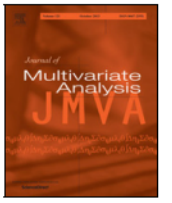
Journal of Multivariate Analysis 202 (2024) 105298

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Estimation of multiple networks with common structures in heterogeneous subgroups

Xing Qin [a], Jianhua Hu [b], Shuangge Ma [c], Mengyun Wu [b,*]

[a] School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai, China
[b] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[c] Department of Biostatistics, Yale School of Public Health, New Haven, USA

ARTICLE INFO

AMS 2020 subject classifications:
primary 62H30
secondary 62H12

Keywords:
Gaussian graphical models
Heterogeneity analysis
High-dimensional data
Network estimation

ABSTRACT

Network estimation has been a critical component of high-dimensional data analysis and can provide an understanding of the underlying complex dependence structures. Among the existing studies, Gaussian graphical models have been highly popular. However, they still have limitations due to the homogeneous distribution assumption and the fact that they are only applicable to small-scale data. For example, cancers have various levels of unknown heterogeneity, and biological networks, which include thousands of molecular components, often differ across subgroups while also sharing some commonalities. In this article, we propose a new joint estimation approach for multiple networks with unknown sample heterogeneity, by decomposing the Gaussian graphical model (GGM) into a collection of sparse regression problems. A reparameterization technique and a composite minimax concave penalty are introduced to effectively accommodate the specific and common information across the networks of multiple subgroups, making the proposed estimator significantly advancing from the existing heterogeneity network analysis based on the regularized likelihood of GGM directly and enjoying scale-invariant, tuning-insensitive, and optimization convexity properties. The proposed analysis can be effectively realized using parallel computing. The estimation and selection consistency properties are rigorously established. The proposed approach allows the theoretical studies to focus on independent network estimation only and has the significant advantage of being both theoretically and computationally applicable to large-scale data. Extensive numerical experiments with simulated data and the TCGA breast cancer data demonstrate the prominent performance of the proposed approach in both subgroup and network identifications.

1. Introduction

Many modern applications often involve analyzing a network structure for a set of high-dimensional variables. For example, biological networks specific to disease contexts explore the patterns of association in molecular data (e.g., genes, proteins, etc.) and play a critical role in understanding the underlying biological progresses [4,16,30]. Several statistical methods have been developed for network estimation. Among them, Gaussian graphical models have been widely employed, where the precision matrix describes the conditional dependencies between variables [8,38]. To capture sparsity patterns in the precision matrix, one strategy is to conduct estimation based on the penalized likelihood function [1,12]. Another strategy is to reduce the estimation of the precision matrix to a collection of sparse regression problems [26], which is not only easy to optimize but also more amenable to theoretical

* Corresponding author.
E-mail address: wu.mengyun@mail.shufe.edu.cn (M. Wu).

https://doi.org/10.1016/j.jmva.2024.105298
Received 15 May 2023; Received in revised form 5 February 2024; Accepted 9 February 2024
Available online 13 February 2024

analysis. Recent methodological developments include CLIME [6], Tiger [24], SCIO [23], and scaled-Lasso [34], among which Tiger is computationally the most efficient due to its square root loss function and asymptotic tuning-free property.

Despite considerable successes, these studies are limited to homogeneous analysis and assume that high-dimensional data has only a single network. In practical applications, it is common that observed data comes from multiple subgroups and has heterogeneous dependencies. For example, in biomedical studies, cancer heterogeneity has received extensive attention, and the biological networks often differ across subtypes of the same cancer while also having certain commonalities. To accommodate both the specific and common information among multiple networks, based on the Gaussian graphical models, significant progresses have been made recently using the group regularization techniques, including the more popular likelihood-based strategies [2,7,14,28] and relatively limited investigations from a sparse linear regression perspective in a column-by-column fashion [25].

However, all the aforementioned approaches rely on prior known subgroup memberships, which are usually not available for practical high-dimensional data with complex and unknown group structures. To accommodate the unknown heterogeneity, the Gaussian mixture model serves as a suitable choice for simultaneously conducting subgroup and network identifications. To this end, [13] develops a joint graphical lasso penalty on multiple precision matrices to extract both homogeneous and heterogeneous information across all subgroups. [15] takes a further step, developing an efficient ECM algorithm and establishing non-asymptotic statistical properties for the estimated networks. Recently, [29] develops a penalized fusion method for heterogeneity analysis based on the Gaussian mixture model, which has the advantage of being capable of automatically determining the number of subgroups. The aforementioned approaches are based on likelihood regularization strategies and are often only applicable to small-scale data with dimensions less than 100.

In this study, we develop a new joint estimation approach for multiple networks with unknown sample heterogeneity. Based on the Gaussian mixture model, we propose estimating multiple precision matrices in a column-by-column manner with a reparameterization technique and introduce a composite minimax concave penalty (MCP) to effectively accommodate both specific and common structures of networks. This is significantly different from the existing heterogeneous network estimation publications [15,29] (which directly maximize a penalized likelihood with respect to large matrices) and has the advantage of automatically imposing an adaptive penalty on the parameters, resulting in scale-invariant estimators and potential tuning-insensitive properties. This study is much more than an extension of the homogeneous column-wise Gaussian graphical model [24,25]. Specifically, significantly advancing from [24] which is based on the square root loss function, the proposed approach is based on ordinary square loss, enjoying simplicity and achieving optimization convexity. In addition, significantly advancing from [25], the proposed approach can provide equivalent estimates under scaling in mixture models, which is much more important than in homogeneous analysis. The proposed approach allows theoretical studies to focus on independent network estimation only (disregarding the diagonal elements of precision matrices) and can take advantage of parallel computation to achieve more efficiency, making it theoretically and computationally feasible for large-scale data. Overall, this study provides a practically useful tool for the joint estimation research of multiple networks from heterogeneous subgroups.

2. Methods

Suppose that there are n independent subjects, which come from K subgroups. For the ith subject, denote $\boldsymbol{x}_i = \left(x_{i,1}, \ldots, x_{i,p}\right)^\top$ as the p-dimensional vector of predictor measurements and $C_i \in \{1, \ldots, K\}$ as the subgroup assignment. Assume that

$$\boldsymbol{x}_i \mid C_i = k \sim \mathcal{N}_p\left(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\right), \tag{1}$$

where $\boldsymbol{\mu}_k = \left(\mu_{k,1}, \ldots, \mu_{k,p}\right)^\top$ and $\boldsymbol{\Omega}_k = \left(\omega_{k,\ell j}\right)_{p\times p}$ are the mean vector and precision matrix for the kth subgroup, respectively. This Gaussian graphical model has been popular in existing network analysis publications. For the kth subgroup, the relationships between the predictors in the network are measured using the subgroup-specific precision matrix $\boldsymbol{\Omega}_k$, which describes the conditional independence between any two predictors given the rest. Specifically, if $\omega_{k,\ell j} \neq 0$, there is a connection between predictors ℓ and j for the kth subgroup, and otherwise if $\omega_{k,\ell j} = 0$.

Based on (1), when $C_i = k$, the conditional distribution of $x_{i,j}$ given $\boldsymbol{x}_{i,\setminus j} = \left\{x_{i,\ell} : \ell \neq j\right\}_{(p-1)\times 1}$ is

$$x_{i,j} \mid \boldsymbol{x}_{i,\setminus j}; C_i = k \sim \mathcal{N}\left(\mu_{k,j} - \sum_{\ell \neq j} \omega_{k,j\ell}/\omega_{k,jj}\left(x_{i,\ell} - \mu_{k,\ell}\right), 1/\omega_{k,jj}\right). \tag{2}$$

For $\ell \neq j$, define

$$\ddot{\beta}_{k,j,\ell} := -\omega_{k,j\ell}/\omega_{k,jj},\ \ddot{\beta}_{k,j,j} := \mu_{k,j} - \sum_{\ell \neq j} \ddot{\beta}_{k,j,\ell}\mu_{k,\ell},\ \text{and}\ \sigma^2_{k,j} := 1/\omega_{k,jj}. \tag{3}$$

Then (2) can be rewritten as:

$$x_{i,j} \mid \boldsymbol{x}_{i,\setminus j}; C_i = k \sim \mathcal{N}\left(\ddot{\beta}_{k,j,j} + \boldsymbol{x}_{i,\setminus j}^\top \ddot{\boldsymbol{\beta}}_{k,j,\setminus j}, \sigma^2_{k,j}\right), \tag{4}$$

where $\ddot{\boldsymbol{\beta}}_{k,j,\setminus j} = \left\{\ddot{\beta}_{k,j,\ell} : \ell \neq j\right\}_{(p-1)\times 1}$. Based on (3), estimation of $\boldsymbol{\Omega}_k$'s can be achieved by considering the conditional distribution (4) for $j \in \{1, \ldots, p\}$, where $\omega_{k,j\ell} \neq 0$ corresponds to $\ddot{\beta}_{k,j,\ell} \neq 0$ for $\ell \neq j$.

In practice, the subgroup assignments C_i's are not always observed. To accommodate the unknown sample heterogeneity, we consider the following conditional Gaussian mixture model:

$$x_{i,j} \mid \boldsymbol{x}_{i,\setminus j} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\ddot{\beta}_{k,j,j} + \boldsymbol{x}_{i,\setminus j}^\top \ddot{\boldsymbol{\beta}}_{k,j,\setminus j}, \sigma^2_{k,j}\right),$$

where $\pi_k = Pr(C_i = k)$ is the probability that a subject belongs to the kth subgroup with $\sum_{k=1}^{K} \pi_k = 1$. For more effective heterogeneous network estimation while accommodating the common structure across the K networks, we consider reparameterization for each $k \in \{1, \ldots, K\}$ and $j, \ell \in \{1, \ldots, p\}$:

$$\tau_{k,j} = 1/\sigma_{k,j}, \beta_{k,j,\ell} = \ddot{\beta}_{k,j,\ell}/\sigma_{k,j}, \tag{5}$$

and propose the following penalized objective function for each $j \in \{1, \ldots, p\}$:

$$L_n(\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k g\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}^{\top} \boldsymbol{\beta}_{k,j,\setminus j} + \beta_{k,j,j}, \tau_{k,j}\right) \right\} - n \sum_{\ell \neq j} \rho \left\{ \sum_{k=1}^{K} \rho\left(\left|\beta_{k,j,\ell}\right|; \lambda, \gamma\right); \lambda, \frac{K\lambda\gamma}{2} \right\}, \tag{6}$$

where $\boldsymbol{\theta}_j = \mathrm{vec}\left(\pi_1, \ldots, \pi_K, \boldsymbol{\alpha}_j\right)$ is the vector of the overall unknown parameters in the jth subproblem with $\boldsymbol{\alpha}_j = \mathrm{vec}\left(\tau_{1,j}, \ldots, \tau_{K,j}, \boldsymbol{\beta}_j\right)$ and $\boldsymbol{\beta}_j = \mathrm{vec}\left(\beta_{1,j,j}, \boldsymbol{\beta}_{1,j,\setminus j}, \ldots, \beta_{K,j,j}, \boldsymbol{\beta}_{K,j,\setminus j}\right)$, $g\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}^{\top} \boldsymbol{\beta}_{k,j,\setminus j} + \beta_{k,j,j}, \tau_{k,j}\right) = \ln \tau_{k,j} - \left(\tau_{k,j} x_{i,j} - \beta_{k,j,j} - \boldsymbol{x}_{i,\setminus j}^{\top} \boldsymbol{\beta}_{k,j,\setminus j}\right)^2 /2$, and $\rho(v; \lambda, \gamma) = \lambda \int_0^v \left\{1 - (\lambda\gamma)^{-1} x\right\}_+ dx$ is the minimax concave penalty (MCP) with tuning parameter λ and regularization parameter γ. Here, following the existing literature [3], the tuning parameter of the outer penalty is chosen to be $K\lambda\gamma/2$, ensuring that the group-level penalty reaches a maximum if and only if each of its components is at its maximum.

The proposed estimation procedure has been motivated by the following considerations. In (6), the first term is a reparameterized version of the log-likelihood function based on the distribution of $x_{i,j}$ given $\boldsymbol{x}_{i,\setminus j}$. Different from the studies that estimate the precision matrix based on the joint distribution of $\boldsymbol{x}_i$ directly, the proposed conditional distribution-based strategy can reduce the estimation problem of a large matrix to a collection of sparse linear regression problems, enjoying not only great computational efficiency but also satisfactory theoretical properties. Specifically, as demonstrated in [33], the scale-invariance considerations and theoretical results indicate the necessity of taking penalty levels proportional to noise levels in high-dimensional problems. Thus, taking advantage of the column-wise strategy, we utilize reparameterization to automatically scale the penalty to the noise level and make the proposed estimator scale-invariant [32,42]. Different from the existing heterogeneous network analysis studies [13,15,29], such a strategy achieves "adaptive adjustment" of the coefficients based on the data characteristics, resulting in a more accurate estimation. The proposed model is closely related to the scaled sparse linear regression [33], which does not require any knowledge of noise and has comparable theoretical properties to studies under known noise. In contrast, penalizing $\ddot{\beta}_{k,j,\ell}$'s directly on the basis of (4) leads to an indirect effect on the estimation of the scale parameters $\sigma_{k,j}$'s and non-convex optimization, which may result in serious consequences in mixture model analysis. Additionally, such a reparameterization strategy enjoys the advantage that the identification and estimation results are insensitive to the choice of tuning parameters, which simplifies the tuning procedure. In the second term, we apply the composite MCP [3] to capture heterogeneity and homogeneity across all K subgroups. According to the relationship between $\beta_{k,j,\ell}$ and $\omega_{k,j\ell}$, the outer MCP encourages the K estimated precision matrices to share a similar pattern of sparsity to accommodate the potential common network structure across the subgroups, where $\beta_{1,j,\ell}, \ldots, \beta_{K,j,\ell}$ will be simultaneously shrunk to zero or not. On the other hand, the inner MCP imposed on $\beta_{k,j,\ell}$'s amounts to further sparsing each precision matrix to accommodate the specific network structure across the subgroups. Here, different from the sparse group lasso penalty commonly used in the literature [13,15], the adopted composite MCP improves accuracy and interpretability of the model by allowing important relationships to have large coefficients without overshrinking. As our primary goal is to estimate subgroup-specific precision matrices and investigate heterogeneous networks, the intercept terms $\beta_{k,j,j}$'s are not subject to penalization. The model can be easily extended to accommodate sparse learning for the subgroup means as well.

2.1. Computation

We develop an Expectation Maximization (EM) algorithm to optimize objective function (6). Specifically, we introduce a latent variable $z_{ki} = I\{C_i = k\}$, where $I\{\cdot\}$ is the indicator function, and then consider the following complete objective function:

$$Q_n(\boldsymbol{\theta}_j) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ki} \ln \left\{ \pi_k g\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}^{\top} \boldsymbol{\beta}_{k,j,\setminus j} + \beta_{k,j,j}, \tau_{k,j}\right) \right\} - n \sum_{\ell \neq j} \rho \left\{ \sum_{k=1}^{K} \rho\left(\left|\beta_{k,j,\ell}\right|; \lambda, \gamma\right); \lambda, \frac{K\lambda\gamma}{2} \right\}. \tag{7}$$

Denote $\boldsymbol{\Psi} = \mathrm{vec}\left(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K, \pi_1, \ldots, \pi_K\right)$ and $\boldsymbol{\alpha} = \mathrm{vec}\left(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p\right)$. The computation proceeds with the following steps.

1. Initialization: Set iteration time $t = 0$. For each $k \in \{1, \ldots, K\}$, initialize $\pi_k^{(0)} = 1/K$ and $\boldsymbol{\Omega}_k^{(0)} = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is an identity matrix of dimension p. Randomly divide the subjects into K subgroups of an equal size. For each $k \in \{1, \ldots, K\}$, take the mean of each subgroup as the initial value $\boldsymbol{\mu}_k^{(0)}$.
2. In the tth E step, the conditional expectation of (7) with respect to $\boldsymbol{\Psi}^{(t-1)}$ obtained at the $(t-1)$th iteration is

$$\mathrm{E}_{\boldsymbol{\Psi}^{(t-1)}}\left\{Q_n(\boldsymbol{\theta}_j)\right\} = \sum_{i=1}^{n} \sum_{k=1}^{K} \rho_{ki}^{(t)} \ln \left\{ \pi_k g\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}^{\top} \boldsymbol{\beta}_{k,j,\setminus j} + \beta_{k,j,j}, \tau_{k,j}\right) \right\} - n \sum_{\ell \neq j} \rho \left\{ \sum_{k=1}^{K} \rho\left(\left|\beta_{k,j,\ell}\right|; \lambda, \gamma\right); \lambda, \frac{K\lambda\gamma}{2} \right\} \tag{8}$$

for $j \in \{1, \ldots, p\}$. Here $\rho_{ki}^{(t)} = Pr\left(z_{ki} = 1 \mid \boldsymbol{x}_i; \boldsymbol{\Psi}^{(t-1)}\right) = \pi_k^{(t-1)} f_k\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Omega}_k^{(t-1)}\right) \Big/ \left\{ \sum_{\ell=1}^{K} \pi_\ell^{(t-1)} f_\ell\left(\boldsymbol{x}_i; \boldsymbol{\mu}_l^{(t-1)}, \boldsymbol{\Omega}_\ell^{(t-1)}\right) \right\}$ with f_k being a multivariate Gaussian density function with a subgroup-specific mean vector $\boldsymbol{\mu}_k^{(t-1)}$ and precision matrix $\boldsymbol{\Omega}_k^{(t-1)}$, where $\boldsymbol{\mu}_k^{(t-1)}$ and $\boldsymbol{\Omega}_k^{(t-1)}$ are computed based on the estimation of $\boldsymbol{\alpha}$ at the $(t-1)$th iteration and Eqs. (3) and (5) (see the supplementary materials for more details). Here, with the consideration of the symmetry of $\boldsymbol{\Omega}_k$'s, following [24], we compute

$\omega_{k,\ell j}^{(t-1)} = \left(\omega_{k,\ell j}^{(t-1)} + \omega_{k,j\ell}^{(t-1)}\right)/2$. To further ensure the positive definiteness of the precision matrices, a small diagonal matrix is added to $\boldsymbol{\Omega}_k$'s [25].

3. In the tth M step,

- Maximize (8) with respect to π_k and compute $\pi_k^{(t)} = \sum_{i=1}^n \rho_{ki}^{(t)}/n$.
- Maximize (8) with respect to $\boldsymbol{\alpha}_j$. For each $j \in \{1, \ldots, p\}$, the following two steps are carried out sequentially,
  (1) With the other parameters fixed, optimizing $Q_n(\boldsymbol{\theta}_j)$ with respect to $\tau_{k,j}$ for each $k \in \{1, \ldots, K\}$ yields

$$\tau_{k,j}^{(t)} = \arg\min_{\tau_{k,j}} \left\{ \frac{\sum_{i=1}^n \rho_{ki}^{(t)} \left(\tau_{k,j} x_{i,j} - \beta_{k,j,j}^{(t-1)} - \boldsymbol{x}_{i,\setminus j}^\top \boldsymbol{\beta}_{k,j,\setminus j}^{(t-1)}\right)^2}{2} - \sum_{i=1}^n \rho_{ki}^{(t)} \ln \tau_{k,j} \right\} = \frac{\tilde{b}_k^{(t)} + \sqrt{\left(\tilde{b}_k^{(t)}\right)^2 + 4n\tilde{a}_k^{(t)}\pi_k^{(t)}}}{2\tilde{a}_k^{(t)}},$$

where $\tilde{a}_k^{(t)} = \sum_{i=1}^n \rho_{ki}^{(t)} x_{i,j}^2$ and $\tilde{b}_k^{(t)} = \sum_{i=1}^n \rho_{ki}^{(t)} \left(\boldsymbol{x}_{i,\setminus j}^\top \boldsymbol{\beta}_{k,j,\setminus j}^{(t-1)} + \beta_{k,j,j}^{(t-1)}\right) x_{i,j}$.
  (2) With the other parameters fixed, optimizing $Q_n(\boldsymbol{\theta}_j)$ with respect to $\boldsymbol{\beta}_j$ yields that

$$\boldsymbol{\beta}_j^{(t)} = \arg\min_{\boldsymbol{\beta}_j} \left[\sum_{i=1}^n \sum_{k=1}^K \frac{\rho_{ki}^{(t)} \left(\tau_{k,j}^{(t)} x_{i,j} - \beta_{k,j,j} - \boldsymbol{x}_{i,\setminus j}^\top \boldsymbol{\beta}_{k,j,\setminus j}\right)^2}{2} + n \sum_{\ell \neq j} \rho \left\{ \sum_{k=1}^K \rho\left(\left|\beta_{k,j,\ell}\right|; \lambda, \gamma\right); \lambda, \frac{K\lambda\gamma}{2} \right\} \right], \tag{9}$$

which is a weighted linear regression problem with composite MCP and can be realized using the R package `grpreg` [3].

By iterating the E and M steps, convergence is achieved usually within a moderate number of iterations, which is concluded in our numerical study if $\sum_{k=1}^K \left\{ \left\|\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_k^{(t-1)}\right\|_2 / \left\|\boldsymbol{\mu}_k^{(t)}\right\|_2 + \left\|\boldsymbol{\Omega}_k^{(t)} - \boldsymbol{\Omega}_k^{(t-1)}\right\|_F / \left\|\boldsymbol{\Omega}_k^{(t)}\right\|_F \right\} \leq 0.01$ with $\left\|\boldsymbol{\Omega}_k^{(t)}\right\|_F = \sqrt{\sum_{i,j} \left(\omega_{k,ij}^{(t)}\right)^2}$ being the Frobenius norm of $\boldsymbol{\Omega}_k^{(t)}$. To improve performance of the proposed EM algorithm, following published literature [42], we consider multiple random initializations of subjects' subgroup memberships and choose the estimate with the smallest AIC defined below as the final estimate.

Following [3], we set $\gamma = 3$ in the proposed algorithm. For λ, following [7], we adopt the AIC criterion defined as $-2\sum_{i=1}^n \ln\left\{\sum_{k=1}^K \hat{\pi}_k f_k\left(\boldsymbol{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Omega}}_k\right)\right\} + \sum_{k=1}^K 2\hat{s}_k$, where $\hat{\pi}_k$'s, $\hat{\boldsymbol{\mu}}_k$'s, and $\hat{\boldsymbol{\Omega}}_k$'s are the final updates from the proposed algorithm, and $\hat{s}_k = \#\left\{(\ell, j) : \hat{\omega}_{k,\ell j} \neq 0, 1 \leq \ell < j \leq p\right\}$. Given fixed tuning parameters, in each iteration of the proposed EM algorithm, for each $j \in \{1, \ldots, p\}$, updating the regression parameters $\boldsymbol{\beta}_j$ costs at most $O\{nK(p-1)\}$ operations, resulting in a complete cost up to $O\{nKp(p-1)\}$ operations. In comparison, the existing popular SCAN approach [15], which conducts heterogeneous network analysis based on the Gaussian mixture model and EM algorithm, has a computational complexity of $O\left\{nKp^2 + Kp^3\right\}$ in each iteration of the EM algorithm. The proposed approach is computationally more feasible for large datasets with $p > n$. Since the optimization of (8) can be realized separately for $j \in \{1, \ldots, p\}$, we develop a parallel computing strategy for the EM algorithm to further improve efficiency. The R package `MultiNet` that implements the proposed approach in a parallel manner is available at https://github.com/mengyunwu2020/MultiNet.

3. Statistical properties

For a set $\mathcal{S}$, denote $\mathcal{S}^c$ and $|\mathcal{S}|$ as its complement and cardinality, respectively. For a vector $\boldsymbol{v}$, denote $\boldsymbol{v}_{\mathcal{S}}$ as the component of $\boldsymbol{v}$ indexed by $\mathcal{S}$. For a matrix $\boldsymbol{M} = (M_{ij})_{p_1 \times p_2}$, denote $\boldsymbol{M}_{\mathcal{S}_1, \mathcal{S}_2}$ as the submatrix of $\boldsymbol{M}$ indexed by $\mathcal{S}_1$ and $\mathcal{S}_2$ and $\|\boldsymbol{M}\|_\infty = \max_{i \in \{1, \ldots, p_1\}} \sum_{j=1}^{p_2} \left|M_{ij}\right|$ as the maximum induced norm of $\boldsymbol{M}$.

Denote $\boldsymbol{\Omega}_k^0$ as the true precision matrix for the kth subgroup and $\boldsymbol{E}_k^0 = \left\{(\ell, j) : 1 \leq \ell \neq j \leq p, \omega_{k,\ell j}^0 \neq 0\right\}$ as the set of the nonzero off-diagonal elements of $\boldsymbol{\Omega}_k^0$. For the jth subproblem, let the vector of the true parameters be $\boldsymbol{\theta}_j^0 = \text{vec}\left(\pi_1^0, \ldots, \pi_K^0, \tau_{1,j}^0, \ldots, \tau_{K,j}^0, \beta_{1,j,j}^0, \ldots, \beta_{K,j,j}^0, \boldsymbol{\beta}_{1,j,\setminus j}^0, \ldots, \boldsymbol{\beta}_{K,j,\setminus j}^0\right)$, and denote $C_j = \left\{q : \theta_{j,q}^0 \neq 0\right\}$ as the non-zero element set, where $\theta_{j,q}^0$ is the qth element of $\boldsymbol{\theta}_j^0$. In addition, for the jth subproblem and kth subgroup, denote $S_{j,k} = \left\{\ell : \beta_{k,j,\ell}^0 \neq 0 \text{ and } \ell \neq j\right\} = \left\{\ell : \omega_{k,j\ell}^0 \neq 0 \text{ and } \ell \neq j\right\}$. Let $s_0 = \max_{j \in \{1, \ldots, p\}, k \in \{1, \ldots, K\}} |S_{j,k}|$.

Let $\boldsymbol{\theta}_{j,C_j}^* = \text{vec}\left(\pi_1^*, \ldots, \pi_K^*, \tau_{1,j}^*, \ldots, \tau_{K,j}^*, \beta_{1,j,j}^*, \ldots, \beta_{K,j,j}^*, \boldsymbol{\beta}_{1,j,S_{j,1}}^*, \ldots, \boldsymbol{\beta}_{K,j,S_{j,K}}^*\right)$ be the maximizer of the oracle counterpart of objective function (6) defined as:

$$\widetilde{L}_n(\boldsymbol{\theta}_{j,C_j}) = \sum_{i=1}^n \ln f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,C_j}\right) = \sum_{i=1}^n \ln \left\{ \sum_{k=1}^K \pi_k g\left(x_{i,j}; \boldsymbol{x}_{i,S_{j,k}}^\top \boldsymbol{\beta}_{k,j,S_{j,k}} + \beta_{k,j,j}, \tau_{k,j}\right) \right\},$$

where $f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,C_j}\right) = \sum_{k=1}^K \pi_k g\left(x_{i,j}; \boldsymbol{x}_{i,S_{j,k}}^\top \boldsymbol{\beta}_{k,j,S_{j,k}} + \beta_{k,j,j}, \tau_{k,j}\right)$.

To establish the theoretical results of the proposed approach in terms of identification and estimation, we need the following technical assumptions.

Assumption 1. The probability density function $f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)$ has a common support and is identifiable with respect to $\boldsymbol{\theta}_j$. Moreover, the first and second derivatives of $\ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)$ satisfy that, for any $q,\ell\in\{1,\ldots,K(p+2)\}$,

$$E\left\{\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)}{\partial\theta_{j,q}}\bigg|_{\boldsymbol{\theta}_j=\boldsymbol{\theta}_j^0}\right\}=0,$$

$$E\left\{\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)}{\partial\theta_{j,q}}\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)}{\partial\theta_{j,\ell}}\bigg|_{\boldsymbol{\theta}_j=\boldsymbol{\theta}_j^0}\right\}=E\left\{-\frac{\partial^2 \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)}{\partial\theta_{j,q}\partial\theta_{j,\ell}}\bigg|_{\boldsymbol{\theta}_j=\boldsymbol{\theta}_j^0}\right\}.$$

Assumption 2. The Fisher information matrix

$$\mathcal{I}(\boldsymbol{\theta}_{j,C_j})=E\left[\left\{\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_{j,C_j}\right)}{\partial\boldsymbol{\theta}_{j,C_j}}\right\}\left\{\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_{j,C_j}\right)}{\partial\boldsymbol{\theta}_{j,C_j}}\right\}^{\mathrm{T}}\right]$$

is finite and positive definite at $\boldsymbol{\theta}_{j,C_j}=\boldsymbol{\theta}_{j,C_j}^0$. Furthermore, there exist finite constants m_1 and m_2 such that for any $q,\ell\in C_j$,

$$E\left\{\frac{\partial^2 \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_{j,C_j}\right)}{\left(\partial\theta_{j,q}\right)^2}\bigg|_{\boldsymbol{\theta}_{j,C_j}=\boldsymbol{\theta}_{j,C_j}^0}\right\}<m_1, E\left[\left\{\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_{j,C_j}\right)}{\partial\theta_{j,\ell}}\frac{\partial \ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_{j,C_j}\right)}{\partial\theta_{j,q}}\right\}^2\bigg|_{\boldsymbol{\theta}_{j,C_j}=\boldsymbol{\theta}_{j,C_j}^0}\right]<m_2.$$

Assumption 3. There exists an open set $\mathcal{N}_0$ containing the true parameter $\boldsymbol{\theta}_j^0$ and satisfies that, for almost all $\boldsymbol{v}_i=\mathrm{vec}(x_{i,j},\boldsymbol{x}_{i,\backslash j})$ and $\boldsymbol{\theta}_j\in\mathcal{N}_0$, the density $f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)$ admits all third derivatives with respect to $\boldsymbol{\theta}_j$. Moreover, there exist integrable functions $M_1(\boldsymbol{v}_i)$ and $M_2(\boldsymbol{v}_i)$, such that for any $\boldsymbol{\theta}_j\in\mathcal{N}_0$ and $q,\ell,m\in\{1,\ldots,K(p+2)\}$,

$$\left|\frac{\partial^2}{\partial\theta_{j,q}\partial\theta_{j,\ell}}\ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)\right|\le M_1(\boldsymbol{v}_i),\quad \left|\frac{\partial^3}{\partial\theta_{j,q}\partial\theta_{j,\ell}\partial\theta_{j,m}}\ln f\left(x_{i,j};\boldsymbol{x}_{i,\backslash j},\boldsymbol{\theta}_j\right)\right|\le M_2(\boldsymbol{v}_i).$$

Assumption 4. For some constant $b\in(0,1/4)$, $K=O(n^b)$ and $s_0=o(n^{1/4-b})$.

Assumption 5. $Ks_0/(\sqrt{n}\lambda^2)\to 0$ and $n^{a/2-1/2}\sqrt{\ln n}/\lambda^2\to 0$, when $n\to\infty$, where $a\in(0,1/2)$.

Assumption 6. $\min_{k\in\{1,\ldots,K\},j\in\{1,\ldots,p\}}\left\{\left|\beta_{k,j,\ell}^0\right|:\ell\in S_{j,k}\right\}>\gamma\lambda$, and $\theta_{j,q}^0$ is bounded for each $j\in\{1,\ldots,p\}$ and $q\in\{1,\ldots,K(p+2)\}$.

Assumption 7. $\ln(p)=O(n^a)$.

Assumptions 1–3 impose conditions on the mixture distribution and have been commonly assumed in the field of mixture modeling [18,20,42]. Assumption 4 makes a constraint on the sparsity parameter s_0 and allows the number of subgroups K to grow slowly with n. It suggests that, for each sub-regression problem, as long as the number of non-zero parameters (i.e., $O(s_0K)$) is equal, the required sample size n is the same. Similar conditions have also been assumed in published multiple network analysis studies, such as Condition (C1) and assumptions in Proposition 1 in [5]. Assumption 5 restricts the rate of λ relative to the sample size. A similar condition is often assumed in high-dimensional studies [10,42]. Assumption 6 imposes constraints on the true parameters, where the first subcondition restricts the rate by which the nonzero coefficients can be distinguished from zero, and the other one restricts the true parameters to a bounded range, which are also considered in the literature [10,15]. Assumption 7 allows the dimension p to grow exponentially.

Theorem 1. *Under Assumptions 1–4, there exists a strict local maximizer $\boldsymbol{\theta}_{j,C_j}^*$ of $\widetilde{L}_n(\boldsymbol{\theta}_{j,C_j})$ such that $\left\|\boldsymbol{\theta}_{j,C_j}^*-\boldsymbol{\theta}_{j,C_j}^0\right\|=O_p(\sqrt{Ks_0/n})$.*

Theorem 2. *Under Assumptions 1–7, the oracle estimator $\widetilde{\boldsymbol{\theta}}_j$ with $\widetilde{\boldsymbol{\theta}}_{j,C_j}=\boldsymbol{\theta}_{j,C_j}^*$ and $\widetilde{\boldsymbol{\theta}}_{j,C_j^c}=\mathbf{0}$ is a strict local maximizer of $L(\boldsymbol{\theta}_j)$ with probability tending to 1.*

Theorem 3. *Under Assumptions 1–7, there exist a strict local maximizer $\hat{\boldsymbol{\theta}}_j$ of $L_n(\boldsymbol{\theta}_j)$ for each $j\in\{1,\ldots,p\}$, where for each $k\in\{1,\ldots,K\}$, the corresponding estimated precision matrix $\hat{\boldsymbol{\Omega}}_k$ obtained from (5) and the resulting edge set $\hat{E}_k=\left\{(\ell,j):1\le\ell\ne j\le p,\ \text{and}\ \hat{\omega}_{k,\ell j}\ne 0\right\}$ satisfy that*

(i) $\max_{k,j,\ell}\left|\hat{\omega}_{k,j\ell}-\omega^0_{k,j\ell}\right| = O_p\left(\sqrt{Ks_0/n}\right)$, $\sum_{k=1}^K\left\|\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}^0_k\right\|_\infty = O_p\left(\sqrt{K^3s_0^3/n}\right)$, *and* $\sum_{k=1}^K\left\|(\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}^0_k)_{\mathcal{A}}\right\|_F = O_p\left(\sqrt{K^3s^2/n}\right)$,

(ii) *with probability tending to 1,* $\hat{\boldsymbol{E}}_k = \boldsymbol{E}^0_k$.

Here, $\mathcal{A} = \{(\ell, j) : 1 \le \ell \ne j \le p\}$ *is the off-diagonal index set for the* $p \times p$ *precision matrix, and* $s = \max_{k\in\{1,\ldots,K\}}\left|\boldsymbol{E}^0_k\right|$ *is the sparsity parameter of the true precision matrices.*

The proofs of Theorems 1, 2, and 3 are provided in Section 7. In our study, we reduce the estimation problem of a large matrix to a collection of sparse linear regression problems. For each regression problem, we examine the theoretical properties based on the oracle estimator. A similar strategy has been considered in published high-dimensional data analysis studies, such as [10,17,42]. Theorem 1 establishes the estimation consistency of the oracle estimator $\boldsymbol{\theta}^*_{j,C_j}$, where the true sparsity structure is known and thus the number of unknown parameters is in the same order as Ks_0. Theorem 2 then demonstrates that the proposed estimator $\hat{\boldsymbol{\theta}}_j$ is asymptotically as efficient as the oracle one. In Theorem 3, we further study the estimation errors of the precision matrices under the elementwise sup-norm and maximum induced norm, as well as that of the off-diagonal elements under the Frobenius norm, which are useful for the graph recovery from the precision matrices.

Remark 1. As the spectral norm is dominated by the maximum induced norm, the estimation error of the proposed estimators under the spectral norm also follows a rate of $O_p\left(\sqrt{K^3s_0^3/n}\right)$, which is important for the consistency of the eigenvalue and eigenvector estimates and can be further used to analyze theoretical properties of downstream inferences.

Remark 2. Considering the goal of network estimation and taking advantage of the column-wise strategy, the proposed theoretical study can focus on the off-diagonal elements of $\boldsymbol{\Omega}_k$'s. Under the assumption $\ln(p) = O(n^a)$, the accuracy of the off-diagonal elements ($\mathcal{A}$) can reach $\sum_{k=1}^K\left\|(\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}^0_k)_{\mathcal{A}}\right\|_F = O_p\left(\sqrt{K^3s^2/n}\right)$, which depends on the sparsity parameter of the true precision matrices, making the proposed approach applicable to large-scale datasets. In contrast, the existing heterogeneous network analysis approaches, including [13,15,29], rely heavily on the estimation properties of the precision matrices as a whole, which requires consideration of the estimation error of the diagonal elements and needs the assumption $p\ln(p) = o(n)$.

4. Simulation

We first consider $p = 100$ and the number of subgroups $K \in \{2, 3, 4\}$. Two settings for the subgroup sizes are considered, where the first one is a balanced design with 200 subjects in each subgroup, and the second one is an imbalanced design with subgroup sizes being (150, 200), (150, 200, 250), and (155, 185, 215, 245) for $K \in \{2, 3, 4\}$, respectively. The networks for the K subgroups are generated as follows. First, following [7], we simulate each network with ten unconnected subnetworks and consider three settings S1-S3 with different levels of similarity across the subgroups. Specifically, under settings S1 and S2, there are two and five subnetworks where all the subgroups share the same sparsity structures, respectively. Under setting S3, for $K = 2$, eight of the ten subnetworks have the same sparsity structures in all the subgroups; and for $K = 3$ and 4, there are five subnetworks with the same sparsity structures in all the subgroups, and another three subnetworks with the same sparsity structures shared only by the first two subgroups. Second, for each subnetwork, we consider three types of network structure: Power-law network, for which the degree distribution follows a power law; Nearest-neighbor network, where $p/10$ points are first randomly generated on a unit square, and based on the calculated $p/10 \times (p/10-1)/2$ pairwise distances, 2 nearest neighbors of each point are found and connected; and Erdös-Rényi network, where the edge between each pair of nodes is added independently with probability 0.2.

For each $k \in \{1, \ldots, K\}$, the observations of the kth subgroup are generated from $\mathcal{N}_p\left(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k^{-1}\right)$, where the first four elements of $\boldsymbol{\mu}_k$ are non-zero and the rest $p-4$ elements are all zero (details are provided in the supplementary materials, see Appendix), and $\boldsymbol{\Omega}_k$'s are generated based on the networks. Specifically, following [24], for the kth network, we generate an adjacency matrix $\boldsymbol{A}_k$ whose non-zero off-diagonal entries (corresponding to edges) are generated from $\text{Uniform}([-0.6, -0.3] \cup [0.3, 0.6])$ and the diagonal entries are set to be 0. Then, $\boldsymbol{\Omega}_k$ is constructed as $\boldsymbol{\Omega}_k = \boldsymbol{D}\left[\boldsymbol{A}_k + \left\{\left|\lambda_{\min}(\boldsymbol{A}_k)\right| + 0.2\right\}\boldsymbol{I}_p\right]\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix whose diagonal elements are $\left(\boldsymbol{1}_5, 3\boldsymbol{1}_5, \boldsymbol{1}_5, 3\boldsymbol{1}_5, \ldots, \boldsymbol{1}_5, 3\boldsymbol{1}_5\right)$ with $\boldsymbol{1}_d$ being a d-dimensional vector of all ones, and $\lambda_{\min}(\boldsymbol{A}_k)$ is the smallest eigenvalue of $\boldsymbol{A}_k$.

In addition to the proposed approach, five alternatives are also considered. SCAN, which is a heterogeneous network analysis approach based on the penalized log-likelihood of the Gaussian mixture model and the EM-ADMM algorithm [15]. Tiger, which is a single network analysis method via column-wise linear regressions and exploits Lasso based on a square root loss function for sparse network estimation [24]. True+Tiger, True+JGL, and True+JSEM, which apply Tiger, JGL, and JSEM for multiple network estimation based on the true subgroup memberships of subjects. Specifically, Tiger is conducted separately for each subgroup. JGL is the likelihood-based joint graph Lasso for multiple network estimation with known subgroup memberships [7]. JSEM conducts a joint analysis of multiple networks using neighborhood selection based on the group lasso penalty [25]. Both SCAN and True+JGL can be realized using the R package JGL, and Tiger and True+Tiger can be realized using the R package huge. The R code implementing True+JSEM can be downloaded from https://github.com/drjingma/JSEM. Among these approaches, SCAN achieves subgroup identification and multiple network estimation simultaneously, as well as accommodating common structures among networks. It is the most direct competitor for the proposed approach. In contrast, the other alternatives were originally developed for homogeneous network analysis or heterogeneous network analysis with known subgroup memberships. Tiger, True+Tiger, and True+JSEM estimate the sparse precision matrices in a column-by-column fashion. The first two perform sparse regression based

on the square root loss function, but they either ignore the differences or the commonalities across subgroups. True+JSEM utilizes the squared loss function and does not consider the sparsity structure within each subgroup.

To evaluate the performance of different approaches, we consider the following measures: (1) Clustering error (CE) for evaluating subgroup identification performance, which calculates the distance between the estimated and true subgroup memberships $\hat{\varphi}$ and φ, defined as $\mathrm{CE} = (C_n^2)^{-1}\left|\left\{(i,j) : I\left\{\hat{\varphi}\left(\boldsymbol{x}_i\right) = \hat{\varphi}(\boldsymbol{x}_j)\right\} \neq I\left\{\varphi\left(\boldsymbol{x}_i\right) = \varphi(\boldsymbol{x}_j)\right\}, i < j\right\}\right|$; (2) Precision matrix square error (PME) for measuring estimation performance, defined as $\mathrm{PME} = \sum_{k=1}^{K}\left\|\hat{\boldsymbol{\Omega}}_k - \boldsymbol{\Omega}_k^0\right\|_F / K$; (3) True and false positive rates (TPR and FPR) for evaluating network identification performance, defined as $\mathrm{TPR} = \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0 \neq 0, \hat{\omega}_{k,\ell j} \neq 0\right\} \Big/ \sum_{\ell<j} I\left\{\omega_{k,\ell j}^0 \neq 0\right\}\right] / K$ and $\mathrm{FPR} = \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0 = 0, \hat{\omega}_{k,\ell j} \neq 0\right\} \Big/ \sum_{\ell<j} I\left\{\omega_{k,\ell j}^0 = 0\right\}\right] / K$.

For each scenario, 100 replicates are conducted. For a fair comparison, for all approaches, the true value of K is considered. The results under the three types of network structure for $K = 3$ are shown in Tables 1–3, respectively. The rest of the results for $K = 2$ and $K = 4$ are provided in the supplementary materials. It can be observed that the proposed approach has superior or competitive performance compared to SCAN in terms of subgroup identification accuracy under all the simulation scenarios. The improvement is more significant under the scenarios with more complex network structures (e.g. power-law network), lower levels of subgroup differences (e.g. S3), or a more imbalanced sample design. For example, for $K = 3$ and the power-law network (Table 1), under the scenario with setting S3 and an imbalanced design, the median CEs are 0.000 (proposed) and 0.166 (SCAN), respectively. Additionally, the proposed approach also performs better in network estimation and identification accuracy. It is able to identify most TPs while keeping FPs much lower than the alternatives. For example, for $K = 3$ and the Erdös-Rényi network (Table 3), under the scenario with setting S2 and a balanced design, the proposed approach has (PME, TPR, FPR)=(16.751, 0.987, 0.039), while (26.975, 0.772, 0.076) for SCAN, (31.025, 0.789, 0.169) for Tiger, (21.104, 0.870, 0.075) for True+Tiger, (20.795, 0.901, 0.184) for True+JGL, and (16.403, 0.965, 0.087) for True+JSEM.

In general, under scenarios with a smaller number of subgroups ($K = 2$), performance of all approaches is significantly improved, with the proposed approach still performing the best and having better performance with increasing similarity across networks (from S1 to S3). The superiority of the proposed approach becomes more prominent under scenarios with a larger number of subgroups, suggesting the validity of the proposed strategy under complex situations. Tiger tends to perform the worst as it ignores heterogeneity. Comparatively, SCAN can simultaneously conduct subgroup identification and multiple network estimation, thus having better performance. Under ideal scenarios with true subgroup memberships, True+Tiger, True+JGL, and True+JSEM perform the second best, with True+JSEM being the most prominent as it conducts joint network analysis with an effective sparse linear programming perspective. In most cases, benefiting from the satisfactory scale-invariant property, the proposed approach, although with unknown sample heterogeneity, can exhibit more accurate network identification than these methods depending on true subgroup memberships. In summary, our approach can effectively capture both shared and unique network structures in heterogeneous network data across diverse scenarios with various degrees of within-group similarity, network structures, and number of subgroups.

In addition to the above analyses, we investigate scenarios where the predictors are higher-dimensional, with $p = 500$ and $p = 1000$. The detailed settings and results (Table S7) are provided in the supplementary materials, where the simulated networks for $p = 500$ are denser than those for $p = 1000$. As it is not computationally feasible to conduct analysis with SCAN for scenarios with $p = 1000$, the corresponding results are not available. For larger-scale data, the performance of those regularized likelihood-based approaches such as SCAN and True+JGL decays as expected, especially for SCAN, whose performance gets dramatically worse and cannot afford the computational cost in the case of $p = 1000$. However, the network analysis approaches based on column-wise linear regressions continue to maintain their satisfactory performance. Under the scenarios with a denser network and high dimension ($p = 500$), the proposed approach exhibits slightly inferior performance compared to True+JSEM, as it is more difficult to accurately identify subgroup memberships for subjects, resulting in reduced network identification accuracy. Under the scenario with sparser networks ($p = 1000$), although with a high dimension, the proposed approach still behaves more favorably. Even when the networks exhibit a higher degree of similarity (S3), the TPR of the proposed approach drops slightly due to the misclassification of subjects, but it still has a smaller number of false positives for the networks.

To gain a deeper insight into the benefit of considering common structures in different networks, we compare the proposed approach with an ad-hoc approach called "ad-hoc MCP" that estimates each network separately using only MCP based on the identified subgroups with the proposed approach. The comparison results under the scenario with $K = 3$ are provided in Supplementary Table S8, which demonstrates that the proposed approach has superior identification and estimation performance compared to the ad-hoc MCP method. Furthermore, we take one replicate under the scenario with setting S3, a balanced design, and the power-law network as an example, and provide the heatmaps of the true sparse structures and estimated results with the proposed approach and the ad-hoc MCP approach in Supplementary Figs. S1 and S2. It is observed that the proposed approach can more accurately identify the true sparse structures and more effectively accommodate the common structures across different networks.

4.1. Computer time

To examine the computational superiority of the proposed approach, we conduct simulations with various values of (n, p, K), which are all implemented on a computer with an Intel Core i7 processor and 24 GB of RAM. The computer time results of the proposed approach and the alternatives with fixed tuning parameters are reported in Supplementary Table S9. In general, analysis

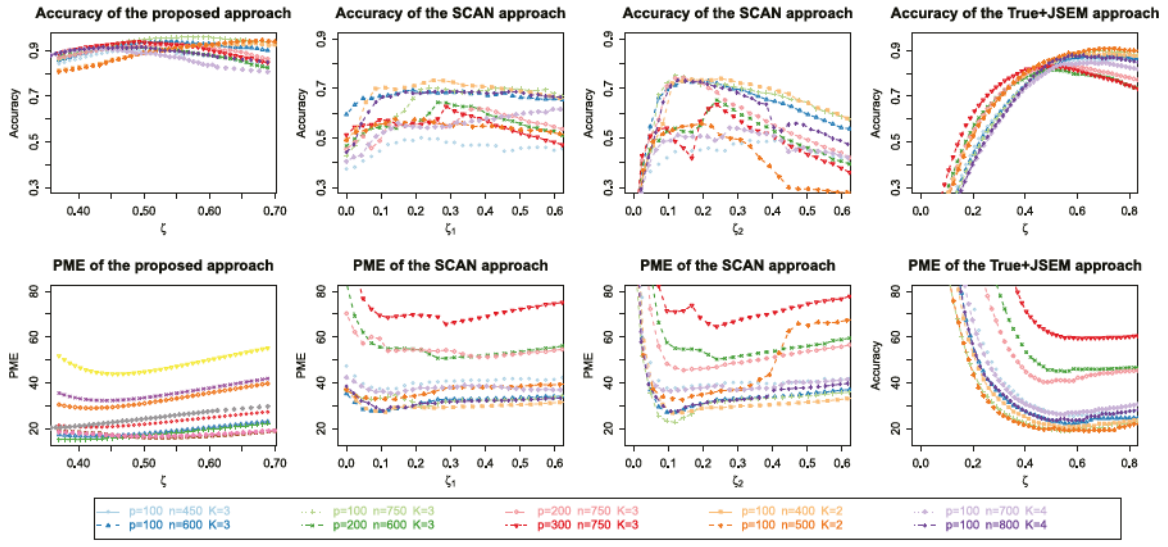


Fig. 1. Column one: Values of Accuracy and PME as a function of ζ for the proposed approach; Columns two and three: Values of Accuracy and PME as a function of ζ_1 for the individual-level sparsity parameter which has form $\zeta_1\sqrt{\ln\{K(p-1)\}/n}$, and ζ_2 for the group-level sparsity parameter which has form $\zeta_2\sqrt{\ln(p-1)/n}$ for the SCAN approach; Column four: Values of Accuracy and PME as a function of ζ for the group-level sparsity parameter which has form $\zeta\sqrt{\ln(p-1)/n}$ for the True+JSEM approach, under the scenario with the power-law network and setting S1.

with the proposed approach is observed to take more time than Tiger and those methods based on known subgroup memberships, which is due to the fact that our approach aims to perform both subgroup and multiple network identifications, while the others either only analyze a single network or rely on prior subgroup information. Compared to the most direct competitor SCAN, which is also based on the Gaussian mixture model and EM algorithm, the proposed approach is significantly faster, especially for large-scale data. For example, under the scenario with $n = 600$, $p = 500$, and $K = 3$, the average computer time is 1794.367 (proposed) and 5169.790 (SCAN) seconds. When the dimension increases to 1000, SCAN is computationally infeasible (more than 24 h) even with a sample size of $n = 300$. However, the proposed algorithm is still computationally affordable, with the computer time being around 2.6 h when $n = 1000$.

The computational cost of the proposed approach can be significantly reduced by using parallel computing, benefiting from the column-wise strategy. For $K = 3$, we compare the computational cost of the single-thread c_{single} and its parallel version $c_{parallel}$ under the scenarios with various sample sizes, dimensions, and number of cores. The cost ratio $r_c = c_{parallel}/c_{single}$ as a function of dimension is provided in Supplementary Fig. S3. As can be observed, the computer time of the proposed approach speeds up over the single-thread version of the algorithm, especially for high-dimensional settings and the use of multi-core processors. The corresponding costs of the proposed analysis with eight cores are also provided in Supplementary Table S9. For high-dimensional data, the proposed approach can sometimes be as efficient as the True+JGL and True+JSEM methods. For example, under the scenario with $n = 300$, $p = 1000$, and $K = 2$, the average computer time is 172.217 s (proposed with parallel), 678.119 s (True+JGL), and 553.785 s (True+JSEM).

4.2. Tuning-insensitive regularization path

We further numerically examine the tuning-insensitive properties of the proposed approach. Motivated by [24,33], for finite samples, we consider the tuning parameter with the form $\zeta\sqrt{\ln\{K(p-1)\}/n}$, where ζ is a positive constant independent of all unknown parameters, and introduce the graph recovery accuracy: Accuracy = TPR − FPR. Taking the simulation scenario under the power-law network and setting S1 as an example, we examine the values of Accuracy and PME as a function of ζ for the proposed approach, as well as SCAN and True+JSEM, in Fig. 1. It is observed that for the proposed approach, the regularization paths are flat without a significant change when ζ lies in the range of $(0.4, 0.7)$, which suggests that the proposed approach is empirically insensitive to the tuning parameter λ. In contrast, for SCAN and True+JSEM, a larger range of (ζ_1, ζ_2) (or ζ) is required in the search of the optimal value, and the paths present more irregular changes, which indicates more sensitivity of SCAN and True+JSEM to the choice of tuning parameters. In general, the proposed approach is easier to implement than SCAN and True+JSEM to find reasonable tuning parameters.

5. Data analysis

We focus on reconstructing gene networks for cancer, which is an important task for better understanding the underlying biological processes. Specifically, we analyze breast cancer data from The Cancer Genome Atlas (TCGA), where the underlying

Table 1
Simulation results under the scenarios with the power-law networks and $K = 3$. In each cell, we show the median (median absolute deviation) of the CE, PME, TPR and FPR values based on 100 replicates. CE $= (C_n^2)^{-1}\left|\{(i,j): I\{\hat{\varphi}(\boldsymbol{x}_i)=\hat{\varphi}(\boldsymbol{x}_j)\}\neq I\{\varphi(\boldsymbol{x}_i)=\varphi(\boldsymbol{x}_j)\}, i<j\}\right|$, PME $= \sum_{k=1}^{K}\left\|\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}_k^0\right\|_F/K$, TPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\{\omega_{k,\ell j}^0\neq 0, \hat{\omega}_{k,\ell j}\neq 0\}\Big/\sum_{\ell<j} I\{\omega_{k,\ell j}^0\neq 0\}\right]/K$, and FPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\{\omega_{k,\ell j}^0=0, \hat{\omega}_{k,\ell j}\neq 0\}\Big/\sum_{\ell<j} I\{\omega_{k,\ell j}^0=0\}\right]/K$.

Approach	CE	PME	TPR	FPR
S1 with the balanced design				
Proposed	0.000(0.000)	20.377(1.058)	0.961(0.004)	0.037(0.001)
SCAN	0.001(0.001)	32.988(0.207)	0.757(0.006)	0.066(0.002)
Tiger	–	36.751(0.268)	0.763(0.016)	0.193(0.004)
True+Tiger	–	28.146(0.608)	0.832(0.008)	0.067(0.002)
True+JGL	–	27.372(0.539)	0.888(0.008)	0.181(0.003)
True+JSEM	–	22.451(0.635)	0.943(0.006)	0.084(0.003)
S2 with the balanced design				
Proposed	0.000(0.000)	19.884(1.528)	0.957(0.010)	0.036(0.002)
SCAN	0.000(0.000)	28.831(0.359)	0.791(0.009)	0.061(0.002)
Tiger	–	34.823(0.302)	0.779(0.012)	0.183(0.006)
True+Tiger	–	25.802(0.289)	0.848(0.005)	0.064(0.001)
True+JGL	–	25.367(0.453)	0.910(0.004)	0.172(0.003)
True+JSEM	–	20.182(0.549)	0.960(0.005)	0.070(0.007)
S3 with the balanced design				
Proposed	0.000(0.000)	21.495(0.770)	0.963(0.006)	0.037(0.001)
SCAN	0.000(0.000)	30.745(0.258)	0.778(0.008)	0.058(0.001)
Tiger	–	36.745(0.320)	0.739(0.012)	0.189(0.005)
True+Tiger	–	28.153(0.319)	0.839(0.006)	0.063(0.001)
True+JGL	–	26.963(0.559)	0.907(0.008)	0.168(0.004)
True+JSEM	–	19.588(0.497)	0.973(0.005)	0.068(0.004)
S1 with the imbalanced design				
Proposed	0.000(0.000)	22.823(1.167)	0.952(0.007)	0.037(0.001)
SCAN	0.002(0.002)	35.378(0.475)	0.739(0.007)	0.058(0.002)
Tiger	–	38.486(0.219)	0.756(0.011)	0.190(0.005)
True+Tiger	–	29.594(0.574)	0.837(0.005)	0.064(0.001)
True+JGL	–	28.999(0.487)	0.900(0.007)	0.173(0.002)
True+JSEM	–	24.722(0.767)	0.940(0.007)	0.079(0.006)
S2 with the imbalanced design				
Proposed	0.000(0.000)	21.155(1.522)	0.949(0.009)	0.037(0.001)
SCAN	0.086(0.084)	31.318(2.751)	0.766(0.025)	0.060(0.005)
Tiger	–	34.915(0.306)	0.796(0.012)	0.187(0.005)
True+Tiger	–	26.661(0.522)	0.846(0.005)	0.067(0.001)
True+JGL	–	26.007(0.662)	0.906(0.006)	0.173(0.003)
True+JSEM	–	20.472(0.552)	0.959(0.004)	0.068(0.012)
S3 with the imbalanced design				
Proposed	0.000(0.000)	22.736(2.614)	0.954(0.009)	0.038(0.002)
SCAN	0.166(0.004)	34.468(4.191)	0.753(0.063)	0.063(0.009)
Tiger	–	37.710(0.241)	0.744(0.010)	0.180(0.004)
True+Tiger	–	28.955(0.590)	0.837(0.006)	0.065(0.002)
True+JGL	–	28.095(0.539)	0.908(0.010)	0.173(0.004)
True+JSEM	–	20.138(0.685)	0.977(0.007)	0.068(0.003)

heterogeneity has been posing an increasing public health concern. The mRNA gene expression measurements are considered, which are downloaded from the TCGA website using the R package `cgdsr`. In total, 1100 breast cancer subjects are available with 18,506 gene expression measurements. As the number of connected genes is not expected to be large, to improve stability as well as reduce computational cost, we conduct prescreening, which has been a common technique in published studies. Specifically, following [25], we focus on the genes in the Wnt signaling, oxidative phosphorylation, Mtor signaling, and citrate cycle tca cycle pathways, which have been demonstrated to play an important role in all cancer types in the literature. This results in 316 genes for downstream analysis.

In practice, the number of subgroups K is usually unknown. In this study, we adopt the gap statistic using the R package *NbClust* to select the optimal value of K. Given the candidate set $K \in \{2,3,4,5,6,7,8\}$, with the gap statistic, the proposed approach identifies two subgroups with group sizes 805 (subgroup 1) and 295 (subgroup 2). In addition, 1171 and 666 edges among 309 and 289 genes are discovered for the two subgroups, respectively, and 222 common edges across the two subgroups are identified. The graphical representation of the identified networks is provided in Fig. 2.

To gain more insight into the identified networks, we examine the related genes' functional and biological connections by conducting Gene Ontology (GO) enrichment analysis, which is implemented using DAVID 2021 [31]. Our analysis first suggests that those genes involved in the common edges are functionally and biologically connected with certain significantly enriched GO terms.

Table 2
Simulation results under the scenarios with the nearest-neighbor networks and $K = 3$. In each cell, we show the median (median absolute deviation) of the CE, PME, TPR and FPR values based on 100 replicates. CE $= (C_n^2)^{-1}\left|\left\{(i,j): I\left\{\hat{\varphi}\left(\boldsymbol{x}_i\right)=\hat{\varphi}(\boldsymbol{x}_j)\right\}\neq I\left\{\varphi\left(\boldsymbol{x}_i\right)=\varphi(\boldsymbol{x}_j)\right\}, i<j\right\}\right|$, PME $= \sum_{k=1}^{K}\left\|\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}_k^0\right\|_F/K$, TPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0\neq 0, \hat{\omega}_{k,\ell j}\neq 0\right\}\Big/\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0\neq 0\right\}\right]/K$, and FPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0= 0, \hat{\omega}_{k,\ell j}\neq 0\right\}\Big/\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0= 0\right\}\right]/K$.

Approach	CE	PME	TPR	FPR
S1 with the balanced design				
Proposed	0.000(0.000)	18.171(0.524)	0.982(0.005)	0.037(0.002)
SCAN	0.000(0.000)	34.619(0.392)	0.710(0.008)	0.076(0.001)
Tiger	–	37.508(0.207)	0.817(0.016)	0.184(0.004)
True+Tiger	–	22.575(0.616)	0.862(0.007)	0.068(0.001)
True+JGL	–	22.317(0.562)	0.904(0.008)	0.178(0.004)
True+JSEM	–	18.468(0.498)	0.964(0.005)	0.084(0.003)
S2 with the balanced design				
Proposed	0.000(0.000)	17.441(0.515)	0.987(0.003)	0.037(0.001)
SCAN	0.002(0.002)	36.546(0.699)	0.713(0.008)	0.073(0.002)
Tiger	–	40.345(0.245)	0.825(0.013)	0.183(0.004)
True+Tiger	–	23.529(0.654)	0.867(0.006)	0.066(0.001)
True+JGL	–	21.983(0.497)	0.925(0.016)	0.190(0.020)
True+JSEM	–	17.508(0.603)	0.980(0.005)	0.071(0.003)
S3 with the balanced design				
Proposed	0.000(0.000)	17.637(0.507)	0.990(0.003)	0.037(0.002)
SCAN	0.002(0.002)	34.577(0.543)	0.739(0.005)	0.078(0.001)
Tiger	–	38.052(0.304)	0.808(0.015)	0.192(0.006)
True+Tiger	–	24.609(0.612)	0.852(0.008)	0.072(0.001)
True+JGL	–	22.986(0.589)	0.919(0.014)	0.215(0.008)
True+JSEM	–	19.003(0.477)	0.957(0.005)	0.073(0.003)
S1 with the imbalanced design				
Proposed	0.000(0.000)	18.582(0.864)	0.977(0.003)	0.037(0.002)
SCAN	0.000(0.000)	33.991(0.492)	0.710(0.008)	0.075(0.001)
Tiger	–	37.336(0.196)	0.824(0.010)	0.181(0.004)
True+Tiger	–	23.089(0.447)	0.862(0.005)	0.071(0.001)
True+JGL	–	22.730(0.562)	0.907(0.006)	0.179(0.002)
True+JSEM	–	18.693(0.413)	0.962(0.005)	0.083(0.004)
S2 with the imbalanced design				
Proposed	0.000(0.000)	17.526(0.909)	0.980(0.008)	0.038(0.002)
SCAN	0.002(0.002)	35.934(0.459)	0.718(0.005)	0.072(0.002)
Tiger	–	40.089(0.176)	0.817(0.008)	0.178(0.003)
True+Tiger	–	24.006(0.455)	0.863(0.007)	0.071(0.002)
True+JGL	–	22.817(0.562)	0.919(0.013)	0.176(0.005)
True+JSEM	–	17.546(0.663)	0.980(0.003)	0.071(0.004)
S3 with the imbalanced design				
Proposed	0.000(0.000)	17.688(0.609)	0.987(0.003)	0.038(0.001)
SCAN	0.002(0.002)	34.453(0.414)	0.739(0.008)	0.076(0.002)
Tiger	–	38.031(0.239)	0.824(0.012)	0.192(0.005)
True+Tiger	–	25.031(0.467)	0.857(0.005)	0.076(0.002)
True+JGL	–	23.854(0.716)	0.921(0.010)	0.219(0.005)
True+JSEM	–	19.367(0.456)	0.957(0.005)	0.074(0.004)

For example, in one common subnetwork, genes ATP6V1A, ATP5F1A, ATP6V1B2, and NDUFS1 are enriched with the ATP metabolic process (GO:0046034, P-value: 7.37×10^{-3}), which has been demonstrated to be closely linked to the selective killing of respiratory competent cancer cells that are critical for tumor progression, including breast cancer [21]. In addition, in this subnetwork, genes ROCK1, ROCK2, and RAC1 are enriched with regulation of stress fiber assembly (GO:0051492, P-value: 1.06×10^{-3}), and studies have shown that stress fiber-mediated cellular stiffness can promote tumor growth in the precancerous stage, including breast cancer [35]. Genes NDUFA11, COX6A1, RPS6KA3, FRAT1, STK11, SOX17, RPS6KA1, EP300, and ATP6V1E1 are enriched with protein binding function (GO:0005515, P-value: 3.14×10^{-12}), which has been reported to play an important role in cancer treatment, and in fact several proteins have been shown to be effective in delivering to tumor sites [37]. Genes GSK3B, PRKAA1, CHD8, TCF7, LRP6, NKD2, SOX17, DVL1, and DVL3 are enriched with the Wnt signaling pathway (GO:0016055, P-value: 1.63×10^{-21}), which has been reported to be activated in over half of breast cancer patients and plays an important role in triple negative breast cancer development [39]. Moreover, genes COX7B, NDUFA11, COX4I1, NDUFA10, COX6A1, COX7C, COX8 A, NDUFC2, NDUFC1, and SDHD are enriched with mitochondrial inner membrane (GO:0005743, P-value: 1.39×10^{-58}), which has been reported to be essential in the regulation of cancer cell migration and invasion [36], including breast cancer [41]. The GO enrichment analysis also finds that the genes involved in the specific edges of different subgroups are associated with some distinct significantly enriched GO terms. For example, genes WNT5 A, TSC2, ULK1, AXIN2, CSNK1E, and TP53 are enriched with the protein localization process (GO:0008104, P-value: 3.58×10^{-3}) in subgroup 1, which has been reported to be implicated in the pathogenesis of human diseases, including breast cancer, and therapeutic strategies targeting protein localization have been conceptualized as promising for the

Table 3
Simulation results under the scenarios with the Erdös-Rényi networks and $K = 3$. In each cell, we show the median (median absolute deviation) of the CE, PME, TPR and FPR values based on 100 replicates. CE $= (C_n^2)^{-1}\left|\{(i,j): I\{\hat{\varphi}(\boldsymbol{x}_i)=\hat{\varphi}(\boldsymbol{x}_j)\}\neq I\{\varphi(\boldsymbol{x}_i)=\varphi(\boldsymbol{x}_j)\}, i<j\}\right|$, PME $= \sum_{k=1}^{K}\left\|\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}_k^0\right\|_F/K$, TPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0\neq 0, \hat{\omega}_{k,\ell j}\neq 0\right\}\Big/\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0\neq 0\right\}\right]/K$, and FPR $= \sum_{k=1}^{K}\left[\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0= 0, \hat{\omega}_{k,\ell j}\neq 0\right\}\Big/\sum_{\ell<j} I\left\{\omega_{k,\ell j}^0= 0\right\}\right]/K$.

Approach	CE	PME	TPR	FPR
S1 with the balanced design				
Proposed	0.002(0.002)	16.805(1.169)	0.988(0.004)	0.040(0.001)
SCAN	0.004(0.002)	26.831(1.337)	0.779(0.016)	0.076(0.004)
Tiger	–	31.927(0.190)	0.797(0.015)	0.176(0.005)
True+Tiger	–	17.662(0.593)	0.918(0.008)	0.073(0.001)
True+JGL	–	18.989(0.619)	0.918(0.008)	0.181(0.003)
True+JSEM	–	17.261(0.455)	0.945(0.008)	0.098(0.003)
S2 with the balanced design				
Proposed	0.002(0.002)	16.751(0.906)	0.987(0.004)	0.039(0.002)
SCAN	0.004(0.002)	26.975(0.412)	0.772(0.007)	0.076(0.001)
Tiger	–	31.025(0.220)	0.789(0.017)	0.169(0.005)
True+Tiger	–	21.104(0.451)	0.870(0.005)	0.075(0.001)
True+JGL	–	20.795(0.460)	0.901(0.020)	0.184(0.021)
True+JSEM	–	16.403(0.586)	0.965(0.009)	0.087(0.002)
S3 with the balanced design				
Proposed	0.002(0.002)	14.699(0.628)	0.996(0.004)	0.039(0.002)
SCAN	0.006(0.006)	17.155(2.467)	0.768(0.027)	0.061(0.008)
Tiger	–	20.312(0.181)	0.796(0.012)	0.186(0.006)
True+Tiger	–	17.670(0.467)	0.919(0.008)	0.074(0.001)
True+JGL	–	17.271(0.459)	0.930(0.010)	0.158(0.005)
True+JSEM	–	15.435(0.433)	0.971(0.004)	0.098(0.003)
S1 with the imbalanced design				
Proposed	0.004(0.002)	16.979(1.322)	0.986(0.006)	0.038(0.003)
SCAN	0.004(0.002)	26.194(0.787)	0.786(0.008)	0.076(0.002)
Tiger	–	31.834(0.169)	0.797(0.012)	0.175(0.004)
True+Tiger	–	18.193(0.564)	0.918(0.005)	0.077(0.002)
True+JGL	–	19.442(0.488)	0.914(0.008)	0.184(0.003)
True+JSEM	–	17.479(0.499)	0.942(0.008)	0.097(0.002)
S2 with the imbalanced design				
Proposed	0.002(0.002)	18.227(1.626)	0.983(0.007)	0.039(0.003)
SCAN	0.004(0.002)	27.080(0.547)	0.764(0.010)	0.076(0.002)
Tiger	–	31.129(0.221)	0.797(0.009)	0.168(0.004)
True+Tiger	–	21.627(0.477)	0.870(0.005)	0.077(0.002)
True+JGL	–	21.139(0.567)	0.893(0.013)	0.182(0.008)
True+JSEM	–	17.005(0.579)	0.962(0.007)	0.088(0.003)
S3 with the imbalanced design				
Proposed	0.002(0.002)	14.946(0.959)	0.992(0.004)	0.039(0.003)
SCAN	0.008(0.004)	25.161(0.934)	0.791(0.016)	0.079(0.003)
Tiger	–	32.592(0.195)	0.794(0.018)	0.159(0.006)
True+Tiger	–	18.603(0.628)	0.919(0.008)	0.078(0.001)
True+JGL	–	18.169(0.440)	0.931(0.009)	0.161(0.004)
True+JSEM	–	15.791(0.668)	0.972(0.004)	0.099(0.002)

treatment of a variety of human diseases [19]. In addition, genes LEF1, PSEN1, DKK1, and LRP6 are enriched with embryonic limb morphogenesis process (GO:0030326, P-value: 0.02) in subgroup 2, where some embryonic genes are significantly upregulated in estrogen receptor negative breast cancer [43]. These biologically sensible findings provide support for the validity of the proposed network identification analysis.

Analyses are also conducted using the alternative approaches with $K = 2$. Here, as the true subgroup memberships are not available, for comparative analyses under known subgroup memberships as well as indirect support for the subgroup identification results, we follow [25] and consider three clinical breast cancer subgroups, namely ER+, ER-, and other unevaluated cases, based on whether the subjects have estrogen receptors. In our data analysis, the sizes of these three subgroups are 812, 238, and 50, respectively. The multiple network estimation procedure using the alternatives is performed based on the ER+ and ER- subgroups. The summary comparison results of the heterogeneity analysis and network analysis are reported in Tables S10 and S11 of the supplementary materials, where the numbers of subjects in each subgroup and edges in each network identified by the different approaches and their overlaps are provided. Here, the subgroups identified using the different approaches are matched by correlation, and the different approaches lead to different subgroup memberships. For ER+ and ER-, we present the proportions of subjects that are identified in different subgroups using the proposed and SCAN approaches in Fig. S4 of the supplementary materials. It is observed that the two approaches can discriminate the ER+ and ER- subgroups effectively, with the proposed approach demonstrating more advantageous results. The P-value of the Chi-square test for the proposed approach is 6.08×10^{-69}. Additionally, by matching the identified subgroups with ER status, the CE of the proposed approach is 0.29, compared to 0.37 for SCAN. These results suggest the effectiveness of the proposed analysis.

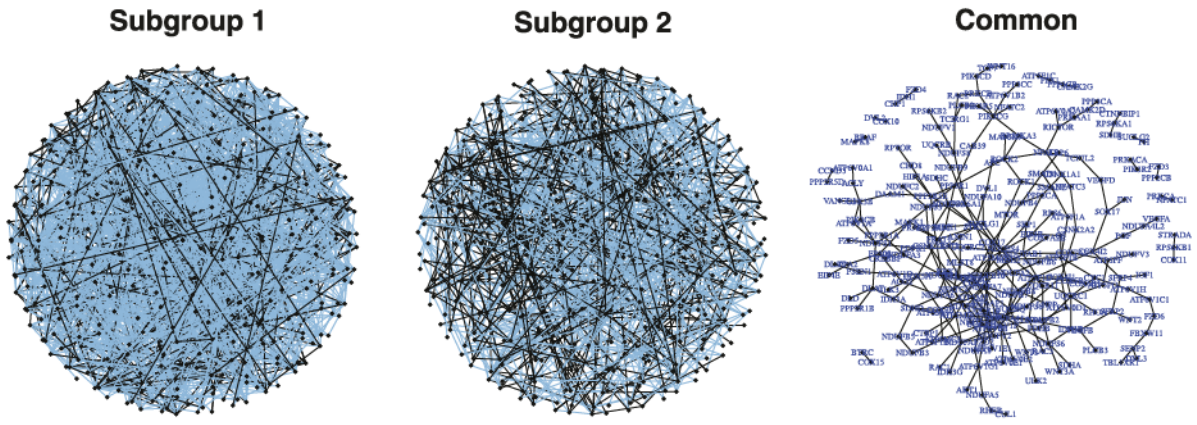


Fig. 2. Data analysis: gene networks for the two subgroups identified with the proposed approach. The black lines represent the common edges shared by the two subgroups, and the blue lines represent the specific edges for each subgroup. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Table S11 of the supplementary materials, we observe that the different approaches identify a moderate number of overlapping edges. To indirectly support the network identification results, following [9], we adopt a resampling strategy and compute the negative log-likelihood statistic (NLS) to evaluate "prediction accuracy", with a smaller value indicating a better performance. Specifically, we first randomly divide the data into a training and a testing set, estimate the parameters on the training set, and finally compute the NLS using the testing set. Based on 100 resamplings, the proposed approach has an average NLS value of 77449.60, compared to 78278.28 for SCAN, 99115.53 for Tiger, 97528.56 for ER+Tiger, 97385.31 for ER+JGL, and 96463.68 for ER+JSEM, with the proposed approach having competitive prediction accuracy.

6. Discussion

Network analysis of heterogeneous subgroups is still a wide-open problem in various research fields. In this article, we have proposed a new joint estimation approach for multiple networks based on the multivariate Gaussian mixture modeling. Different from the existing regularized likelihood-based approaches, which are usually only applicable to small-scale data, the proposed approach provides a more effective and useful tool for estimating multiple precision matrices by solving a collection of simpler sparse regression subproblems. Specifically, based on the Gaussian graphical model, we have proposed a reparameterized mixture of regression models with composite MCP to effectively accommodate both similarities and differences across undiscovered distinct subgroups. Such a strategy enjoys the scale-invariant and tuning-insensitive properties and can be solved in parallel to largely reduce computational cost. The theoretical properties of the proposed approach are investigated, indicating that the proposed estimator enjoys the oracle property. A number of numerical experiments demonstrate the superior performance of the proposed approach in terms of subgroup and network identification accuracy. The application to a TCGA data exploits different gene networks for breast cancer and rediscovers biologically sensible gene relationships associated with the heterogeneous subgroups.

In the proposed approach, we have adopted the composite MCP to accommodate the common and specific structures among different networks. Other penalties, such as $\rho\left(\left\|\text{vec}\left(\beta_{1,j,\ell},\ldots,\beta_{K,j,\ell}\right)\right\|_2;\lambda_1,\gamma_1\right)+\sum_{k=1}^{K}\rho\left(\left|\beta_{k,j,\ell}\right|;\lambda_2,\gamma_2\right)$ (sparse group MCP) can also achieve identification at both the group and individual levels and may have satisfactory statistical and numerical properties. Both composite MCP and sparse group MCP have been adopted in published studies, such as [3,17] for composite MCP, and [22] for sparse group MCP. It is expected that performance of these two penalties may depend on the underlying model, data settings, and other factors. We adopt the composite MCP, as it has been popular in the literature and leads to satisfactory numerical performance. In future works, some other penalties, including this sparse group MCP, can be further investigated. In our data analysis, the TCGA breast cancer data has been analyzed, which has often been used for Gaussian distribution-based network analysis [27,40]. We have followed these studies and conducted analysis without data transformation or normalization, and some results with important biological implications have been found. We acknowledge the complexity of gene expression data with potential non-Gaussian properties. It would be of interest to implement robust techniques like the t distribution mixture model or some other nonparanormal mixture graphical models to further study non-Gaussian distributed data. We have mainly focused on the identification of gene networks using expression data. Many other biological measurements, such as mutation and DNA methylation, can be further used for a better understanding of the mechanisms of cancer.

7. Technical details

Proof of Theorem 1. Recall that $\widetilde{L}_n(\boldsymbol{\theta}_{j,C_j})=\sum_{i=1}^{n}\ln f\left(x_{i,j};\boldsymbol{x}_{i,\setminus j},\boldsymbol{\theta}_{j,C_j}\right)$. Let $\delta_n=\sqrt{Ks_0/n}$ and $\boldsymbol{h}$ be a s_j-dimensional vector, where $s_j=3K+\sum_{k=1}^{K}|S_{j,k}|$ is the nonsparsity size for the jth subproblem. It suffices to show that $\widetilde{L}_n(\boldsymbol{\theta}^0_{j,C_j}+\delta_n\boldsymbol{h})<\widetilde{L}_n(\boldsymbol{\theta}^0_{j,C_j})$ everywhere

on the boundary $\{\boldsymbol{h} : \|\boldsymbol{h}\| = C\}$, where C is a sufficiently large positive constant. By Taylor expansion, we have:

$$\begin{aligned}
&\widetilde{L}_n(\boldsymbol{\theta}^0_{j,\mathcal{C}_j} + \delta_n \boldsymbol{h}) - \widetilde{L}_n(\boldsymbol{\theta}^0_{j,\mathcal{C}_j}) \\
&= \delta_n \boldsymbol{h}^\top \left\{ \frac{\partial \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} \right\} + \frac{1}{2}\delta_n^2 \boldsymbol{h}^\top \left\{ \frac{\partial^2 \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial^2 \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} \right\} \boldsymbol{h} + \frac{\delta_n^3}{6} \sum_{q,\ell,m \in \mathcal{C}_j} \frac{\partial^3 \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial \theta_{j,q} \partial \theta_{j,\ell} \partial \theta_{j,m}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \mathring{\boldsymbol{\theta}}_{j,\mathcal{C}_j}} h_q h_\ell h_m \\
&:= I + II + III,
\end{aligned}$$

where $\mathring{\boldsymbol{\theta}}_{j,\mathcal{C}_j}$ lies on the line segment connecting $\boldsymbol{\theta}^0_{j,\mathcal{C}_j} + \delta_n \boldsymbol{h}$ and $\boldsymbol{\theta}^0_{j,\mathcal{C}_j}$.

For I, by Assumption 1, as $n \to \infty$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \frac{\partial \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,\mathcal{C}_j}\right)}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} \right\} \\
&= \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \ln f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,\mathcal{C}_j}\right)}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} - E\left\{ \frac{\partial \ln f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,\mathcal{C}_j}\right)}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} \right\} \right] \to \mathcal{N}\left(\boldsymbol{0}, \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j})\right).
\end{aligned}$$

Thus,

$$\frac{\partial \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} = O_p(\sqrt{n}).$$

Then, it is easy to see that

$$|I| \le O_p(n\delta_n^2)\|\boldsymbol{h}\|.$$

For II,

$$II = -\frac{1}{2} n\delta_n^2 \boldsymbol{h}^\top \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j}) \boldsymbol{h} + \frac{1}{2} n\delta_n^2 \boldsymbol{h}^\top \left\{ \frac{1}{n} \frac{\partial^2 \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial^2 \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} + \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j}) \right\} \boldsymbol{h}.$$

Following similar arguments as in the proof of Lemma 8 in [11], with Assumptions 1–2, we have:

$$\left\| \frac{1}{n} \frac{\partial^2 \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial^2 \boldsymbol{\theta}_{j,\mathcal{C}_j}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}} + \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j}) \right\| = o_p(1/s_j).$$

Therefore,

$$II = -\frac{1}{2} n\delta_n^2 \boldsymbol{h}^\top \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j}) \boldsymbol{h} + \frac{1}{2} n\delta_n^2 \|\boldsymbol{h}\|^2 \times o_p(1).$$

For III, by Assumption 3 and applying Cauchy–Schwartz inequality, we have:

$$\begin{aligned}
|III| &= \frac{\delta_n^3}{6} \left| \sum_{q,\ell,m \in \mathcal{C}_j} \frac{\partial^3 \widetilde{L}_n(\boldsymbol{\theta}_{j,\mathcal{C}_j})}{\partial \theta_{j,q} \partial \theta_{j,\ell} \partial \theta_{j,m}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \mathring{\boldsymbol{\theta}}_{j,\mathcal{C}_j}} h_q h_\ell h_m \right| = \frac{\delta_n^3}{6} \left| \sum_{q,\ell,m \in \mathcal{C}_j} \sum_{i=1}^n \frac{\partial^3 \ln f\left(x_{i,j}; \boldsymbol{x}_{i,\setminus j}, \boldsymbol{\theta}_{j,\mathcal{C}_j}\right)}{\partial \theta_{j,q} \partial \theta_{j,\ell} \partial \theta_{j,m}} \bigg|_{\boldsymbol{\theta}_{j,\mathcal{C}_j} = \mathring{\boldsymbol{\theta}}_{j,\mathcal{C}_j}} h_q h_\ell h_m \right| \\
&\le \frac{\delta_n^3}{6} \sum_{i=1}^n \left\{ \sum_{q,\ell,m \in \mathcal{C}_j} M_2^2(\boldsymbol{v}_i) \right\}^{1/2} \|\boldsymbol{h}\|^3 = O_p(s_j^{3/2}\delta_n) \times n\delta_n^2 \times \|\boldsymbol{h}\|^2.
\end{aligned}$$

By Assumption 4, for $b \in (0, 1/4)$, $K = O(n^b)$ and $s_0 = o(n^{1/4-b})$, then $(Ks_0)^2 = o(n^{1/2})$. Thus,

$$III = o_p(n\delta_n^2)\|\boldsymbol{h}\|^2.$$

Due to the positive definiteness of the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta}_{j,\mathcal{C}_j})$ at $\boldsymbol{\theta}_{j,\mathcal{C}_j} = \boldsymbol{\theta}^0_{j,\mathcal{C}_j}$, for a sufficiently large C, the quadratic function $-n\delta_n^2 \boldsymbol{h}^\top \mathcal{I}(\boldsymbol{\theta}^0_{j,\mathcal{C}_j})\boldsymbol{h}/2$ in II dominates I and III. This completes the proof.

Proof of Theorem 2. To prove that $\widetilde{\boldsymbol{\theta}}_j$ is a local maximizer of $L_n\left(\boldsymbol{\theta}_j\right)$, we consider $\boldsymbol{\theta}_j$ in a small neighborhood of $\widetilde{\boldsymbol{\theta}}_j$ such that $\left\|\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j\right\| = O\left(n^{a/2-1/2}\sqrt{\ln n/(Kp)} + s_0\sqrt{K/(pn)}\right)$, and let $\underline{\boldsymbol{\theta}}_j$ have $\underline{\boldsymbol{\theta}}_{j,\mathcal{C}_j} = \boldsymbol{\theta}_{j,\mathcal{C}_j}$ and $\underline{\boldsymbol{\theta}}_{j,\mathcal{C}_j^c} = \boldsymbol{0}$. According to Theorem 1, we have

$L_n(\underline{\theta}_j) < L_n(\widetilde{\theta}_j)$. Hence it suffices to show $L_n(\underline{\theta}_j) > L_n(\theta_j)$. Notice that

$$L_n(\theta_j) - L_n(\underline{\theta}_j) = \widetilde{L}_n(\theta_j) - \widetilde{L}_n(\underline{\theta}_j) - n\sum_{\ell\neq j}\left[\rho\left\{\sum_{k=1}^{K}\rho\left(\left|\beta_{k,j,\ell}\right|;\lambda,\gamma\right);\lambda,\frac{K\lambda\gamma}{2}\right\} - \rho\left\{\sum_{k=1}^{K}\rho\left(\left|\underline{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right);\lambda,\frac{K\lambda\gamma}{2}\right\}\right]$$
$$:= L_{n1} + L_{n2}.$$

For L_{n1}, we have:

$$L_{n1} = \widetilde{L}_n(\theta_j) - \widetilde{L}_n(\underline{\theta}_j) = \sum_k\sum_{\ell\in S^c_{j,k}}\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\bar{\theta}_j}\beta_{k,j,\ell} = \sum_k\sum_{\ell\in S^c_{j,k}}\left\{\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\widetilde{\theta}_j} + (\bar{\theta}_j-\widetilde{\theta}_j)^{\mathsf{T}}\left.\frac{\partial^2\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_j}\right|_{\theta_j=\dot{\theta}_j}\right\}\beta_{k,j,\ell},$$

where $\bar{\theta}_j$ lies on the line segment connecting θ_j and $\underline{\theta}_j$, and $\dot{\theta}_j$ lies on the line segment connecting $\bar{\theta}_j$ and $\widetilde{\theta}_j$.

First consider $\left\{\partial\widetilde{L}_n(\theta_j)/\partial\beta_{k,j,\ell}\right\}\Big|_{\theta_j=\bar{\theta}_j}$. With a second order Taylor expansion, for $\ell\in S^c_{j,k}$, we have:

$$\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\bar{\theta}_j} = \left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\theta^0_j} + \left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right)^{\mathsf{T}}\left.\frac{\partial^2\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_{j,C_j}}\right|_{\theta_j=\theta^0_j} + \left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right)^{\mathsf{T}}\left.\frac{\partial^3\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial^2\theta_{j,C_j}}\right|_{\theta_j=\ddot{\theta}_j}\left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right) \tag{10}$$
$$:= I + II + III,$$

where $\ddot{\theta}_j$ lies on the line segment connecting $\widetilde{\theta}_j$ and θ^0_j.

For I in (10), define the following event:

$$\Omega_1 = \left\{\max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\left|\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\theta^0_j}\right| \le \zeta_n\sqrt{n}\right\},$$

where $\zeta_n = n^{a/2}\sqrt{\ln n}$ with $a\in(0,1/2)$. By Assumption 3 and $\ln(p) = O(n^a)$ in Assumption 7, together with Bernstein's inequality, we can obtain that, when $n\to\infty$, there exists $\kappa>0$ such that:

$$Pr(\Omega_1) = 1 - Pr\left\{\max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\left|\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\theta^0_j}\right| > \zeta_n\sqrt{n}\right\} \ge 1 - \sum_{k=1}^{K}\sum_{\ell\in S^c_{j,k}}Pr\left\{\left|\left.\frac{1}{\sqrt{n}}\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\theta^0_j}\right| > \zeta_n\right\}$$
$$\ge 1 - 2(Kp - \sum_{k=1}^{K}|S_{j,k}|)\exp\left(-\frac{\zeta_n^2}{2\kappa}\right) \ge 1 - 2Kp\exp\left(-\frac{\zeta_n^2}{2\kappa}\right)\to 1.$$

Thus, with probability tending to 1,

$$\max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\left|\left.\frac{\partial\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}}\right|_{\theta_j=\theta^0_j}\right| = O(n^{a/2+1/2}\sqrt{\ln n}).$$

For II in (10), applying Assumption 3 and Cauchy–Schwartz inequality yields that:

$$\max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\left|\left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right)^{\mathsf{T}}\left.\frac{\partial^2\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_{j,C_j}}\right|_{\theta_j=\theta^0_j}\right| \le \max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\sum_{i=1}^{n}\left|\sum_{m\in C_j}\left.\frac{\partial^2\ln f(x_{i,j};\boldsymbol{x}_{i,\setminus j},\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_{j,m}}\right|_{\theta_j=\theta^0_j}\left(\widetilde{\theta}_{j,m}-\theta^0_{j,m}\right)\right|$$
$$\le \max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\sum_{i=1}^{n}\left[\sum_{m\in C_j}\left\{\left.\frac{\partial^2\ln f(x_{i,j};\boldsymbol{x}_{i,\setminus j},\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_{j,m}}\right|_{\theta_j=\theta^0_j}\right\}^2\right]^{\frac{1}{2}}\left\|\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right\| \tag{11}$$
$$\le \sum_{i=1}^{n}\left[s_j\left\{M_1(\boldsymbol{v}_i)\right\}^2\right]^{\frac{1}{2}}\left\|\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right\| = O_p(K\sqrt{n}s_0).$$

Similarly, for III in (10), we have:

$$\max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\left|\left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right)^{\mathsf{T}}\left.\frac{\partial^3\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial^2\theta_{j,C_j}}\right|_{\theta_j=\ddot{\theta}_j}\left(\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right)\right|$$
$$\le \max_{k\in\{1,\dots,K\}}\max_{\ell\in S^c_{j,k}}\sum_{i=1}^{n}\left[\sum_{q,m\in C_j}\left\{\left.\frac{\partial^3\ln f(x_{i,j};\boldsymbol{x}_{i,\setminus j},\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_{j,q}\partial\theta_{j,m}}\right|_{\theta_j=\ddot{\theta}_j}\right\}^2\right]^{\frac{1}{2}}\left\|\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right\|^2 \le \sum_{i=1}^{n}\left[s_j^2\left\{M_2(\boldsymbol{v}_i)\right\}^2\right]^{\frac{1}{2}}\left\|\widetilde{\theta}_{j,C_j}-\theta^0_{j,C_j}\right\|^2$$
$$= o_p(K\sqrt{n}s_0).$$

Then, we have:

$$\max_{k\in\{1,\ldots,K\}}\max_{\ell\in S^c_{j,k}}\left|\frac{\partial L_n(\theta_j)}{\partial\beta_{k,j,\ell}}\bigg|_{\theta_j=\widetilde{\theta}_j}\right|\le O_p(n^{a/2+1/2}\sqrt{\ln n})+O_p(K\sqrt{n}s_0)+o_p(K\sqrt{n}s_0).$$

Moreover, using similar discussion as that in (11), we can obtain that:

$$\max_{k\in\{1,\ldots,K\}}\max_{\ell\in S^c_{j,k}}(\bar{\theta}_j-\widetilde{\theta}_j)^{\top}\frac{\partial^2\widetilde{L}_n(\theta_j)}{\partial\beta_{k,j,\ell}\partial\theta_j}\bigg|_{\theta_j=\bar{\theta}_j}\le\sum_{i=1}^{n}\left[K(p+2)\left\{M_1(v_i)\right\}^2\right]^{\frac{1}{2}}\left\|\bar{\theta}_j-\widetilde{\theta}_j\right\|=O\left(n\sqrt{Kp}\right)\left\|\bar{\theta}_j-\widetilde{\theta}_j\right\|.$$

Together with $\left\|\bar{\theta}_j-\widetilde{\theta}_j\right\|\le\left\|\theta_j-\widetilde{\theta}_j\right\|=O\left(n^{a/2-1/2}\sqrt{\ln n/(Kp)}+s_0\sqrt{K/(pn)}\right)$, we obtain that:

$$L_{n1}=\widetilde{L}_n(\theta_j)-\widetilde{L}_n(\underline{\theta}_j)=\sum_k\sum_{\ell\in S^c_{j,k}}\left\{O_p\left(n^{a/2+1/2}\sqrt{\ln n}\right)+O_p\left(K\sqrt{n}s_0\right)\right\}\beta_{k,j,\ell}. \tag{12}$$

For L_{n2}, let $\dot{\rho}(\beta;\lambda,\gamma)$ be the derivative of $\rho(\beta;\lambda,\gamma)$ with respect to β, more specifically,

$$\dot{\rho}(\beta;\lambda,\gamma)=\left(\lambda-\frac{\beta}{\gamma}\right)I\{\beta\le\lambda\gamma\}.$$

By Taylor expansion, we have:

$$L_{n2}=-n\sum_{k=1}^{K}\sum_{\ell\in S^c_{j,k}}\dot{\rho}\left\{\sum_{k'=1}^{K}\rho\left(\left|\breve{\beta}_{k',j,\ell}\right|;\lambda,\gamma\right);\lambda,\frac{K\lambda\gamma}{2}\right\}\dot{\rho}\left(\left|\breve{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right)\left|\beta_{k,j,\ell}\right|, \tag{13}$$

where $\breve{\beta}$ lies between $\underline{\beta}$ and β. Result (13) also depends on the following fact: by Assumption 6, $\breve{\beta}_{k,j,\ell}>\gamma\lambda$, and thus $\dot{\rho}\left(\left|\breve{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right)=0$ for $\ell\in S_{j,k}$.

Note that when $\ell\in S^c_{j,k}$, when $n\to\infty$,

$$\dot{\rho}\left(\left|\breve{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right)\to\lambda,$$

and

$$\sum_{k'=1}^{K}\rho\left(\left|\breve{\beta}_{k',j,\ell}\right|;\lambda,\gamma\right)=\sum_{k'\ne k}^{K}\rho\left(\left|\breve{\beta}_{k',j,\ell}\right|;\lambda,\gamma\right)+\rho\left(\left|\breve{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right)\to m_{j,\ell}\frac{\gamma\lambda^2}{2},$$

where $m_{j,\ell}=\#\left\{k':\beta^0_{k',j,\ell}\ne0\right\}<K$ is the number of the nonzero elements in $\{\beta_{1,j,\ell},\ldots,\beta_{K,j,\ell}\}$. Therefore, for $\ell\in S^c_{j,k}$, when $n\to\infty$,

$$\dot{\rho}\left\{\sum_{k'=1}^{K}\rho\left(\left|\breve{\beta}_{k',j,\ell}\right|;\lambda,\gamma\right);\lambda,\frac{K\lambda\gamma}{2}\right\}\dot{\rho}\left(\left|\breve{\beta}_{k,j,\ell}\right|;\lambda,\gamma\right)\to\left(1-\frac{m_{j,\ell}}{K}\right)\lambda^2. \tag{14}$$

Following similar arguments as in the proof of Theorem 1 in [10], and combining with the results of (12) and (14), we can obtain that when $n\to\infty$,

$$L_n(\theta_j)-L_n(\underline{\theta}_j)=L_{n1}+L_{n2}\to\sum_k\sum_{\ell\in S^c_{j,k}}\left\{O_p\left(n^{a/2+1/2}\sqrt{\ln n}\right)+O_p\left(Ks_0\sqrt{n}\right)\right\}\beta_{k,j,\ell}-n\left(1-\frac{m_{j,\ell}}{K}\right)\lambda^2\left|\beta_{k,j,\ell}\right|.$$

By Assumption 5, $n^{a/2-1/2}\sqrt{\ln n}/\lambda^2\to0$ and $Ks_0/(\sqrt{n}\lambda^2)\to0$ when $n\to\infty$, we can prove that $L_n(\theta_j)-L_n(\underline{\theta}_j)<0$, which indicates that $\widetilde{\theta}_j$ is a local maximizer of $L_n(\theta_j)$.

Proof of Theorem 3. From Theorem 2, we know that there exists a strict local maximizer $\hat{\theta}_j$ of $L_n(\theta_j)$, which is asymptotically as efficient as $\widetilde{\theta}_j$. Next we consider the properties of the estimators $\hat{\Omega}_k$'s obtained from (5) and the resulting edge set $\hat{E}_k=\left\{(\ell,j):1\le\ell\ne j\le p,\text{ and }\hat{\omega}_{k,\ell j}\ne0\right\}$. Recall that

$$\beta_{k,j,\ell}=-\omega_{k,j\ell}/\sqrt{\omega_{k,jj}},\tau_{k,j}=\sqrt{\omega_{k,jj}}.$$

Thus,

$$\omega_{k,j\ell}=-\beta_{k,j,\ell}\tau_{k,j},\omega_{k,jj}=\tau^2_{k,j}.$$

Together with the result in Theorem 2, it is obvious that with probability tending to 1, $\hat{E}_k=E^0_k$. Furthermore, since $\left|\hat{\tau}_{k,j}-\tau^0_{k,j}\right|\le O_p(\sqrt{Ks_0/n})$ and $\tau^0_{k,j}$ is bounded by Assumption 6, $\left|\hat{\tau}_{k,j}/\tau^0_{k,j}-1\right|\le O_p(\sqrt{Ks_0/n})$. Thus for the diagonal element $\omega_{k,jj}$, we have:

$$\left|\hat{\omega}_{k,jj}-\omega^0_{k,jj}\right|=\left|\hat{\tau}^2_{k,j}-\left(\tau^0_{k,j}\right)^2\right|=\left|\hat{\tau}_{k,j}-\tau^0_{k,j}\right|\left(\hat{\tau}_{k,j}+\tau^0_{k,j}\right)=O_p\left\{\sqrt{Ks_0/n}\left(\max_k\max_j\tau^0_{k,j}\right)\right\}=O_p\left(\sqrt{Ks_0/n}\right).$$

Together with $\left|\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right| \le O_p(\sqrt{Ks_0/n})$, for the off-diagonal element $\omega_{k,j\ell}$ with $j \neq \ell$, we have:

$$\left|\hat{\omega}_{k,j\ell}-\omega^0_{k,j\ell}\right| = \left|\hat{\beta}_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\tau^0_{k,j}\right| \le \left(\left|\hat{\beta}_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\hat{\tau}_{k,j}\right| + \left|\beta^0_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\tau^0_{k,j}\right|\right)$$
$$\le \hat{\tau}_{k,j}\left|\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right| + \left|\frac{\hat{\tau}_{k,j}}{\tau^0_{k,j}}-1\right|\left|\omega^0_{k,j\ell}\right| = O_p\left(\sqrt{Ks_0/n}\right).$$

Thus, we conclude that $\max_{k,j,\ell}\left|\hat{\omega}_{k,j\ell}-\omega^0_{k,j\ell}\right| = O_p\left(\sqrt{Ks_0/n}\right)$.

Similarly,

$$\sum_{k=1}^K \left\|\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}^0_k\right\|_\infty = \sum_{k=1}^K \max_j\left\{\sum_{\ell=1}^p\left|\hat{\omega}_{k,j\ell}-\omega^0_{k,j\ell}\right|\right\} = \sum_{k=1}^K \max_j\left\{\sum_{\ell\neq j}\left|\hat{\beta}_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\tau^0_{k,j}\right| + \left|\hat{\tau}^2_{k,j}-\left(\tau^0_{k,j}\right)^2\right|\right\}$$
$$\le \sum_{k=1}^K \max_j\left\{\sum_{\ell\neq j}\left(\left|\hat{\beta}_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\hat{\tau}_{k,j}\right| + \left|\beta^0_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\tau^0_{k,j}\right|\right) + \left|\hat{\tau}_{k,j}-\tau^0_{k,j}\right|\left(\hat{\tau}_{k,j}+\tau^0_{k,j}\right)\right\}$$
$$\le \sum_{k=1}^K \max_j\left\{\left|\hat{\tau}_{k,j}-\tau^0_{k,j}\right|\sum_{\ell\neq j}\left|\beta^0_{k,j,\ell}\right| + \hat{\tau}_{k,j}\sum_{\ell\neq j}\left|\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right| + \left|\hat{\tau}_{k,j}-\tau^0_{k,j}\right|\left(\hat{\tau}_{k,j}+\tau^0_{k,j}\right)\right\}$$
$$\le O_p\left(\sqrt{Ks_0/n}\max_j\sum_{k=1}^K\sum_{\ell\neq j}\left|\beta^0_{k,j,\ell}\right| + Ks_0/\sqrt{n} + \sqrt{K^3s_0/n}\right) = O_p\left(\sqrt{K^3s_0^3/n}\right).$$

Besides,

$$\sum_{k=1}^K \left\|\left(\hat{\boldsymbol{\Omega}}_k-\boldsymbol{\Omega}^0_k\right)_{\mathcal{A}}\right\|_F = \sum_{k=1}^K\left\{\sum_{j=1}^p\sum_{\ell\neq j}\left(\hat{\omega}_{k,j\ell}-\omega^0_{k,j\ell}\right)^2\right\}^{1/2} = \sum_{k=1}^K\left\{\sum_{j=1}^p\sum_{\ell\neq j}\left(\hat{\beta}_{k,j,\ell}\hat{\tau}_{k,j}-\beta^0_{k,j,\ell}\tau^0_{k,j}\right)^2\right\}^{1/2}$$
$$\le \sum_{k=1}^K\left\{2\sum_{j=1}^p\left(\hat{\tau}_{k,j}-\tau^0_{k,j}\right)^2\sum_{\ell\neq j}\left(\beta^0_{k,j,\ell}\right)^2 + 2\sum_{j=1}^p\hat{\tau}^2_{k,j}\sum_{\ell\neq j}\left(\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right)^2\right\}^{1/2}$$
$$\le \sum_{k=1}^K\left\{2\sum_{j=1}^p\left(\frac{\hat{\tau}_{k,j}}{\tau^0_{k,j}}-1\right)^2\sum_{\ell\neq j}\left(\omega^0_{k,j\ell}\right)^2 + 2\sum_{j=1}^p\hat{\tau}^2_{k,j}\sum_{\ell\neq j}\left(\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right)^2\right\}^{1/2}$$
$$\le O_p\left(\sqrt{Ks_0/n}\right)\sum_{k=1}^K\left\{\sum_{j=1}^p\sum_{\ell\neq j}\left(\omega^0_{k,j\ell}\right)^2\right\}^{1/2} + O_p\left[\sum_{k=1}^K\left\{\sum_{j=1}^p\sum_{\ell\neq j}\left(\hat{\beta}_{k,j,\ell}-\beta^0_{k,j,\ell}\right)^2\right\}^{1/2}\right] \le O_p\left(\sqrt{K^3s_0s/n}\right) \le O_p\left(\sqrt{K^3s^2/n}\right).$$

CRediT authorship contribution statement

Xing Qin: Methodology, Data curation, Formal analysis Writing – original draft, Review & editing. **Jianhua Hu:** Writing – original draft, Review & editing. **Shuangge Ma:** Methodology, Writing – original draft, Review & editing. **Mengyun Wu:** Conceptualization, Methodology, Writing – original draft, Review & editing.

Acknowledgments

The authors thank the editors and reviewers for their careful review and insightful comments. This work was supported by National Natural Science Foundation of China [12071273], Shanghai Rising-Star Program [22QA1403500], Shanghai Research Center for Data Science and Decision Technology, National Institutes of Health [CA204120], National Science Foundation [2209685], Shanghai Science and Technology Development Funds [23JC1402100], and Science Research Project of Hebei Education Department [ZD2021043].

Appendix A. Supplementary data

The Supplementary Materials contain additional results of the proposed algorithm, simulations, and data analysis, which can be found online.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2024.105298.

References

[1] O. Banerjee, L. El Ghaoui, A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, J. Mach. Learn. Res. 9 (2008) 485–516.

[2] A.E. Bilgrau, C.F. Peeters, P.S. Eriksen, M. Bø gsted, W.N. van Wieringen, Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes, J. Mach. Learn. Res. 21 (26) (2020) 1–52.

[3] P. Breheny, J. Huang, Penalized methods for bi-level variable selection, Stat. Interface 2 (3) (2009) 369–380.

[4] P. Buphamalai, T. Kokotovic, V. Nagy, J. Menche, Network analysis reveals rare disease signatures across multiple levels of biological organization, Nature Commun. 12 (1) (2021) 1–15.
[5] T. Cai, H. Li, W. Liu, J. Xie, Joint estimation of multiple high-dimensional precision matrices, Statist. Sinica 26 (2016) 445–464.
[6] T. Cai, W. Liu, X. Luo, A constrained l_1 minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 594–607.
[7] P. Danaher, P. Wang, D.M. Witten, The joint graphical lasso for inverse covariance estimation across multiple classes, J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (2) (2014) 373–397.
[8] S. Engelke, A.S. Hitz, Graphical models for extremes, J. R. Stat. Soc. Ser. B Stat. Methodol. 82 (4) (2020) 871–932.
[9] X. Fan, K. Fang, S. Ma, S. Wang, Q. Zhang, Assisted graphical model for gene expression data analysis, Stat. Med. 38 (13) (2019) 2364–2380.
[10] J. Fan, J. Lv, Nonconcave penalized likelihood with NP-dimensionality, IEEE Trans. Inform. Theory 57 (8) (2011) 5467–5484.
[11] J. Fan, H. Peng, Nonconcave penalized likelihood with a diverging number of parameters, Ann. Statist. 32 (3) (2004) 928–961.
[12] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.
[13] C. Gao, Y. Zhu, X. Shen, W. Pan, Estimation of multiple networks in Gaussian mixture models, Electron. J. Stat. 10 (2016) 1133–1154.
[14] A.J. Gibberd, J.D. Nelson, Regularized estimation of piecewise constant Gaussian graphical models: The group-fused graphical lasso, J. Comput. Graph. Statist. 26 (3) (2017) 623–634.
[15] B. Hao, W.W. Sun, Y. Liu, G. Cheng, Simultaneous clustering and estimation of heterogeneous graphical models, J. Mach. Learn. Res. 18 (2018) 1–58.
[16] S.M. Hill, L.M. Heiser, T. Cokelaer, M. Unger, N.K. Nesser, D.E. Carlin, Y. Zhang, A. Sokolov, E.O. Paull, C.K. Wong, et al., Inferring causal molecular networks: Empirical assessment through a community-based effort, Nature Methods 13 (4) (2016) 310–318.
[17] Y. Huang, Q. Zhang, S. Zhang, J. Huang, S. Ma, Promoting similarity of sparsity structures in integrative analysis with penalization, J. Amer. Statist. Assoc. 112 (517) (2017) 342–350.
[18] F.K.C. Hui, D.I. Warton, S.D. Foster, Multi-species distribution modeling using penalized mixture of regressions, Ann. Appl. Stat. 9 (2) (2015) 866–882.
[19] M.-C. Hung, W. Link, Protein localization in disease and therapy, J. Cell Sci. 124 (20) (2011) 3381–3392.
[20] A. Khalili, S. Lin, Regularization in finite mixture of regression models with diverging number of parameters, Biometrics 69 (2) (2013) 436–446.
[21] M.S. Kim, R. Gernapudi, Y.C. Cedeño, B.M. Polster, R. Martinez, P. Shapiro, S. Kesari, E. Nurmemmedov, A. Passaniti, Targeting breast cancer metabolism with a novel inhibitor of mitochondrial ATP synthesis, Oncotarget 11 (43) (2020) 3863–3885.
[22] J. Liu, J. Huang, Y. Xie, S. Ma, Sparse group penalized integrative analysis of multiple cancer prognosis datasets, Genetics Res. 95 (2–3) (2013) 68–77.
[23] W. Liu, X. Luo, Fast and adaptive sparse precision matrix estimation in high dimensions, J. Multivariate Anal. 135 (2015) 153–162.
[24] H. Liu, L. Wang, Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models, Electron. J. Stat. 11 (1) (2017) 241–294.
[25] J. Ma, G. Michailidis, Joint structural estimation of multiple graphical models, J. Mach. Learn. Res. 17 (1) (2016) 5777–5824.
[26] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[27] Y. Niu, Y. Ni, D. Pati, B.K. Mallick, Covariate-assisted Bayesian graph learning for heterogeneous data, J. Amer. Statist. Assoc. (2023) 1–15.
[28] B.S. Price, A.J. Molstad, B. Sherwood, Estimating multiple precision matrices with cluster fusion regularization, J. Comput. Graph. Statist. 30 (4) (2021) 823–834.
[29] M. Ren, S. Zhang, Q. Zhang, S. Ma, et al., Gaussian graphical model-based heterogeneity analysis via penalized fusion, Biometrics 78 (2) (2022) 524–535.
[30] J. Shen, F. Liu, Y. Tu, C. Tang, Finding gene network topologies for given biological function with recurrent neural network, Nature Commun. 12 (1) (2021) 1–10.
[31] B.T. Sherman, M. Hao, J. Qiu, X. Jiao, M.W. Baseler, H.C. Lane, T. Imamichi, W. Chang, DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic Acids Res. 10 (2022) W216–W221.
[32] N. Städler, P. Bühlmann, S. Van De Geer, ℓ_1-Penalization for mixture regression models, Test 19 (2010) 209–256.
[33] T. Sun, C.-H. Zhang, Scaled sparse linear regression, Biometrika 99 (4) (2012) 879–898.
[34] T. Sun, C.-H. Zhang, Sparse matrix inversion with scaled lasso, J. Mach. Learn. Res. 14 (1) (2013) 3385–3418.
[35] S. Tavares, A.F. Vieira, A.V. Taubenberger, M. Araújo, N.P. Martins, C. Brás-Pereira, A. Polónia, M. Herbig, C. Barreto, O. Otto, et al., Actin stress fiber organization promotes cell stiffening and proliferation of pre-invasive breast cancer cells, Nature Commun. 8 (1) (2017) 1–18.
[36] D.C. Wallace, Mitochondria and cancer, Nat. Rev. Cancer 12 (10) (2012) 685–698.
[37] X. Wang, S. Li, S. Wang, S. Zheng, Z. Chen, H. Song, Protein binding nanoparticles as an integrated platform for cancer diagnosis and treatment, Adv. Sci. 9 (29) (2022) 2202453.
[38] H. Yi, Q. Zhang, C. Lin, S. Ma, Information-incorporated Gaussian graphical model for gene expression data, Biometrics 78 (2) (2022) 512–523.
[39] T. Zhan, N. Rindtorff, M. Boutros, Wnt signaling in cancer, Oncogene 36 (11) (2017) 1461–1473.
[40] X.-F. Zhang, L. Ou-Yang, T. Yan, X.T. Hu, H. Yan, A joint graphical model for inferring gene networks across multiple subpopulations and data types, IEEE Trans. Cybern. 51 (2) (2019) 1043–1055.
[41] J. Zhao, J. Zhang, M. Yu, Y. Xie, Y. Huang, D.W. Wolff, P.W. Abel, Y. Tu, Mitochondrial dynamics regulates migration and invasion of breast cancer cells, Oncogene 32 (40) (2013) 4814–4824.
[42] T. Zhong, Q. Zhang, J. Huang, M. Wu, S. Ma, Heterogeneity analysis via integrating multi-sources high-dimensional data with applications to cancer studies, Statist. Sinica 33 (2023) 729–758.
[43] M. Zvelebil, E. Oliemuller, Q. Gao, O. Wansbury, A. Mackay, H. Kendrick, M.J. Smalley, J.S. Reis-Filho, B.A. Howard, Embryonic mammary signature subsets are activated in Brca1-/- and basal-like breast cancers, Breast Cancer Res. 15 (2) (2013) 1–17.