5.1 Example 1: Analysis of the Cell-Signalling Data

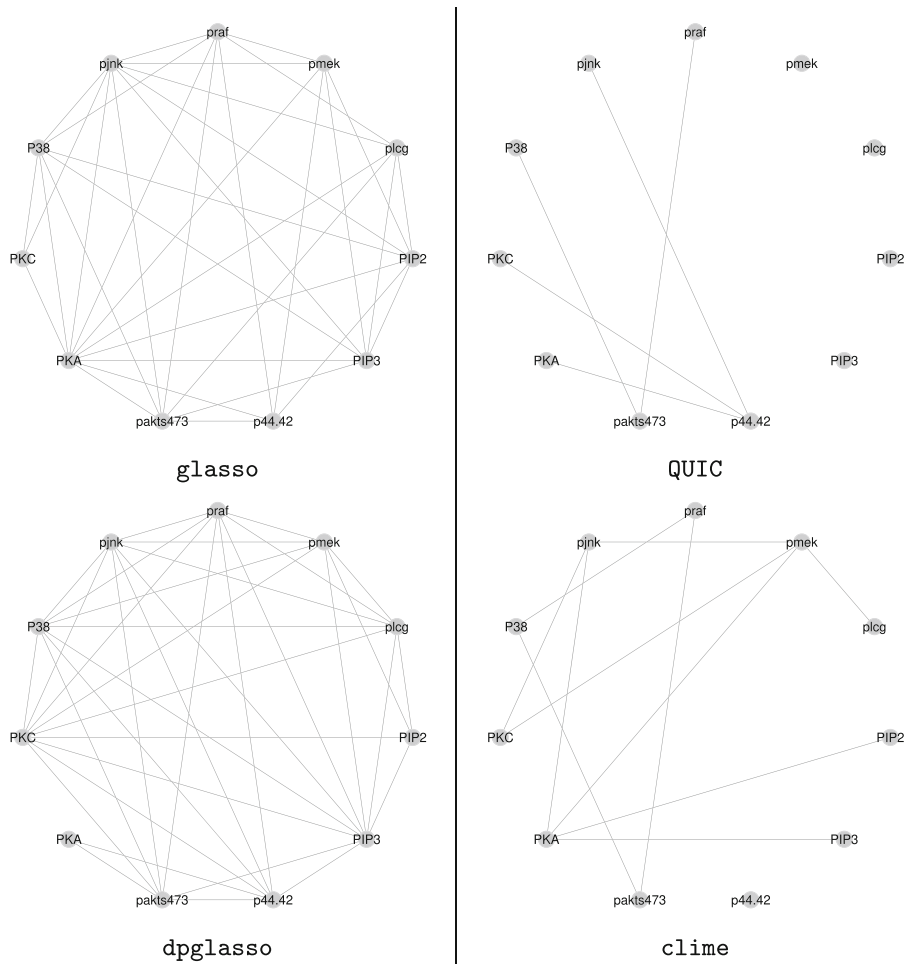
In the first data application, we study the cell-signalling dataset (e.g. Sachs *et al.*, 2005), which contains  $p = 11$  proteins and  $n = 7466$  cells. In this study, our primary interest is to understand the relationship among various signalling proteins by investigating signalling pathways and the dependence structure of proteins. As a result, the goal is to adopt estimation methods in Sections 2 and 3 to characterise the network structure of 11 proteins.

In our analysis, we start by considering the parametric estimation in Section 2.2 and primarily implement some existing packages summarised in Section 2.3. Specifically, for the GLASSO based approaches, we examine the R packages `glasso`, `QUIC`, `dpglasso` and `clime`; the resulting graphs are displayed in Figure 3. For the C.I. approach, we employ the estimation methods in Section 2.2.2, which can be implemented by the R packages `XMRF`, `space` and `gconcord`, respectively. The resulting graphs are displayed in Figure 4. In Figure 3, we observe that two packages `glasso` and `dpglasso` produce graphs with complex edges, while graphs determined by `QUIC` and `clime` contain less edges. Similarly, in Figure 4, the `XMRF` method has the most complex network structure and a graph derived by `gconcord` is most sparse. From the first glance in Figures 3 and 4, it is interesting to see that no pair of variables with/without edge is commonly detected by those seven methods. For example, a pair (praf, pakts.473) can not be detected by the `space` method only; a pair (PKC, PKA) can only be detected by the `glasso` method. It shows that the estimation results in this real dataset are sensitive under various estimation strategies.

Next, we relax assumptions under Section 2.1. The first extension is the non-parametric setting in (20). We primarily adopt the R package `huge` to identify the network structure, which is displayed in Figure 5. It is clear to see that a graph determined by `huge` is more complex than other graphs in Figures 3 and 4 as most variables are linked with edges, except for some pairs, such as (plcg, PIP2) and (PIP2, PIP3).

The second extension is the consideration of measurement error. We primarily examine the SIMEX approach proposed by Chen & Yi (2022). In the presence of measurement error, to implement estimation method to address measurement error effects, we employ sensitivity analyses and specify the covariance matrix  $\Sigma_{\epsilon} = (Q^{-1} - 1)\hat{\Sigma}_{X^*}$  with  $Q = 0.65, 0.75$  and  $0.85$  reflecting different magnitudes of measurement error effects, where  $\hat{\Sigma}_{X^*}$  is the empirical estimate of the covariance matrix based on the data. Here, we display the same result derived by Chen & Yi (2022) in Figure 6.

We observe that more edges are detected when  $Q$  is increasing, such as two additionally identified pairs (PIP3, praf) and (pjnk, praf) between  $Q = 0.75$  and  $Q = 0.65$ , and another two pairs (pakts473, pjnk) and (pakts473, praf) when  $Q$  is increasing from 0.75 to 0.85. On the other hand, network structures in Figures 3 and 4, which can be regarded as the *naive* analysis by using error-prone variables, have different results from the SIMEX method which accounts for measurement error effects. For example, `glasso`, `dpglasso`, `XMRF` and `space` produce more complex network structures. While `QUIC`, `clime`, `gconcord` and the SIMEX method



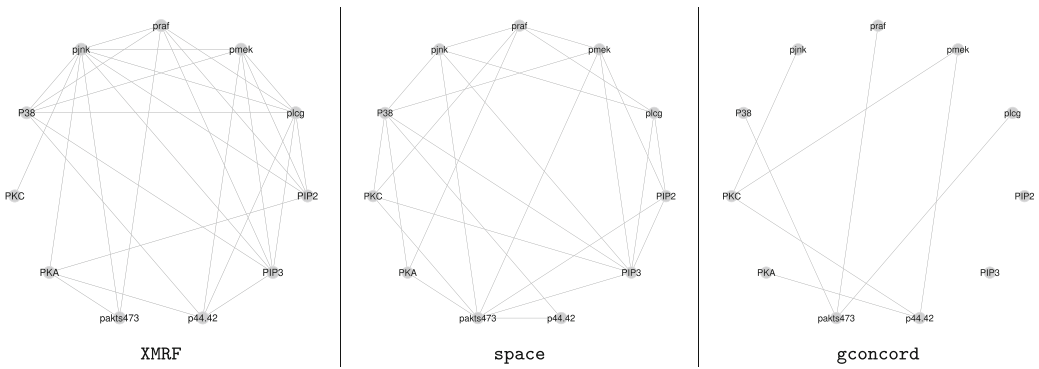
**Figure 3.** Data application in Section 5.1: network structures determined four different methods in Section 2.2.1

with different values of  $R$  provide sparse network structures, it is interesting to see that some edges are detected by the SIMEX method only, such as  $(PIP3, pjnk)$ ,  $(PIP3, plc9)$  and  $(praf, pmek)$ . On the contrary, a pair  $(pakt473, P38)$  can be identified by QUIC, clime and gconcord only, but is not available in Figure 6. The examination of measurement error simply demonstrates that in the presence of measurement error in the variables, ignoring the feature of mismeasurement may produce spurious correlation structures among the variables.

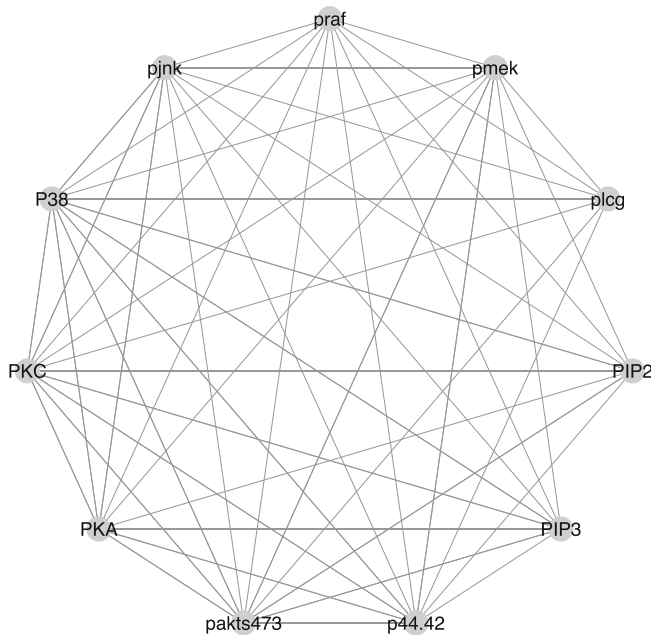
## 5.2 Example 2: Classification With Gene Expression Data

In this section, we analyse the gene expression data collected by Golub *et al.* (1999) and compare the performance of network-based classification methods in Section 4.2.

The dataset contains 7128 genes that were measured using Affymetrix oligonucleotide arrays and the binary outcome including acute myeloid leukaemia (AML, labelled as ‘+1’) and acute lymphoblastic leukaemia (ALL, labelled as ‘−1’). According to the description of Golub *et al.* (1999), the purpose of this study is to identify gene signature for the distinction between AML and ALL. The sample size in the data is 72, coming from the two classes, with 47 specimens



**Figure 4.** Data application in Section 5.1: network structures determined three different methods in Section 2.2.2



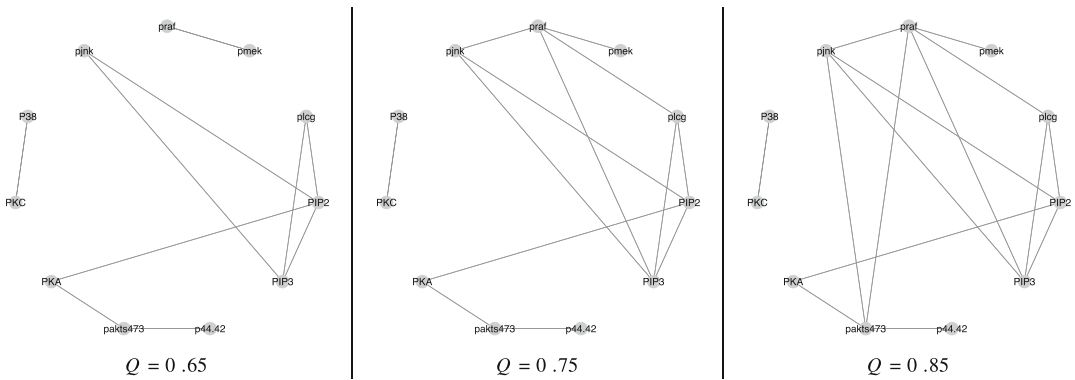
**Figure 5.** Data application in Section 5.1: a network structure determined by the package huge

in class ALL and 25 specimens in class AML. In particular, according to the study design, those 72 samples are composed of the training data of 38 specimens (27 in class ALL and 11 in class AML) and the testing data of 34 specimens (20 in class ALL and 14 in class AML).

As commented by Chen (2018) and Grimes *et al.* (2019), network structure is ubiquitous in biological data, it motivates us to construct classification models with network structures of gene expressions incorporated. We mainly adopt the network-based classification methods in Section 4.2, including the logistic regression (Chen *et al.*, 2019), discriminant analysis (Chen, 2022c) and SVM (He *et al.*, 2019) methods, to examine the gene expression analysis and make comparisons among those methods.

Noting that the number of genes is extremely larger than the sample size, to make analysis more stable and reasonable, it is necessary to remove non-informative features before applying learning algorithms. To detect important genes, we adopt the distribution-free feature screening method proposed by Chen (2023), which is a powerful approach to address





**Figure 6.** Data application in Section 5.1: network structures determined by the SIMEX method proposed by Chen and Yi (2022) with different values of  $Q$

ultrahigh-dimensional data and is valid to detect informative gene expression values for binary/categorical responses (e.g. Chen, 2022b). It turns out that 14 genes are selected from the 38 training sample data, which are strongly correlated to the response of different leukaemia types, and those selected genes are labelled by their own ID numbers.

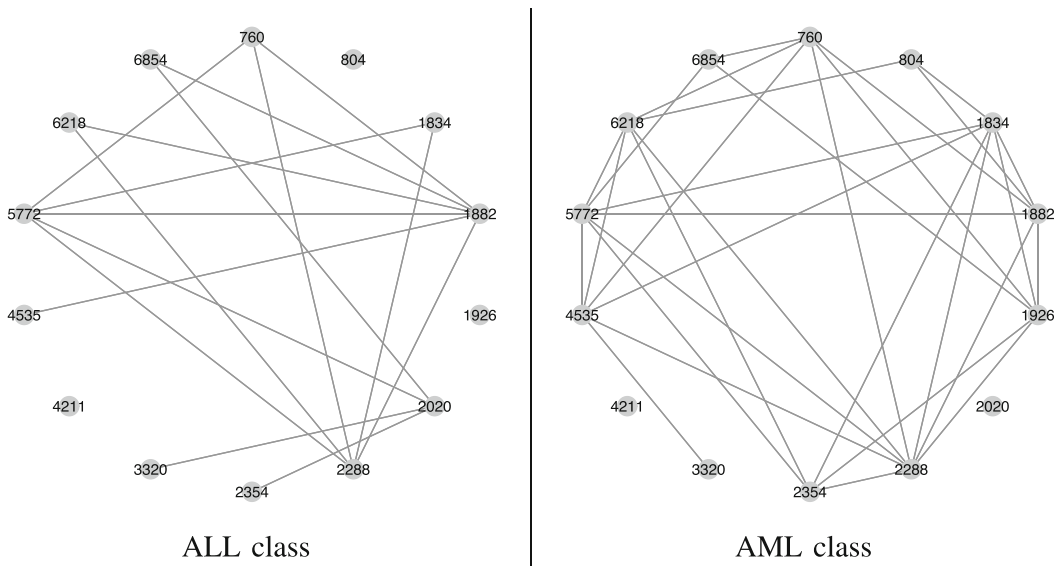
We now adopt classification models in Section 4.2 to fit the training data. Specifically, we first use the package `XMRF` to identify network structures with (or without) involvement of the binary response, and the results are displayed in Figures 7 and 8, respectively. In addition, we adopt the R package `NetDA` to determine class-dependent and pooled-sample network structures, and display them in Figures 9 and 10, respectively. Based on two classes, Figures 7 and 9 show that network structures of gene expressions are different from each other, and it suggests that specific network structure may reflect the associated class. For the comparison among Figures 7–10, we can observe that network structures displayed in Figures 9 and 10 look more complex than those summarised in Figures 7 and 8.

After that, we construct three models `LR-ClassGraph`, `SVM-ClassVertex` and `SVM-ClassEdge` based on graphs in Figure 7, where `SVM-ClassVertex` and `SVM-ClassEdge` denote the SVM method with the predictors being replaced by surrogate predictors (60) and (61) based on the subgraphs in Figure 7, respectively. In addition, based on the network structure in Figure 8, we derive `LR-HomoGraph`, `SVM-HomoVertex` and `SVM-HomoEdge`, where `SVM-HomoVertex` and `SVM-HomoEdge` approaches follow the similar definitions of `SVM-ClassVertex` and `SVM-ClassEdge` but are derived based on the network structure in Figure 8. Moreover, to implement the `NetQDA` and `NetLDA` methods, we adopt network structures in Figures 9 and 10 and derive estimates (66) and (67), respectively.

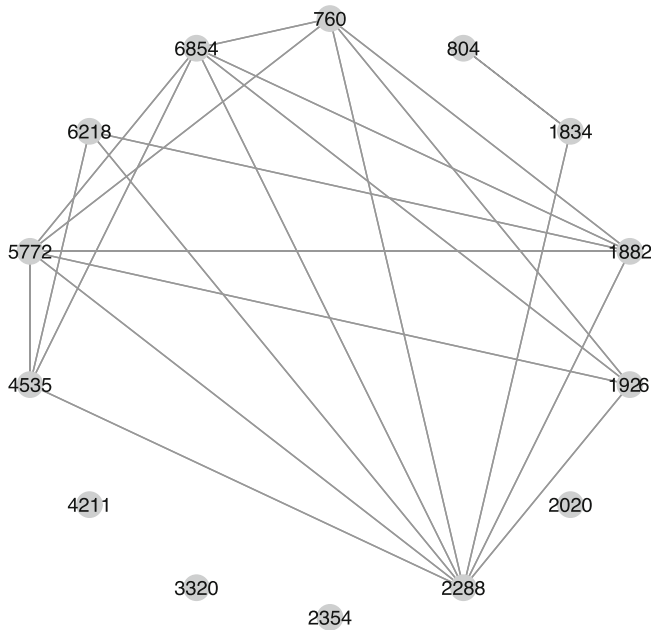
Finally, when the fitted models based on the training data are constructed, we further use the gene expressions in the testing data to do the prediction. To assess the performance of prediction, we primarily examine the F-score. Specifically, for subject  $j$  in the testing data with  $j = 1, \dots, 34$ , let  $\hat{y}_j$  denote the predicted class label and let  $y_j$  denote the true class label. For class  $i \in \{-1, 1\}$ , we calculate the number of the *true positives* (TP), the number of the *false positives* (FP), and the number of the *false negatives* (FN) as follows:

$$\text{TP} = \sum_{j=1}^{34} \mathbb{I}(y_j = +1, \hat{y}_j = +1), \quad \text{FP} = \sum_{j=1}^{34} \mathbb{I}(y_j = -1, \hat{y}_j = +1),$$

and



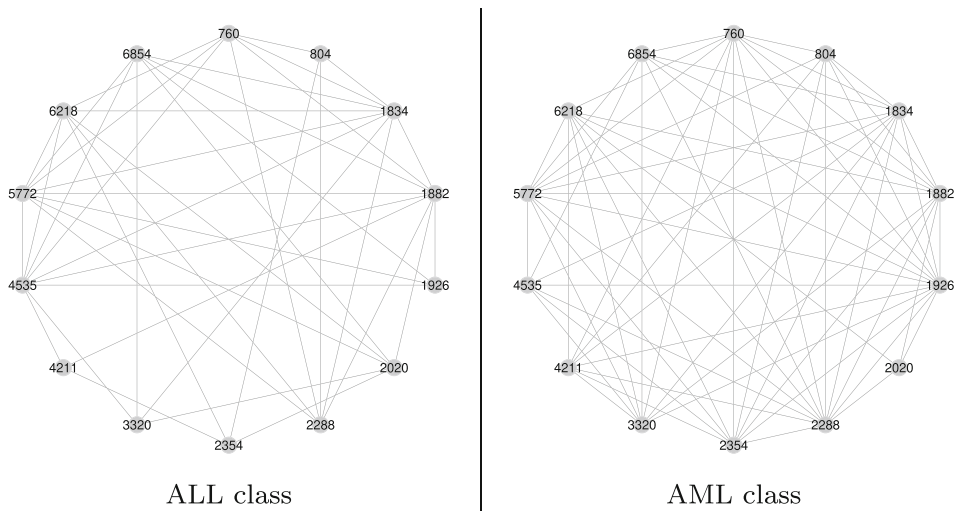
**Figure 7.** Data application in Section 5.2: network structures based on two different classes for the LR-ClassGraph, SVM-ClassVertex and SVM-ClassEdge methods.



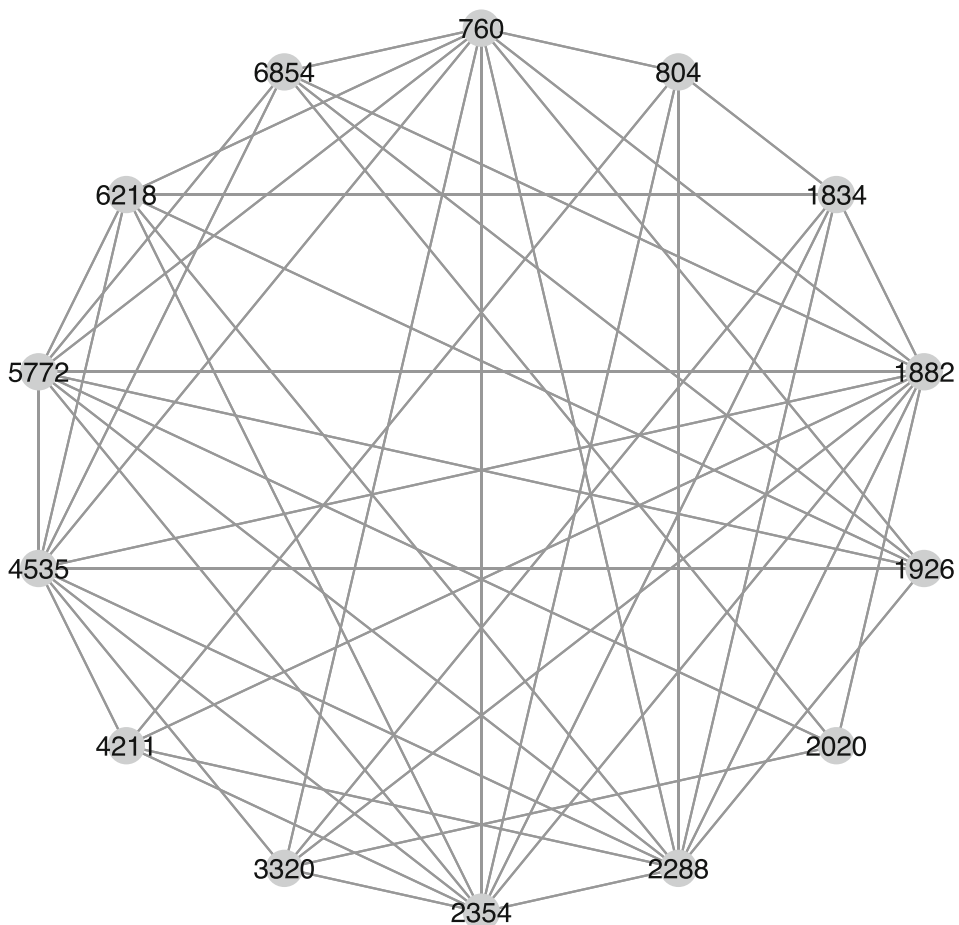
**Figure 8.** Data application in Section 5.2: a network structure determined for the LR-HomoGraph, SVM-HomoVertex and SVM-HomoEdge methods.

$$\text{FN} = \sum_{j=1}^{34} \mathbb{I}(y_j = +1, \hat{y}_j = -1).$$

Then *precision*, *recall* and *F-score* are respectively given by



**Figure 9.** Data application in Section 5.2: network structures based on two different classes for the NetQDA method.



**Figure 10.** Data application in Section 5.2: a network structure determined for the NetLDA method.

Table 4. Data application in Section 5.2: overall performance of classification methods applied to gene expression data.

Methods	PRE	REC	F-score
LR-HomoGraph	1.000	0.928	0.965
LR-ClassGraph	1.000	0.928	0.965
SVM-HomoVertex	0.911	1.000	0.953
SVM-HomoEdge	1.000	0.911	0.953
SVM-ClassVertex	0.933	1.000	0.965
SVM-ClassEdge	1.000	0.928	0.965
NetLDA	0.911	0.911	0.911
NetQDA	1.000	1.000	1.000

$$PRE = \frac{TP}{TP + FP}, REC = \frac{TP}{TP + FN}, \text{ and F-score} = 2 \times \frac{PRE \times REC}{PRE + REC}.$$

In principle, PRE, REC and F-score are between zero to one. Higher values of PRE, REC and F-score reflect better prediction.

Numerical results for the prediction results, including PRE, REC and F-scores, are summarised in Table 4. We observe that the LR-HomoGraph and LR-ClassGraph methods have the same result. With class-dependent network structures accommodated, SVM-ClassVertex and SVM-ClassEdge outperform SVM-HomoVertex and SVM-HomoEdge, and NetQDA is more accurate than NetLDA. This phenomenon indicates that the class can be reflected by the corresponding network structure. For the comparisons among methods, we find that NetLDA is slightly worse than LR-HomoGraph, SVM-HomoVertex and SVM-HomoEdge when the network structure is estimated by the pooled sample. On the contrary, with the information of classes accommodated, it is interesting to see that NetQDA has the most accurate prediction and outperforms LR-ClassGraph, SVM-ClassVertex and SVM-ClassEdge.

6 Summary

Graphical models are useful tools to analyse the dependence structure among high-dimensional variables and are widely used in many research areas. In this paper, we overview important topics in the developments of graphical models. We focus the discussion on the estimation procedures and computations. We also summarise fruitful research results for regression models and classification with network structures accommodated. In addition, some information related to existing R packages is also provided in this paper.

Even though estimation methods of graphical models have been explored, some research gaps still remain. For example, in addition to regression models mentioned in Section 4, network structures frequently appear in other types of models as well as data structures. It is expected to extend the graphical structures to deal with other types of data or complex settings, such as non-parametric or semiparametric models. Regarding the machine learning frameworks, it is interesting and challenging to explore other settings and approaches, such as boosting or neural network methods with graphical structures accommodated. Finally, while several methods have been established, the relevant computational packages have not been available to public users. It is also important to develop R packages for public to implement the estimation methods and further data analysis. Those topics are also the potential research projects in the future.

In this paper, we pay our attention on discussing the framework of graphical models. In statistical analysis, *network data* analysis is the other relevant topic and typically includes technological, biological and information network (e.g. Kolaczyk, 2009; Newman, 2018). Here, we briefly comment the difference between graphical models and network data. In graphical models, known as

probabilistic modellings and referred to the basic setup in Section 2.1, vertices are formulated by random variables with a specific distribution, and edges connecting to pairs of random variables are deterministic and reflect conditional dependence among random variables. On the contrary, in the network data, vertices can be subjects, such as people, and the interest of network data is to explore the connections among subjects, where edges can be regarded as relationship, such as friendship. As a result, randomness would be defined in edges. For example, let  $A_{kl} = A_{lk}$  denote a binary random variable reflecting the presence or absence of an edge between two vertices  $k$  and  $l$  in  $V$ . The matrix  $\mathbb{A} = [A_{kl}]$  is thus the (random) adjacency matrix for a graph.

The second and key difference is the model structure and estimation. As introduced in preceding sections, the developments of graphical models aim to estimate parameters, such as precision matrices, associated with edges of random variables. For the network data, random graphs, referred to a model specifying a collection of possible graphs, are perhaps primary tools to characterise network structures. In the development of random graphs, the exponential random graph model (ERGM, Kolaczyk, 2009, Section 6.5) is one of popular approaches, which is defined as

$$\mathbb{P}_{\alpha}(X) = \exp \left\{ \sum_{j=1}^J \alpha_j T_j(X) - \varphi(\alpha) \right\},$$

where  $\alpha \triangleq (\alpha_1, \dots, \alpha_J)^{\top}$  is a vector of unknown parameters,  $\varphi(\alpha)$  is the normalising constant, and  $T_1(X), \dots, T_J(X)$  are functions of a random vector  $X$  on the space of graphs that could be the number of edges, triangles or stars. There are several approaches for the estimation of  $\alpha$ , including the stochastic approximation under the Robbins–Monro algorithm (e.g. Snijders, 2002) and the importance sampling based on the Geyer–Thompson algorithm (e.g. Handcock, 2003; Hunter & Handcock, 2006). More detailed discussions or relevant developments of network data analysis can be found in some research papers (e.g. Chatterjee *et al.*, 2011; Chatterjee & Diaconis, 2013; Yan & Xu, 2013; Yan *et al.*, 2015, 2016) and monographs (e.g. Crane, 2018; Kolaczyk, 2009, 2017; Newman, 2018); and fundamental computation implementations are summarised by Kolaczyk & Csárdi (2014).

## Endnotes

<sup>1</sup>This package has been archived on <https://cran.r-project.org/src/contrib/Archive/QUIC/>.

<sup>2</sup>This package has been archived on <https://cran.r-project.org/src/contrib/Archive/dpglasso/>.

<sup>3</sup>This package has been archived on <https://cran.r-project.org/src/contrib/Archive/XMRF/>.

<sup>4</sup>This package has been archived on <https://cran.r-project.org/src/contrib/Archive/space/>.

<sup>5</sup>This package has been archived on <https://cran.r-project.org/src/contrib/Archive/gconcord/>.

## ACKNOWLEDGEMENTS

The author would like to thank the Editor, the Associate Editor, and one referee for their valuable comments that significantly improve the initial manuscript. In addition, the author also appreciates his research assistant, Mr. Min-Yi Chen, for helping search for references and run numerical studies during the revision process. Chen's research was supported by National Science and Technology Council with grant ID 110-2118-M-004-006-MY2 and 112-2118-M-004-005-MY2.

## Data availability statement

Data available on request from the authors.

## References

- Agresti, A. (2012). *Categorical Data Analysis*. New York: Wiley.
- Ali, A., Kolter, J.Z. & Tibshirani, R.J. 2016. The multiple quantile graphical model. In *NIPS 2016: Advances in Neural Information Processing Systems 29*, Eds. D.D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett, Barcelona: Spain.
- Avella-Medina, M., Battey, H.S., Fan, J. & Li, Q. (2018). Robust estimation of high dimensional covariance and precision matrices. *Biometrika*, **105**, 271–284.
- Bandara, S., Schlöder, J.P., Eils, R., Bock, H.G. & Meyer, T. (2009). Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput. Biol.*, **5**, e1000558.
- Basu, S., Li, X. & Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Trans. Signal Process.*, **67**(5), 1207–1222.
- Belloni, A., Chen, M. & Chernozhukov, V. 2019. Quantile graphical models: prediction and conditional independence with applications to systemic risk. arXiv:1607.00286v3.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Series B*, **36**, 192–236.
- Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Machine Learn.*, **3**, 1–122.
- Cai, W., Guan, G., Pan, R., Zhu, X. & Wang, H. (2018). Network linear discriminant analysis. *Comput. Stat. Data Anal.*, **117**, 32–44.
- Cai, T.T., Li, H., Liu, W. & Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, **100**, 139–156.
- Cai, T.T., Liu, W. & Luo, X. (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, **106**, 594–607.
- Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Model*. New York: Chapman and Hall.
- Chandrasekaran, V., Parrilo, P.A. & Willsky, A.S. (2012). Latent variable graphical model selection via convex optimization. *Annals Stat.*, **40**, 1935–1967.
- Chatterjee, S. & Diaconis, P. (2013). Estimating and understanding exponential random graph models. *Annals Stat.*, **41**, 2428–2461.
- Chatterjee, S., Diaconis, P. & Sly, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.*, **21**, 1400–1435.
- Chen, L.-P. (2018). Multiclassification to gene expression data with some complex features. *Biostat. Biom. Open Access J.*, **9**(1), 555751. <https://doi.org/10.19080/BBOAJ.2018.09.555751>
- Chen, L.-P. (2022a). NetDA: An R package for network-based discriminant analysis subject to multi-label classes. *J. Probab. Stat.*, **1–14**, 1041752.
- Chen, L.-P. (2022b). Classification and prediction for multi-cancer data with ultrahigh-dimensional gene expressions. *PLOS ONE*, **17**(9), e0274440.
- Chen, L.-P. (2022c). Network-based discriminant analysis for multiclassification. *J. Classif.*, **39**, 410–431.
- Chen, S., Witten, D.M. & Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, **102**, 47–64.
- Chen, L.-P. (2023). A note of feature screening via a rank-based coefficient of correlation. *Biom. J.*, **65**, 2100373.
- Chen, L.-P. & Yi, G.Y. (2021a). Analysis of noisy survival data with graphical proportional hazards measurement error models. *Biometrics*, **77**, 956–969.
- Chen, L.-P., & Yi, G. Y. (2022). De-noising analysis of noisy data under mixed graphical models. *Electron. J. Stat.*, **16**(2), 3861–3909.
- Chen, L.-P. & Yi, G.Y. (2021b). Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error. *Annals Inst. Stat. Math.*, **73**, 481–517.
- Chen, L.-P., Yi, G.Y., Zhang, Q. & He, W. (2019). Multiclass analysis and prediction with network structured covariates. *J. Stat. Distribut. Appl.*, **6**(6).
- Cheng, J., Li, T., Levina, E. & Zhu, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Stat.*, **26**, 367–378.
- Chun, H., Lee, M.H., Kim, S.-H. & Oh, J. (2018). Robust precision matrix estimation via weighted median regression with regularization. *The Canadian J. Stat.*, **46**, 265–278.
- Crane, H. (2018). *Probabilistic Foundations of Statistical Network Analysis*. Boca Raton, FL: CRC Press.
- Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, **51**, 157–172.
- Dalal, O. & Rajaratnam, B. (2017). Sparse Gaussian graphical model estimation via alternating minimization. *Biometrika*, **104**, 379–395.
- Danaher, P., Wang, P. & Witten, D.M. (2014). The joint graphical LASSO for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Series B: Statistical Methodology*, **76**, 373–397.



- De Angelis, P.L., Pardalos, P.M. & Toraldo, G. (1997). Quadratic Programming with Box Constraints. In *Developments in Global Optimization. Nonconvex Optimization and Its Applications*, Eds. I.M. Bomze, T. Csendes, R. Horst & P.M. Pardalos, Vol. **18**. Boston, MA: Springer. [https://doi.org/10.1007/978-1-4757-2600-8\\_5](https://doi.org/10.1007/978-1-4757-2600-8_5)
- Edwards, D. (2000). *Introduction to Graphical Modelling*. New York: Springer.
- Fan, J., Liu, H., Ning, Y. & Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc., Series B*, **79**, 405–421.
- Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical LASSO. *Bio-statistics*, **9**, 432–441.
- Golub, L., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Gough, D., Borgatti, S., Everett, M. & Johnson, J. (2018). *Analyzing Social Networks*. Sage Publications.
- Grimes, T., Potter, S.S. & Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. *Sci. Rep.*, **9**, 5479.
- Guha, N., Baladandayuthapani, V. & Mallick, B.K. (2020). Quantile graphical models: a Bayesian approach. *J. Mach. Learn. Res.*, **21**, 1–47.
- Guillot, D., Rajaratnam, B., Rolfs, B.T., Maleki, A. & Wong, I. (2012). Iterative thresholding algorithm for sparse inverse covariance estimation. arXiv:1211.2532v3.
- Guo, J., Levina, E., Michailidis, G. & Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, **98**, 1–15.
- Han, F. & Liu, H. (2013). Transition matrix estimation in high dimensional time series. In *ICML:2013: Proceedings of the 30th International Conference on Machine Learning*, Eds. D. McAllester & S. Dasgupta, pp. 172–180. Atlanta, Georgia, USA: PMLR.
- Handcock, M.S. (2003). Assessing degeneracy in statistical models of social networks. In *Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington*. Available online at <https://csss.uw.edu/Papers/wp39.pdf>
- Haslbeck, J.M.B. & Waldorp, L.J. (2020). Estimating time-varying mixed graphical models in high-dimensional data. *J. Stat. Softw.*, **93**, 1–46.
- Hastie, T., Tibshirani, R. & Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer.
- Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning with Sparsity the Lasso and Generalizations*. New York: Chapman and Hall/CRC.
- He, S., Yin, J., Li, H. & Wang, X. (2014). Graphical model selection and estimation for high-dimensional tensor data. *J. Multivar. Anal.*, **128**, 165–185.
- He, W., Yi, G. Y., & Chen, L.-P. (2019). Support vector machine with graphical network structures in features. In *Proceedings, Machine Learning and Data Mining in Pattern Recognition, 15th International Conference on Machine Learning and Data Mining, MLDM 2019*, Vol. **II**, pp. 557–570.
- Højsgaard, S., Edwards, D. & Lauritzen, S. (2012). *Graphical Models with R*. New York: Springer.
- Hsieh, C.-J., Sustik, M.A., Dhillon, I.S. & Ravikumar, P. (2014). QUIC Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, **15**, 2911–2947.
- Huang, K. (1987). *Statistical Mechanics*. New York: Wiley.
- Hunter, D.R. & Handcock, M.S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Stat.*, **15**, 565–583.
- Ising, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z. Physik*, **31**, 253–258.
- Jordan, M.I. (1999). *Learning in Graphical Models*. Boston: MIT Press.
- Jordan, M.I. (2004). Graphical models. *Stat. Sci.*, **19**, 140–155.
- Katenka, N. & Kolaczyk, E.D. (2012). Inference and characterization of multi-attribute networks with application to computational biology. *Annals Appl. Stat.*, **6**, 1068–1094.
- Khare, K., Oh, S.-Y. & Rajaratnam, B. (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. Royal Stat. Society, Series B*, **77**, 803–825.
- Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.
- Kolaczyk, E.D. (2017). *Topics at the Frontier of Statistics and Network Analysis: (Re)visiting the Foundations*. New York: Cambridge University Press.
- Kolaczyk, E.D. & Csárdi, G. (2014). *Statistical Analysis of Network Data with R*. New York: Springer.
- Kolar, M., Liu, H. & Xing, E.P. (2014). Graph estimation from multi-attribute data. *J. Machine Learn. Res.*, **15**, 1713–1750.
- Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. New York: MIT Press.
- Kumar, S., Lun, X.-K., Bodenmiller, B., Rodríguez Martínez, M. & Koeppl, H. (2020). Stabilized reconstruction of signaling networks from single-cell cue-response data. *Sci. Rep.*, **10**, 1233. <https://doi.org/10.1038/s41598-019-56444-5>



- Lafferty, J., Liu, H. & Wasserman, L. (2012). Sparse nonparametric graphical models. *Stat. Sci.*, **27**, 519–537.
- Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.
- Lee, J. & Hastie, T.J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Stat.*, **24**, 230–253.
- Lee, W. & Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivar. Anal.*, **111**, 241–255.
- Lee, W. & Liu, Y. (2015). Joint estimation of multiple precision matrices with common structures. *J. Mach. Learn. Res.*, **16**, 1035–1062.
- Leng, C. & Tang, C.Y. (2012). Sparse matrix graphical models. *J. Am. Stat. Assoc.*, **107**, 1187–1200.
- Li, B., Chun, H. & Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Am. Stat. Assoc.*, **107**, 152–167.
- Lin, J., Basu, S., Banerjee, M. & Michailidis, G. (2016). Penalized maximum likelihood estimation of multi-layered Gaussian graphical models. *J. Machine Learn. Res.*, **17**, 5097–5147.
- Liu, W. (2017). Structural similarity and difference testing on multiple sparse Gaussian graphical models. *Ann. Stat.*, **45**, 2680–2707.
- Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Annals Stat.*, **40**, 2293–2326.
- Liu, H., Lafferty, J. & Wasserman, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Machine Learn. Res.*, **10**, 2295–2328.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
- Ma, J. & Michailidis, G. (2016). Joint structural estimation of multiple graphical models. *J. Mach. Learn. Res.*, **17**, 5777–5824.
- Ma, S., Xue, L. & Zou, H. (2013). Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Comput.*, **25**, 2172–2198.
- Maathuis, M., Drton, M., Lauritzen, S. & Wainwright, M. (2019). *Handbook of Graphical Models*. Boca Raton, FL: CRC Press.
- Mazumder, R. & Hastie, T. (2012a). The graphical LASSO: new insights and alternatives. *Electron. J. Stat.*, **6**, 2125–2149.
- Mazumder, R. & Hastie, T. (2012b). Exact covariance thresholding into connected components for large-scale graphical LASSO. *J. Mach. Learn. Res.*, **13**, 781–794.
- Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *The Annals Stat.*, **34**, 1436–1462.
- Meng, Z., Eriksson, B. & Hero, III A.O. (2014). Learning latent variable Gaussian graphical models. In *Proceedings of the 31st International Conference on Machine Learning*: Beijing, China. JMLR: W&CP volume 32.
- Newman, M. (2018). *Networks*. New York: Oxford University Press.
- Okabe, A. 2017. Spatial analysis along networks. In *Encyclopedia of GIS*, pp. 1938–1948. Cham: Springer.
- Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.*, **104**, 735–746.
- Ravikumar, P., Wainwright, M.J. & Lafferty, J. (2010). High-dimensional Ising model selection using  $\ell_1$ -regularized logistic regression. *Ann. Stat.*, **38**, 1287–1319.
- Ravikumar, P., Wainwright, M.J., Raskutti, G. & Yu, B. (2011). High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electron. J. Stat.*, **5**, 935–980.
- Rocke, D.M. & Durbin, B. (2001). A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.
- Rothman, A., Levina, E. & Zhu, J. (2010). Sparse multivariate regression with covariate estimation. *J. Comput. Graph. Stat.*, **19**, 947–26.
- Roy, A. & Dunson, D.B. (2020). Nonparametric graphical model for counts. *J. Mach. Learn. Res.*, **21**, 1–22.
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. & Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.
- Scutari, M. & Denis, J.-B. (2014). *Bayesian Networks: With Examples in R*. New York: Chapman and Hall/CRC.
- Sedgewick, A.J., Shi, I., Donovan, R.M. & Benos, P.V. (2016). Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinform.*, **17**(1), 12.
- Sinoquet, C. & Mourad, R. (2014). *Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics*. Oxford: Oxford University Press.
- Skripnikov, A. & Michailidis, G. (2019). Regularized joint estimation of related vector autoregressive models. *Comput. Stat. Data Anal.*, **139**, 164–177.
- Snijders, T.A.B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Social Struct.*, **3**, 1–40.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Series B*, **58**, 267–288.

- van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.
- Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. New York: Cambridge University Press.
- Wan, Y.-W., Allen, G.I., Baker, Y., Yang, E., Ravikumar, P., Anderson, M. & Liu, Z. (2016). XMRF: an R package to fit Markov Networks to high-throughput genetics data. *BMC Syst. Biology*, **10**(Suppl 3), 69.
- Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Stat. Sin.*, **25**, 831–851.
- Wilson, G.T., Reale, M. & Haywood, J. (2016). *Models for Dependent Time Series*. Boca Raton, FL: CRC Press.
- Witten, D.M., Friedman, J.H. & Simon, N. (2011). New insights and faster computations for the graphical LASSO. *J. Comput. Graph. Stat.*, **20**, 892–900.
- Won, J.-H., Lim, J., Kim, S.-J. & Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *J. R. Stat. Society, Series B*, **75**, 427–450.
- Xie, Y., Liu, Y. & Valdar, W. (2016). Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika*, **103**, 493–511.
- Xue, L. & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Stat.*, **40**, 2541–2571.
- Yan, T., Leng, C. & Zhu, J. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Annals Statistics*, **44**, 31–57.
- Yan, T. & Xu, J. (2013). A central limit theorem in the  $\beta$ -model for undirected random graphs with a diverging number of vertices. *Biometrika*, **100**, 519–524.
- Yan, T., Zhao, Y. & Qin, H. (2015). Asymptotic normality in the maximum entropy distributions on graphs with an growing number of parameters. *J. Multivariate Anal.*, **133**, 61–76.
- Yang, E., Baker, Y., Ravikumar, P., Allen, G.I. & Liu, Z. (2014). Mixed graphical models via exponential families. In *Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Eds. S. Kaski & J. Corande, Reykjavik, Iceland.
- Yang, E., Ravikumar, P., Allen, G.I. & Liu, Z. (2015). Graphical models via univariate exponential family distribution. *J. Mach. Learn. Res.*, **16**, 3813–3847.
- Yang, Z., Ning, Y. & Liu, H. (2018). On semiparametric exponential family graphical models. *J. Machine Learn. Res.*, **19**, 1–59.
- Yi, G.Y. (2017). *Statistical Analysis with Measurement Error and Misclassification: Strategy, Method and Application*. New York: Springer.
- Yin, J. & Li, H. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by  $\ell_1$ -penalization. *J. Multivar. Anal.*, **116**, 365–381.
- Yörük, E., Ochs, M.F., Geman, D. & Younes, L. (2011). A comprehensive statistical model for cell signaling and protein activity inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 592–606.
- Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.
- Zhao, T., Liu, H., Roeder, K., Lafferty, J. & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.*, **13**, 1059–1062.
- Zhou, S. (2014). Gemini graph estimation with matrix variate normal instances. *Annals Stat.*, **42**, 532–562.
- Zhou, S., Rütimann, P., Xu, M. & Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.*, **12**, 2975–3026.
- Zhou, S., van de Geer, S. & Bühlmann, P. (2009). Adaptive LASSO for high dimensional regression and Gaussian graphical modeling. arXiv:0903.2515.
- Zhu, Y., Shen, X. & Pan, W. (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinforma.*, **10**, S21. <https://doi.org/10.1186/1471-2105-10-S1-S21>
- Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429.

[Received July 2022; accepted August 2023]