Biostatistics, 2025, **26**(1), kxae028
https://doi.org/10.1093/biostatistics/kxae028
Article

Incorporating prior information in gene expression network-based cancer heterogeneity analysis

Rong Li[1], Shaodong Xu[2], Yang Li[2], Zuojian Tang[3], Di Feng[3], James Cai[3], Shuangge Ma[1,*]

[1]Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, 06511, CT, United States
[2]Center for Applied Statistics and School of Statistics, Renmin University of China, 59 Zhongguancun Street, 100872, Beijing, China
[3]Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharmaceuticals Inc., 900 Ridgebury Road, Ridgefield, 06877, CT, United States

*Corresponding author: Email: shuangge.ma@yale.edu

SUMMARY

Cancer is molecularly heterogeneous, with seemingly similar patients having different molecular landscapes and accordingly different clinical behaviors. In recent studies, gene expression networks have been shown as more effective/informative for cancer heterogeneity analysis than some simpler measures. Gene interconnections can be classified as "direct" and "indirect," where the latter can be caused by shared genomic regulators (such as transcription factors, microRNAs, and other regulatory molecules) and other mechanisms. It has been suggested that incorporating the regulators of gene expressions in network analysis and focusing on the direct interconnections can lead to a deeper understanding of the more essential gene interconnections. Such analysis can be seriously challenged by the large number of parameters (jointly caused by network analysis, incorporation of regulators, and heterogeneity) and often weak signals. To effectively tackle this problem, we propose incorporating prior information contained in the published literature. A key challenge is that such prior information can be partial or even wrong. We develop a two-step procedure that can flexibly accommodate different levels of prior information quality. Simulation demonstrates the effectiveness of the proposed approach and its superiority over relevant competitors. In the analysis of a breast cancer dataset, findings different from the alternatives are made, and the identified sample subgroups have important clinical differences.

KEYWORDS: gene expression network; heterogeneity analysis; prior information; regulation.

1. INTRODUCTION

Heterogeneity analysis has been extensively conducted in the research and clinical treatment of cancer (and many other complex diseases). In such analysis, the goal is to separate seemingly similar patients into subgroups that may have different behaviors. Analysis has been conducted based on a wide range of factors such as demographic and clinical characteristics, pathological images, immunological features, and others. With the maturity of high-throughput profiling techniques,

Received: January 26, 2024. **Revised:** July 4, 2024. **Accepted:** July 8, 2024

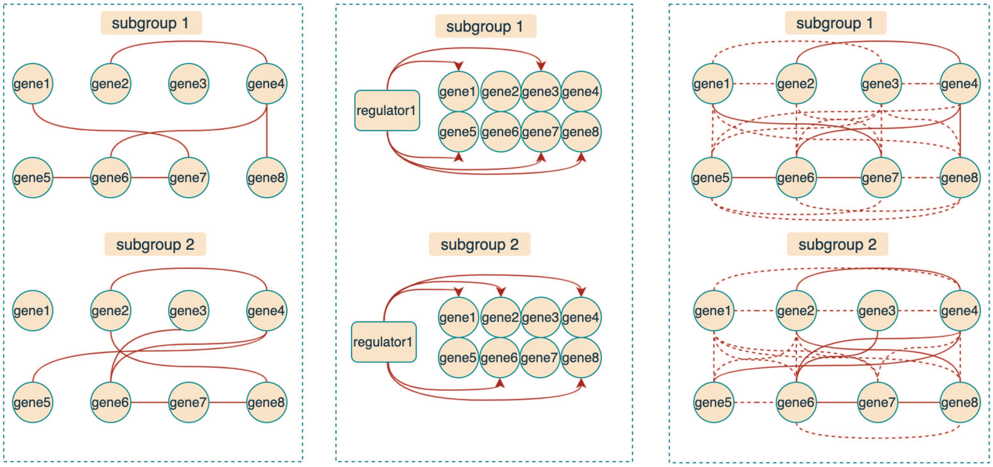

Figure 1. Schematic presentation of gene expression networks under heterogeneity. Left: gene expression networks with direct interconnections only; Middle: regulations of gene expressions by regulators; Right: gene expression networks with both direct and indirect interconnections.

molecular measurements have been effectively integrated into heterogeneity analysis (Navin et al. 2010; Meeks et al. 2020). In particular, a series of studies have shown that gene expression-based heterogeneity analysis can be highly successful (Budinska et al. 2013; Church et al. 2019). Some early studies are based on simple measures, such as the means and variances of gene expressions. In recent studies, it has been shown that gene expression network analysis, which takes a systemic perspective, can generate more insights into patient heterogeneity patterns (Tang et al. 2018; Pio et al. 2022). Here, it is noted that measures such as mean and variance can be easily incorporated into network-based analysis.

As noted in the literature (Kang et al. 2015), the interconnection between two gene expressions can be roughly classified as indirect and direct. An indirect interconnection can occur when for example two gene expressions are mediated by shared genomic regulators, such as transcription factors, microRNAs, and other regulatory molecules. In contrast, a direct interconnection does not involve such shared regulators and may describe a more essential interconnection. A schematic presentation is provided in Fig. 1. The gene expression networks with only direct interconnections (left panel) are usually sparser. The middle panel describes the regulations of gene expressions by regulators, and such regulation relationships are usually sparse. When both direct and indirect interconnections are included, as shown in the right panel, the networks are denser, which may mask the more essential gene interconnections. For the small example in Fig. 1, the networks in the right panel for the two subgroups have 78.3% overlapping edges, while the networks in the left panel have 46.2% overlapping edges—making them easier to be distinguished. Incorporating regulators in gene expression analysis has been made possible by the increasing popularity of multiomics studies, which collect gene expression and regulator data on the same subjects (Kagohara et al. 2018; Lee et al. 2021). It is especially worth noting that there have been a handful of gene expression network analyses incorporating regulators, including heterogeneity analysis (Li et al. 2023).

Gene expression data analysis is "traditionally" challenged by high data dimensionality and weak signals. Network analysis, heterogeneity analysis, and incorporating regulators each and all make these challenges more serious. To improve estimation, here we adopt the strategy of borrowing strengths from prior information, in particular, that contained in published literature. The proposed analysis involves the interconnections among gene expressions and the regulations of gene expressions by regulators. Our preliminary exploration suggests that there are relatively fewer published findings on the regulations of gene expressions. As such, in this study, we focus on the prior information on gene expression interconnections. It is noted that, with relatively straightforward extensions, the proposed analysis can incorporate prior information on the regulations. To fix ideas, we mine prior information by searching PubMed. There are more than 20,000 published studies that simultaneously include "EGFR," "ERBB2," and "breast cancer." In contrast, there are only 18 published studies that simultaneously include "WLPH," "ARAF," and "breast cancer." Based on this observation, it is sensible to conjecture that for breast cancer, EGFR and ERBB2 have a higher likelihood of being interconnected than WLPH and ARAF. Here, it is noted that this search is coarse. There are more refined ways to mine published literature (Lee et al. 2020), however, they may involve complicated algorithms/coding. Additionally, it is noted that PubMed (or any other literature database) does not include all published findings, published findings can be partial or even wrong, and the cooccurrence of two genes in a published literature does not necessarily suggest that they are interconnected. As such, the prior information can be useful but cannot be "fully trusted."

With the consideration of the quality of the prior information, our proposal is to data-dependently "balance" between the prior information and the information contained in the observed data. The proposed strategy shares some similar spirit with that in Jiang et al. (2016), Li et al. (2022), and Wang et al. (2023), which have considerably simpler settings. Broadly speaking, incorporating prior information is not a new strategy. The most natural may be Bayesian analysis (Zhao et al. 2019), which adopts significantly different techniques. Highly curated biological information has also been used (Fan et al. 2016)—it is noted that such information can often be fully trusted and that a higher quality often means a smaller amount of information.

In this study, the goal is to develop a new technique that can further advance cancer heterogeneity analysis. Built on the existing literature, this study advances in multiple important aspects. First, compared to heterogeneity analysis that is based on simpler measures, it is based on gene expression networks as well as regulations of gene expressions by regulators. Second, different from most of the existing multiomics analyses, heterogeneity analysis is conducted, which is critical for cancer and many other diseases. Third, the type of prior information used is significantly different from that in Burrell et al. (2013), Fan et al. (2016), and many others. The data/model settings are much more complicated than in Jiang et al. (2016), Li et al. (2022), and Wang et al. (2023). The adopted analysis strategy significantly differs from Bayesian analysis. The proposed method incorporates prior information to conduct network-based unsupervised clustering, which demands more challenging computation. It does not specify prior information regarding the number of subgroups, which is very challenging to obtain in practice. Instead, it adopts a fusion technique to determine the number of subgroups along with model parameter estimation in a fully data-driven manner. Last but not least, as can be partly seen from our data analysis, this study can also deliver a practically useful tool that can lead to new insights into the heterogeneity of cancer (and some other diseases).

2. METHODS

The proposed analysis is unsupervised and takes measurements on gene expressions and their regulators as input. As argued in the published multiomics studies—in particular including those involving network analysis and heterogeneity analysis (Tarazona et al. 2021; Henao et al. 2023), the collection of regulators does not need to be "complete"—some types of regulators or some components of a specific type of regulators may not be available. The overall overflow is shown in Fig. 2. The first step is to obtain the prior information. After that, the analysis contains two steps. In the first step of prior information-guided analysis, the prior information is "fully trusted." In the second step of prior information-incorporated analysis, we take into account the varying quality of the prior information and balance between the prior information and observed data.

2.1. Extracting prior information

There are many sources of prior information. In this study, we focus on that contained in published studies, which can be broad and of relatively high quality. In particular, we use PubMed, which is one

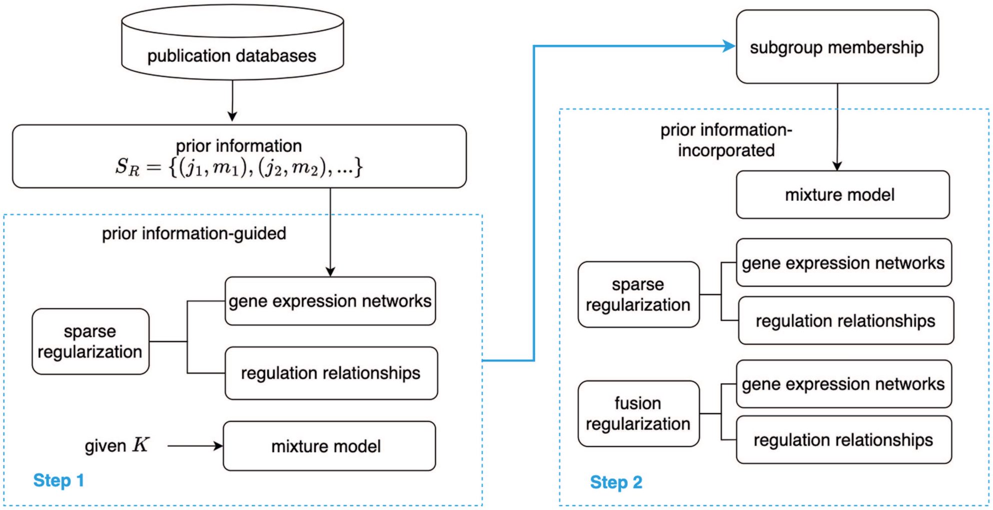

Figure 2. Flowchart of the proposed analysis.

of the most comprehensive publication databases. For a specific cancer (for example, breast cancer), we search PubMed for the cooccurrence of two genes (for example, EGFR and ERBB2). This can be realized using software such as PubMatrix and easyPubMed. Then a threshold is imposed to the counts of cooccurrence to retain the strongest evidence. Two genes are considered as having prior information of being interconnected if they have a nonzero count of cooccurrence (after the thresholding). It is noted that this coarse text mining can be potentially improved (for example, by normalizing using the occurrence counts of individual genes) and that the proposed analysis does not demand prior information to be fully accurate. Additional discussions are provided in the last section.

Denote $\mathcal{S}_R$ as the collection of gene expression relationships. In particular, if there exists prior information for the jth gene and mth gene, then $(j, m) \in \mathcal{S}_R$ (which corresponds to a network edge in the first step of analysis). We also set $(j, j) \in \mathcal{S}_R$. Here, it is noted that in the literature there have been relatively limited studies on heterogeneity analysis and so the same prior information $\mathcal{S}_R$ is shared by all sample subgroups.

2.2. Prior information-guided heterogeneity analysis

Denote n as the number of subjects. For the ith subject, denote $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^T$ as the gene expression measurements and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})^T$ as their regulators, which can be copy number variations, DNA methylation, microRNAs, and others. Here, if there are multiple types of regulators, we stack them together. Multiple published studies have shown that this simple strategy has satisfactory performance (Seal et al. 2020). Denote $\boldsymbol{X}$ as the design matrix composed of $\boldsymbol{x}_i$'s and an intercept. For heterogeneity analysis, denote K_0 as the number of sample subgroups. Different from some studies, it is not assumed that K_0 is known. To start with, we consider a mixture model with $K(\geq K_0)$ subgroups. In practice, although K_0 is unknown, specifying an "upper bound" is usually not difficult. Our numerical exploration suggests that the value of K is not critical as long as it is large enough. Then, the probability density function of $\boldsymbol{y}_i$, $i = 1, \ldots, n$ is:

$$\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}_i) = \sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l), \tag{2.1}$$

where $\boldsymbol{\Theta}_l$ represents the gene expression network in the lth subgroup, $\boldsymbol{\Gamma}_l$ represents the regulation relationships in the lth subgroup, and π_l is the mixing proportion satisfying $0 \leq \pi_l \leq 1$ and

$\sum_{l=1}^{K} \pi_l = 1$, $l = 1, \ldots, K$. This belongs to the mixture modeling framework, which has been extensively adopted for heterogeneity analysis (Hao et al. 2018; Ren et al. 2022). In the proposed analysis, each component of the mixture model and hence the heterogeneity structure is defined by a conditional Gaussian graphical model (CGGM) (Yin and Li 2011) incorporating the regulation relationships in the gene expression networks.

Consider linear regulations and Gaussian distributions for gene expressions. That is, in the lth subgroup, $\boldsymbol{y}_i = \boldsymbol{\Gamma}_l \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Theta}_l^{-1})$. Then, $f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l) = (2\pi)^{-2/p} |\boldsymbol{\Theta}_l|^{1/2} \exp\{-(\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)^T \boldsymbol{\Theta}_l (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)/2\}$. $\boldsymbol{\Theta}_l$ is the precision matrix (inverse of the covariance matrix), and its sparsity structure directly describes the gene expression network structure. $\boldsymbol{\Gamma}_l$ is the coefficient matrix and describes the regulation relationships. It is noted that the Gaussian assumption and GGM model have been extensively adopted for gene expressions (Wang and Huang 2014), and it is possible to relax such assumptions. Additionally, although nonlinear regulations have been considered, considering the high dimensionality and satisfactory performance, we follow the literature (Bersanelli et al. 2016) and consider linear regulations.

In the first step, we propose estimation:

$$(\hat{\boldsymbol{\Omega}}^p, \hat{\boldsymbol{\pi}}^p) = \arg\max_{\boldsymbol{\Omega}, \boldsymbol{\pi}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \left(\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_l) \right) - \mathcal{P}_{prior}(\boldsymbol{\Omega}) \right\}, \tag{2.2}$$

where $\boldsymbol{\Omega}_l = \text{vec}(\boldsymbol{\Theta}_l, \boldsymbol{\Gamma}_l)^T = (\theta_{11,l}, \ldots, \theta_{1p,l}, \ldots, \theta_{pp,l}, \gamma_{11,l}, \ldots, \gamma_{1(q+1),l}, \ldots, \gamma_{p(q+1),l})^T$, $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)^T$, and the penalty function is defined as:

$$\mathcal{P}_{prior}(\boldsymbol{\Omega}) = \sum_{l=1}^{K} \sum_{(j,m) \notin \mathcal{S}_R} p(|\theta_{jm,l}|, \alpha_1) + \sum_{l=1}^{K} \sum_{(j,m)} p(|\gamma_{jm,l}|, \alpha_2), \tag{2.3}$$

where $p(\cdot, \alpha)$ is a base penalty function with regularization parameter $\alpha > 0$. Convenient choices include MCP and SCAD. Note that, as $(j, j) \in \mathcal{S}_R$ for $j = 1, \ldots, p$, the diagonal elements in the precision matrices are not penalized. In practical data analysis, p and q are usually of the same order, which can be partly seen in our data analysis. Additionally, it is expected that a certain proportion of the genes have prior information. As such, the two terms in (2.3) are likely to have similar scales.

Here, we adopt a finite mixture modeling strategy with a prefixed number of subgroups. The objective function has two terms. The first term is the log-likelihood and measures goodness-of-fit. The second term is the sparsity penalty, which conducts regularized estimation and identification of important network connections and regulations. Here, we "fully trust" the prior information—the network edges in the prior information set are not subject to selection. Then the sparsity penalty searches for additional signals. With this step of estimation, we obtain K sample subgroups, the gene expression network and regulation relationships for each subgroup, and the mixture probabilities.

2.3. Prior information-incorporated heterogeneity analysis

With estimation (2.2) and the Bayesian rule, we can also obtain the subgroup membership for each subject. Consider the $n \times K$ membership matrix with each row corresponding to the subgroup identified for each subject. This matrix is denoted as $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$. For $i = 1, \ldots, n$ and $l = 1, \ldots, K$,

$$z_{il} = \begin{cases} 1, & l = \arg\max_{1 \le l \le K} \{\hat{\pi}_l^p f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^p)\}; \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

We propose objective function:

$$\mathcal{L}(\mathbf{\Omega}, \boldsymbol{\pi} | \mathbf{Y}, \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} (1-\eta) \log \left(\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Omega}_l) \right) + \frac{1}{n} \sum_{i=1}^{n} \eta \log \left(\sum_{l=1}^{K} z_{il} f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \mathbf{\Omega}_l) \right) - \mathcal{P}(\mathbf{\Omega}), \tag{2.5}$$

where $0 \leq \eta \leq 1$ is a data-dependent weighting parameter and the penalty:

$$\mathcal{P}(\mathbf{\Omega}) = \sum_{l=1}^{K} \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) + \sum_{l=1}^{K} \sum_{(j,m)} p(|\gamma_{jm,l}|, \lambda_2) + \sum_{l<l'} p\left((\|\mathbf{\Theta}_l - \mathbf{\Theta}_{l'}\|_F^2 + \|\mathbf{\Gamma}_l - \mathbf{\Gamma}_{l'}\|_F^2)^{1/2}, \lambda_3 \right).$$

Consider the estimate:

$$(\hat{\mathbf{\Omega}}, \hat{\boldsymbol{\pi}}) = \arg \max_{\mathbf{\Omega}, \boldsymbol{\pi}} \mathcal{L}(\mathbf{\Omega}, \boldsymbol{\pi} | \mathbf{Y}, \mathbf{X}). \tag{2.6}$$

It is noted that, with the fusion penalization, $\hat{\mathbf{\Omega}} = (\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_K)^T$ may contain identical values. Denote the number of unique values of $\hat{\mathbf{\Omega}}$ as $\hat{K}_0$, which provides an estimate of the number of subgroups. From this estimation, we can also obtain the estimated gene expression network and regulation relationships for each subgroup, mixture probabilities, as well as subgroup membership of each subject.

In (2.5), the first two terms measure goodness-of-fit and balance between the observed data and the prior information. The balancing is achieved with η. Intuitively, when the prior information has a low quality, $\eta \to 0$, and the analysis will be heavily based on the observed data. On the other hand, when the prior information has a high quality, $\eta \to 1$, and it puts more emphasis on borrowing strength from the prior information.

The proposed penalty has two major components. The first two terms achieve sparsity. Here, the edges in the prior information set are also subject to selection, which allows the proposed approach to screen out wrong prior information. The last term adopts a penalized fusion strategy for heterogeneity analysis. The intuition is that, if two of the K subgroups are "close enough," their parameters will be shrunk to be the same, and the two subgroups can be combined together. It is noted that both the gene expression networks and the regulation relationships are included in the fusion penalty and used for defining the subgroups, which significantly differs from the gene network-only heterogeneity analysis. $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ are treated as a group and simultaneously used to promote similarity, which can be more effective than being analyzed separately.

2.4. Computation

The two optimization problems (2.2) and (2.6) are solved sequentially. The objective function in (2.6) requires the solution of (2.2). For each problem, the expectation-maximization (EM) technique is adopted, with computing the conditional expectation of the complete data log-likelihood function in the expectation step and updating $(\hat{\mathbf{\Theta}}, \hat{\mathbf{\Gamma}})$ iteratively in the maximization step. The alternating direction method of multipliers (ADMM) technique (Boyd et al. 2011) is adopted, and the algorithm is summarized in Algorithm 1. The details for the M-steps are provided in Supplementary Materials.

Algorithm 1 Computational algorithm for the proposed method

```
Require: the observed data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$'s, $K$, and prior information $\mathcal{S}_R$.
Ensure: $\hat{K}_0$, mixture probabilities $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_{\hat{K}_0})$, and $(\hat{\boldsymbol{\Theta}}_1, \dots, \hat{\boldsymbol{\Theta}}_{\hat{K}_0}, \hat{\boldsymbol{\Gamma}}_1, \dots, \hat{\boldsymbol{\Gamma}}_{\hat{K}_0})$.
Initialization: $t = 0$, $\hat{\boldsymbol{\Omega}}^{(0)} = (\hat{\boldsymbol{\Theta}}^{(0)}, \hat{\boldsymbol{\Gamma}}^{(0)})$, and $\hat{\boldsymbol{\pi}}^{(0)}$.
The first step for solving (2.2):
while $\|\hat{\boldsymbol{\Omega}}^{(t)} - \hat{\boldsymbol{\Omega}}^{(t-1)}\|_2 \geq 10^{-4}$, do
    $t = t + 1$;
    E-step: calculate conditional expectation:

    $$\mathbb{E}_{\boldsymbol{L}|\boldsymbol{y},\boldsymbol{x},\hat{\boldsymbol{\Omega}}^{(t-1)}}[\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi}|\boldsymbol{Y}, \boldsymbol{X})] = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K} L_{il}^{(t)}[\log \pi_l + \log f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_l)] - \mathcal{P}_{prior}(\boldsymbol{\Omega}), \tag{2.7}$$

    where $L_{il}^{(t)} = \hat{\pi}_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^{(t-1)}\right) / \sum_{l=1}^{K} \hat{\pi}_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^{(t-1)}\right)$ depends on the estimates from the $(t-1)$th step.
    M-step: Update $\hat{\pi}_l^{(t)} = \sum_{i=1}^{n} L_{il}^{(t)}/n$, and update $\hat{\boldsymbol{\Gamma}}^{(t)} = (\hat{\boldsymbol{\Gamma}}_1^{(t)}, \dots, \hat{\boldsymbol{\Gamma}}_K^{(t)})$ and $\hat{\boldsymbol{\Theta}}^{(t)} = (\hat{\boldsymbol{\Theta}}_1^{(t)}, \dots, \hat{\boldsymbol{\Theta}}_K^{(t)})$ iteratively. For $\boldsymbol{\Gamma}_l$, $l = 1, \dots, K$, maximizing (2.7) is equivalent to solving:

    $$\left\{\hat{\boldsymbol{\Gamma}}^{(t)}\right\} = \arg\min_{\boldsymbol{\Gamma}} \left( \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{K} L_{il}^{(t)} \left\{ (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)^\top \hat{\boldsymbol{\Theta}}_l^{(t-1)} (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i) \right\} + \sum_{l=1}^{K}\sum_{(j,m)} p(|\boldsymbol{\gamma}_{jm,l}|, \alpha_2) \right).$$

    It can be achieved using the local quadratic approximation technique. For $\boldsymbol{\Theta}_l$, $l = 1, \dots, K$, it is equivalent to solving:

    $$\left\{\hat{\boldsymbol{\Theta}}^{(t)}\right\} = \arg\min_{\boldsymbol{\Theta}} \left( \sum_{l=1}^{K} n_l^{(t)} \left[ -\log\det(\boldsymbol{\Theta}_l) + \mathrm{tr}(\mathbf{S}_{\Gamma l}^{(t)} \boldsymbol{\Theta}_l) \right] + \sum_{l=1}^{K} \sum_{j \neq m, (j,m) \notin \mathcal{S}_R} p(|\theta_{jm,l}|, \alpha_1) \right),$$

    where $n_l^{(t)} = \sum_{i=1}^{n} L_{il}^{(t)}$ and $\mathbf{S}_{\Gamma l}^{(t)}$ is based on $L_{il}^{(t)}$ and $\hat{\boldsymbol{\Gamma}}^{(t)}$. It can be achieved using the ADMM technique.
end while
Obtain $\hat{\boldsymbol{\pi}}^p = \hat{\boldsymbol{\pi}}^{(t)}$ and $\hat{\boldsymbol{\Omega}}^p = \hat{\boldsymbol{\Omega}}^{(t)}$. Calculate $z_{il}$ according to (2.4) for $i = 1, \dots, n$ and $l = 1, \dots, K$. Revert $t = 0$, $\hat{\boldsymbol{\Omega}}^{(0)} = (\hat{\boldsymbol{\Theta}}^{(0)}, \hat{\boldsymbol{\Gamma}}^{(0)})$ and $\hat{\boldsymbol{\pi}}^{(0)}$ as the initial value.
The second step for solving (2.6):
while $\|\hat{\boldsymbol{\Omega}}^{(t)} - \hat{\boldsymbol{\Omega}}^{(t-1)}\|_2 \geq 10^{-4}$, do
    $t = t + 1$;
    E-step: calculate conditional expectation:

    $$\begin{aligned} \mathbb{E}_{\boldsymbol{L}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Omega}^{(t-1)}}[\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi}|\boldsymbol{Y}, \boldsymbol{X})] =& \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K} (1-\eta) L_{il}^{(t)} [\log \pi_l + \log f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l)] \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K} \eta z_{li} \log f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l) - \mathcal{P}(\boldsymbol{\Omega}), \end{aligned} \tag{2.8}$$

    where $L_{il}^{(t)} = \hat{\pi}_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^{(t-1)}\right) / \sum_{l=1}^{K} \hat{\pi}_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^{(t-1)}\right)$.
    M-step: Update $\hat{\pi}_l^{(t)} = \sum_{i=1}^{n} L_{il}^{(t)}/n$, and update $\hat{\boldsymbol{\Gamma}}^{(t)} = (\hat{\boldsymbol{\Gamma}}_1^{(t)}, \dots, \hat{\boldsymbol{\Gamma}}_K^{(t)})$ using the local quadratic approximation technique and $\hat{\boldsymbol{\Theta}}^{(t)} = (\hat{\boldsymbol{\Theta}}_1^{(t)}, \dots, \hat{\boldsymbol{\Theta}}_K^{(t)})$ using the ADMM algorithm iteratively.
end while
Denote the number of distinct values of $(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{\boldsymbol{\Gamma}}^{(t)})$ as $\hat{K}_0$. $\hat{\boldsymbol{\Omega}}_k = (\hat{\boldsymbol{\Theta}}_k^{(t)}, \hat{\boldsymbol{\Gamma}}_k^{(t)})$, $k = 1, \dots, \hat{K}_0$, and $\hat{\pi}_k = \sum_{l=1}^{K} I(\hat{\boldsymbol{\Omega}}_l^{(t)} = \hat{\boldsymbol{\Omega}}_k)\hat{\pi}_l^{(t)}$, $k = 1, \dots, \hat{K}_0$.
```

For selecting the optimal tunings α_1 and α_2 in the first step, we adopt the following BIC criterion and a grid search:

$$BIC = -2\sum_{i=1}^{n} \log\left[\sum_{l=1}^{K} \hat{\pi}_l^p f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \hat{\boldsymbol{\Omega}}_l^p)\right] + \sum_{l=1}^{K} \log(n) df_l, \tag{2.9}$$

where df_l is the total number of nonzero parameters in $\hat{\boldsymbol{\Omega}}_l^p$, $l = 1, \ldots, K$. In the second step, we need to determine $\lambda_1, \lambda_2, \lambda_3$, and η. We propose first fixing η and, for each candidate value of η, selecting the optimal $(\lambda_1, \lambda_2, \lambda_3)$ using the BIC criterion and a grid search. Then the optimal η can be selected also using the BIC criterion. It is noted that, with a more complex analysis goal, the tunings needed are more than some of the existing studies. However, published studies and our own experience suggest that such tuning parameter selection is feasible and generates reliable results. The code for the proposed algorithm is publicly available at https://github.com/lirong95/prior-cggm.

3. SIMULATION

Simulation is conducted to assess the performance of the proposed method and compare it against relevant alternatives. We set the true number of subgroups $K_0 = 3$, where different subgroups have distinct networks and regulation relationships. For the dimensions, we consider $p = q = 100$ and $p = q = 200$. For the sample sizes, we consider a balanced case with all the subgroups having sample sizes of 500 and an imbalanced case with the three subgroups having sample sizes of 250, 300, and 350. For the prior information, we consider a correctly specified case (denoted as T, where $\mathcal{S}_R$ is the intersection of the nonzero elements in the K_0 subgroups) and a partially mis-specified case (denoted as F, where the entries of $\mathcal{S}_R$ are selected at random following true/false positive rates (TPR/FPR) being 0.6/0.1).

3.1. Settings

The following two network structures are considered. ST1: The first subgroup has an upper-tridiagonal precision matrix with the diagonal elements equal to 1 and the nonzero off-diagonal elements equal to 0.4. The second subgroup has a lower-tridiagonal precision matrix with the diagonal elements equal to 1 and the nonzero off-diagonal elements equal to 0.4. The third subgroup has a diagonal precision matrix with the nonzero elements equal to 1. ST2: The precision matrices are generated by the nearest-neighbor networks. Specifically, each network consists of 10 equally-sized disjoint subnetworks (modules), among which eight are shared by the three sample subgroups. Additionally, the first subgroup shares one module with the second subgroup and another one with the third subgroup. The second subgroup and the third subgroup also have a unique module of their own. The structure of each module is generated by a nearest-neighbor network. We first generate $p/10$ points randomly on a unit square, calculate all $p/10 \times (p/10 - 1)/2$ pairwise distances, and select $m = 2$ nearest neighbors of each point besides itself. The nonzero off-diagonal elements of the precision matrices are located at which the corresponding two points are among the m nearest neighbors of each other. The nonzero values are generated from $\text{Unif}(-0.4, -0.1) \cup \text{Unif}(0.1, 0.4)$. The diagonal elements are all set as 1. ST1 has a chain-like structure, and ST2 has a module structure. They are graphically presented in Supplementary Fig. S1 (Supplementary Materials).

We simulate $\boldsymbol{x}_i$ as having a normal distribution $N(\boldsymbol{0}, \boldsymbol{I}_p)$ and a categorical distribution, where x_{ij} is generated randomly from $\{0, 1, 2\}$ with equal probabilities. In terms of regulations, the positions of the nonzero entries are randomly selected, and each entry has a probability proportional to $1/q$ of being nonzero. The nonzero values are generated from the uniform distribution $\text{Unif}(-1, -0.7) \cup \text{Unif}(0.7, 1)$.

The simulation settings are comprehensive. In particular, two types of network structures are considered, both of which are popular in the literature. Both continuous and categorical regulators are considered to mimic the distributions of genomic regulators encountered in practice. Two levels of prior information quality are considered, which may test the "robustness" of the proposed approach. It is noted that although the data dimensions may not seem that high, with the networks, regulations, and heterogeneity, the number of unknown parameters is significantly larger than the sample sizes. To better gauge performance, we consider the following relevant alternatives. (i) A two-step procedure. In the first step, a clustering method is used to generate subgroups. Here, we consider both the K-means clustering and a nonparametric clustering method (Chauveau and Hoang 2016). Both clustering methods are conducted on (X, Y) and only Y. The number of subgroups is set as $K = 2, 3, 6$. In the second step, we apply the CGGM approach with Lasso penalization (cglasso) (Yin and Li 2011). We denote them as K-cglasso-X, K-cglasso-(XY), np-cglasso-X and np-cglasso-(XY), respectively. (ii) HeteroGGM. The heterogeneous Gaussian graphical model via penalized fusion (HeteroGGM) approach (Ren et al. 2022) is applied. It can simultaneously achieve subgroup membership identification and precision matrix estimation. The number of subgroups is automatically determined by fusion regularization. It does not accommodate the regulations of $\boldsymbol{x}$ on $\boldsymbol{y}$ or the prior information. (iii) RI-HeteroGGM. We conduct the regulation-incorporated network-based heterogeneity analysis (Li et al. 2023). This method extends HeteroGGM to incorporate heterogeneous regulation relationships and can simultaneously obtain subgroup memberships and determine the number of subgroups, precision matrices, and coefficient matrices. It does not accommodate prior information. This alternative may be the closest to the proposed approach. (d) PI-CGGM. This heterogeneity analysis approach accommodates prior information. It conducts the mixture modeling + CGGM analysis with a fixed number of subgroups. It is the first step of the proposed approach and solves objective function (2.2). To facilitate comparison, the number of subgroups is set as $K = 6$.

3.2. Results

When implementing the proposed approach, we set $K = 6$. Similar results are obtained under other K values. We adopt the following measures to evaluate performance. For subgrouping accuracy, we consider $\hat{K}_0$ and adjusted Rand index (RI), which measures the similarity between the estimated and true subgrouping structures. For estimation accuracy, we consider root mean squared error (RMSE). Specifically, for the precision matrices,

$$\mathrm{RMSE}(\boldsymbol{\Theta}) = \begin{cases} \frac{1}{K_0}\sum_{k=1}^{K_0} \|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k^*\|_F & \hat{K}_0 = K_0, \\ \frac{1}{\hat{K}_0}\sum_{l=1}^{\hat{K}_0}\sum_{k=1}^{K_0} \|\hat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_k^*\|_F \cdot I & \\ \left(k = \arg\min_{k'}\{\|\hat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_{k'}^*\|_F^2 + \|\hat{\boldsymbol{\Gamma}}_l - \boldsymbol{\Gamma}_{k'}^*\|_F^2\}\right) & \hat{K}_0 \neq K_0. \end{cases}$$

For variable selection accuracy, we consider true/false positive rates (TPR/FPR):

$$\mathrm{TPR}(\boldsymbol{\Theta}) = \begin{cases} \frac{1}{K_0}\sum_{k=1}^{K_0} \frac{\sum_{j<m} I(\theta_{jm,k}^* \neq 0, \hat{\theta}_{jm,k} \neq 0)}{\sum_{j<m} I(\theta_{jm,k}^* \neq 0)} & \hat{K}_0 = K_0, \\ \frac{1}{\hat{K}_0}\sum_{l=1}^{\hat{K}_0}\sum_{k=1}^{K_0} \frac{\sum_{j<m} I(\theta_{jm,k}^* \neq 0, \hat{\theta}_{jm,l} \neq 0)}{\sum_{j<m} I(\theta_{jm,k}^* \neq 0)} \cdot I & \\ \left(k = \arg\min_{k'}\{\|\hat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_{k'}^*\|_F^2 + \|\hat{\boldsymbol{\Gamma}}_l - \boldsymbol{\Gamma}_{k'}^*\|_F^2\}\right) & \hat{K}_0 \neq K_0, \end{cases}$$

$$\mathrm{FPR}(\boldsymbol{\Theta}) = \begin{cases} \frac{1}{K_0}\sum_{k=1}^{K_0} \frac{\sum_{j<m} I(\theta_{jm,k}^* = 0, \hat{\theta}_{jm,k} \neq 0)}{\sum_{j<m} I(\theta_{jm,k}^* = 0)} & \hat{K}_0 = K_0, \\ \frac{1}{\hat{K}_0}\sum_{l=1}^{\hat{K}_0}\sum_{k=1}^{K_0} \frac{\sum_{j<m} I(\theta_{jm,k}^* = 0, \hat{\theta}_{jm,l} \neq 0)}{\sum_{j<m} I(\theta_{jm,k}^* = 0)} \cdot I & \\ \left(k = \arg\min_{k'}\{\|\hat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_{k'}^*\|_F^2 + \|\hat{\boldsymbol{\Gamma}}_l - \boldsymbol{\Gamma}_{k'}^*\|_F^2\}\right) & \hat{K}_0 \neq K_0. \end{cases}$$

The above measures are defined accordingly for $\boldsymbol{\Gamma}$. Performance is evaluated for $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ separately.

The simulation results for ST1, the normal distribution, and $p = q = 100$ are summarized in Table 1, and the other results are presented in Supplementary Tables S1–S7 (Supplementary Materials). The proposed method demonstrates competitive performance in subgrouping,

Table 1. Simulation results under ST1 with $\boldsymbol{x} \sim N(0, \boldsymbol{I}_p)$ and $p = q = 100$. In each cell, mean (SD).

n	Method	Prior		RMSE	TPR	FPR	RI	$\hat{K}_0$
(250,300,350)	Proposed	T	$\boldsymbol{\Theta}$	1.925(0.399)	1.000(0.001)	0.030(0.020)	0.988(0.036)	3.00(0.00)
			$\boldsymbol{\Gamma}$	2.746(0.804)	0.975(0.042)	0.009(0.008)		
		F	$\boldsymbol{\Theta}$	2.260(0.963)	0.995(0.013)	0.042(0.045)	0.898(0.240)	3.35(0.93)
			$\boldsymbol{\Gamma}$	3.359(2.131)	0.823(0.210)	0.016(0.010)		
	K-cglasso-Y ($K = 2$)	-	$\boldsymbol{\Theta}$	4.068(0.023)	1.000(0.000)	0.103(0.003)	0.001(0.000)	2(0)
			$\boldsymbol{\Gamma}$	7.481(0.030)	0.981(0.008)	0.039(0.001)		
	K-cglasso-Y ($K = 3$)	-	$\boldsymbol{\Theta}$	4.801(0.030)	0.938(0.014)	0.141(0.003)	0.003(0.003)	3(0)
			$\boldsymbol{\Gamma}$	7.743(0.128)	0.909(0.024)	0.060(0.001)		
	K-cglasso-Y ($K = 6$)	-	$\boldsymbol{\Theta}$	3.722(0.151)	0.971(0.014)	0.209(0.003)	0.013(0.007)	6(0)
			$\boldsymbol{\Gamma}$	7.971(0.198)	0.887(0.024)	0.110(0.002)		
	K-cglasso-(X, Y) ($K = 2$)	-	$\boldsymbol{\Theta}$	4.078(0.022)	1.000(0.000)	0.103(0.003)	0.001(0.000)	2(0)
			$\boldsymbol{\Gamma}$	7.480(0.024)	0.982(0.007)	0.039(0.001)		
	K-cglasso-(X, Y) ($K = 3$)	-	$\boldsymbol{\Theta}$	4.815(0.037)	0.939(0.013)	0.143(0.003)	0.003(0.002)	3(0)
			$\boldsymbol{\Gamma}$	7.751(0.120)	0.902(0.018)	0.059(0.001)		
	K-cglasso-(X, Y) ($K = 6$)	-	$\boldsymbol{\Theta}$	3.764(0.126)	0.974(0.011)	0.211(0.003)	0.014(0.008)	6(0)
			$\boldsymbol{\Gamma}$	7.921(0.231)	0.887(0.026)	0.109(0.003)		
	np-cglasso-Y ($K = 2$)	-	$\boldsymbol{\Theta}$	4.102(0.242)	0.999(0.002)	0.096(0.010)	0.101(0.161)	2(0)
			$\boldsymbol{\Gamma}$	6.594(1.230)	0.991(0.009)	0.036(0.005)		
	np-cglasso-Y ($K = 3$)	-	$\boldsymbol{\Theta}$	4.696(0.151)	0.952(0.019)	0.138(0.005)	0.043(0.073)	3(0)
			$\boldsymbol{\Gamma}$	7.207(0.507)	0.930(0.035)	0.058(0.003)		
	np-cglasso-Y ($K = 6$)	-	$\boldsymbol{\Theta}$	3.858(0.166)	0.958(0.019)	0.209(0.002)	0.016(0.008)	6(0)
			$\boldsymbol{\Gamma}$	7.800(0.221)	0.894(0.026)	0.110(0.002)		
	np-cglasso-(X, Y) ($K = 2$)	-	$\boldsymbol{\Theta}$	4.110(0.139)	0.998(0.007)	0.103(0.004)	0.002(0.004)	2(0)
			$\boldsymbol{\Gamma}$	7.431(0.094)	0.984(0.010)	0.039(0.001)		
	np-cglasso-(X, Y) ($K = 3$)	-	$\boldsymbol{\Theta}$	4.819(0.032)	0.939(0.015)	0.142(0.005)	0.003(0.003)	3(0)
			$\boldsymbol{\Gamma}$	7.690(0.129)	0.902(0.020)	0.060(0.001)		
	np-cglasso-(X, Y) ($K = 6$)	-	$\boldsymbol{\Theta}$	3.779(0.150)	0.965(0.016)	0.212(0.003)	0.011(0.007)	6(0)
			$\boldsymbol{\Gamma}$	7.998(0.183)	0.875(0.022)	0.110(0.001)		
	HeteroGGM	-	$\boldsymbol{\Theta}$	4.654(0.286)	0.999(0.004)	0.839(0.031)	0.178(0.260)	3.15(1.56)
			$\boldsymbol{\Gamma}$	-	-	-		
	RI-HeteroGGM	-	$\boldsymbol{\Theta}$	2.596(0.518)	0.999(0.001)	0.042(0.013)	0.689(0.161)	2.20(0.41)
			$\boldsymbol{\Gamma}$	3.545(0.702)	0.914(0.051)	0.011(0.004)		
	PI-CGGM	T	$\boldsymbol{\Theta}$	3.169(0.220)	0.999(0.002)	0.054(0.009)	0.801(0.046)	6(0)
			$\boldsymbol{\Gamma}$	5.523(0.653)	0.661(0.117)	0.008(0.006)		
		F	$\boldsymbol{\Theta}$	5.355(0.383)	0.954(0.028)	0.277(0.025)	0.538(0.206)	6(0)
			$\boldsymbol{\Gamma}$	6.664(1.015)	0.536(0.093)	0.008(0.013)		

Continued

Table 1. Continued

n	Method	Prior		RMSE	TPR	FPR	RI	$\hat{K}_0$
(500,500,500)	Proposed	T	$\boldsymbol{\Theta}$	0.855(0.121)	1.000(0.000)	0.004(0.001)	1.000(0.000)	3.00(0.00)
			$\boldsymbol{\Gamma}$	0.909(0.124)	0.991(0.006)	0.001(0.001)		
		F	$\boldsymbol{\Theta}$	1.066(0.275)	1.000(0.000)	0.003(0.001)	1.000(0.000)	3.00(0.00)
			$\boldsymbol{\Gamma}$	1.056(0.341)	0.985(0.011)	0.002(0.001)		
	K-cglasso-Y $(K = 2)$	-	$\boldsymbol{\Theta}$	4.376(0.346)	0.994(0.007)	0.061(0.003)	0.001(0.000)	2(0)
			$\boldsymbol{\Gamma}$	7.511(0.226)	0.994(0.006)	0.057(0.009)		
	K-cglasso-Y $(K = 3)$	-	$\boldsymbol{\Theta}$	4.376(0.346)	0.994(0.007)	0.061(0.003)	0.002(0.003)	3(0)
			$\boldsymbol{\Gamma}$	7.511(0.226)	0.994(0.006)	0.057(0.009)		
	K-cglasso-Y $(K = 6)$	-	$\boldsymbol{\Theta}$	4.414(0.149)	0.971(0.013)	0.162(0.002)	0.014(0.006)	6(0)
			$\boldsymbol{\Gamma}$	7.656(0.189)	0.934(0.020)	0.070(0.001)		
	K-cglasso-(X, Y) $(K = 2)$	-	$\boldsymbol{\Theta}$	4.338(0.339)	0.995(0.009)	0.063(0.002)	0.001(0.000)	2(0)
			$\boldsymbol{\Gamma}$	7.579(0.257)	0.988(0.012)	0.054(0.012)		
	K-cglasso-(X, Y) $(K = 3)$	-	$\boldsymbol{\Theta}$	4.979(0.024)	0.972(0.013)	0.090(0.002)	0.002(0.003)	3(0)
			$\boldsymbol{\Gamma}$	7.675(0.100)	0.958(0.011)	0.037(0.001)		
	K-cglasso-(X, Y) $(K = 6)$	-	$\boldsymbol{\Theta}$	4.452(0.170)	0.965(0.015)	0.164(0.003)	0.013(0.006)	6(0)
			$\boldsymbol{\Gamma}$	7.679(0.208)	0.926(0.023)	0.069(0.002)		
	np-cglasso-Y $(K = 2)$	-	$\boldsymbol{\Theta}$	3.990(0.693)	0.998(0.004)	0.046(0.010)	0.284(0.198)	2(0)
			$\boldsymbol{\Gamma}$	5.381(1.253)	0.996(0.007)	0.034(0.017)		
	np-cglasso-Y $(K = 3)$	-	$\boldsymbol{\Theta}$	2.946(0.818)	0.999(0.004)	0.038(0.022)	0.597(0.235)	3(0)
			$\boldsymbol{\Gamma}$	3.272(1.586)	0.998(0.006)	0.020(0.018)		
	np-cglasso-Y $(K = 6)$	-	$\boldsymbol{\Theta}$	4.370(0.184)	0.977(0.010)	0.152(0.007)	0.071(0.048)	6(0)
			$\boldsymbol{\Gamma}$	6.695(0.548)	0.972(0.011)	0.065(0.004)		
	np-cglasso-(X, Y) $(K = 2)$	-	$\boldsymbol{\Theta}$	4.597(0.326)	0.995(0.005)	0.060(0.004)	0.006(0.007)	2(0)
			$\boldsymbol{\Gamma}$	7.218(0.387)	0.994(0.011)	0.056(0.011)		
	np-cglasso-(X, Y) $(K = 3)$	-	$\boldsymbol{\Theta}$	4.951(0.051)	0.978(0.009)	0.090(0.003)	0.009(0.010)	3(0)
			$\boldsymbol{\Gamma}$	7.457(0.226)	0.968(0.017)	0.036(0.001)		
	np-cglasso-(X, Y) $(K = 6)$	-	$\boldsymbol{\Theta}$	4.496(0.164)	0.966(0.014)	0.164(0.004)	0.015(0.006)	6(0)
			$\boldsymbol{\Gamma}$	7.620(0.218)	0.928(0.023)	0.070(0.001)		
	HeteroGGM	-	$\boldsymbol{\Theta}$	4.416(0.077)	1.000(0.000)	0.834(0.045)	0.690(0.207)	4.85(1.31)
			$\boldsymbol{\Gamma}$	-	-	-		
	RI-HeteroGGM	-	$\boldsymbol{\Theta}$	0.984(0.150)	1.000(0.000)	0.009(0.003)	1.000(0.000)	3.00(1.00)
			$\boldsymbol{\Gamma}$	1.204(0.374)	0.977(0.016)	0.001(0.000)		
	PI-CGGM	T	$\boldsymbol{\Theta}$	4.798(4.831)	0.999(0.001)	0.061(0.014)	0.805(0.060)	6(0)
			$\boldsymbol{\Gamma}$	3.267(0.701)	0.861(0.066)	0.003(0.001)		
		F	$\boldsymbol{\Theta}$	5.196(0.613)	0.983(0.015)	0.234(0.032)	0.757(0.043)	6(0)
			$\boldsymbol{\Gamma}$	3.319(0.630)	0.858(0.041)	0.004(0.001)		

selection, and estimation, across the whole spectrum of simulation scenarios. Specifically, when the prior information is correct, the proposed method can accurately identify the number of subgroups and achieve desirable estimation and selection accuracy. When the prior information is partially misspecified, the performance remains competitive compared to the approaches without incorporating prior information (i.e. RI-HeteroGGM). This suggests that the proposed approach has the "robustness" property—it can data-dependently adjust the impact of prior information. The alternative methods have inferior performance. HeteroGGM, which does not take the regulations into account, tends to over-estimate the number of subgroups and over-select the nonzero elements in the precision matrices. The estimation of the two-step procedure heavily depends on the subgrouping results, and it performs acceptably only when the number of subgroups is correctly specified—this is highly challenging in practice.

We also conduct a simulation experiment to assess the effects of prior information on the final estimation. We consider the setting with continuous regulators, ST1 precision matrices, dimensions $p = q = 100$, and sample sizes $n = (300, 300, 300)$. We consider varying prior information quality, as measured by the TPR/FPR values, with a larger TPR and a smaller FPR indicating a higher quality of prior information. The results are summarized in Supplementary Table S8 (Supplementary Materials). It is observed that the performance of subgrouping, estimation, and selection deteriorates as the degree of misspecification in prior information increases, which is as expected. It is also observed that, even when the prior information is completely wrong (with TPR=0 and FRP=1), the proposed method still performs comparably to those without accommodating prior information. This is due to the weighting strategy. Additional simulations are conducted to more deeply comprehend the impact of weight η. The results in Supplementary Table S9 (Supplementary Materials) suggest that higher-quality prior information corresponds to a larger η, which is highly sensible.

4. BREAST CANCER DATA ANALYSIS

Breast cancer is among the most extensively studied using high-throughput profiling techniques, and there have been a handful of multiomics breast cancer studies. Here, we analyze the METABRIC data and refer to the original publications (Curtis et al. 2012; Pereira et al. 2016; Rueda et al. 2019) for details on the study and experimental designs. The dataset contains gene expression and copy number variation measurements on 1,898 subjects. Copy number variation has been long recognized as a critical regulator of gene expression. Although in principle the proposed analysis can be conducted using a large number of genes, to generate more reliable results, we focus on the "most interesting" genes. In particular, we consider genes in the PAM50 set (which has been manually curated and suggested as highly relevant for breast cancer subtyping) and in the KEGG "breast cancer" pathway—this leads to 154 genes that are highly likely to be relevant for breast cancer biology. We then identify the corresponding copy number variations. For prior information mining, we use the R package `easyPubMed`. The search involves breast cancer and any two genes out of the 154. The cooccurrence counts are presented in Fig. 3. Prior information is available for about 30% of the gene pairs. To focus on more reliable prior information, we impose a threshold of 10, which leads to a total of 195 gene pairs. More detailed information is available from the authors. Here, it is noted that the cutoff of 10 can be somewhat subjective. However, this may not pose a serious concern with the flexibility of the proposed method.

When implementing the proposed method, we set $K = 6$. A total of four subgroups are identified, with sizes 782, 448, 439 and 229, respectively. The detailed membership information is available from the authors. The estimated gene expression networks and regulation relationships are shown in Supplementary Fig. S2 (Supplementary Materials). In Supplementary Table S11 (Supplementary Materials), we present the numbers of network edges and the numbers of overlapping edges. We further present the DeltaCon distances in parentheses to measure the similarity of the corresponding two networks (Tantardini et al. 2019). It is observed that the four subgroups have significantly different network structures. Subgroups 1 and 3 are the most similar in terms of gene

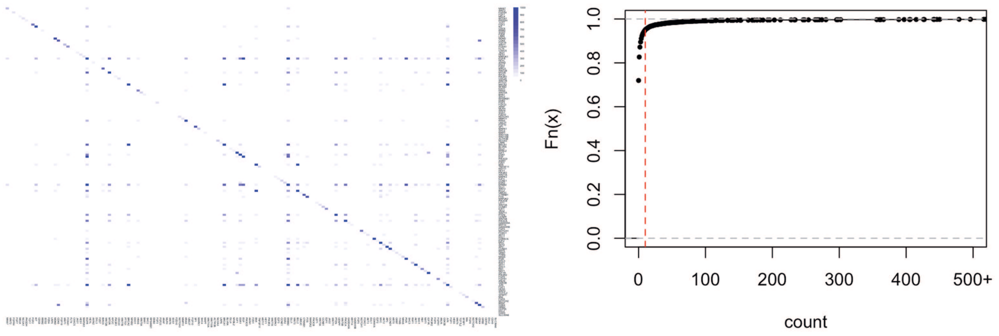

Figure 3. Prior information. Left: heatmap; Right: cumulative distribution function (red dotted line corresponds to count=10).

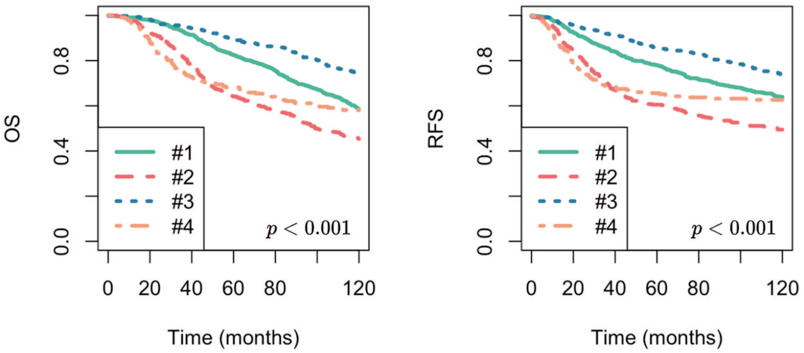

Figure 4. Comparison of survival for the four subgroups. Left: overall survival (OS); Right: relapse free survival (RFS).

expression networks. More comparisons between the networks are presented in Supplementary Fig. S3. It is observed that the network of subgroup 2 has more high-degree nodes, which indicates higher direct connectivity. The network of subgroup 3 has more high-betweeness nodes, which indicates higher influence of information passing. Across the four subgroups, on average, about half of the edges in $\mathcal{S}_R$ are identified. Additionally, the likelihood for an edge to be identified is correlated with the cooccurrence count.

The proposed analysis is unsupervised, and there is a lack of an objective way to evaluate subgrouping accuracy. To get additional insights, we compare some key clinical features across the four subgroups. Supplementary Table S10 (Supplementary Materials) presents the results of the Chi-squared tests with FDR adjustment for the Nottingham prognostics index (NPI), number of lymph nodes examined positive (LNP), and age at diagnosis (Age). In Fig. 4, we further compare overall survival (OS) and relapse free survival (RFS). Significant differences are observed across the four subgroups, which can provide "indirect support" to the sensibility of analysis. Breast cancer has been subtyped based on molecular biomarkers. A commonly adopted is the Claudin subtyping, under which breast cancer is classified as Basal-like, Her2, Luminal A, Luminal B, and Normal-like. We compare the obtained subgrouping with the Claudian subtyping, and the results are summarized in Supplementary Table S12 (Supplementary Materials). Consider Basal-like and Luminal A, both of which are divided into two of the identified subgroups. To get further insights, we compare the Basal-like subtype within subgroup 2 and subgroup 4, as well as the Luminal A subtype within subgroup 1 and subgroup 2. The results are presented in Supplementary Fig. S4 (Supplementary Materials). It is observed that the Luminal A subtype within subgroup 3 has a significantly better prognosis than that within subgroup 1. Additionally, the Basal-like subtype within subgroup 2 has a significantly poorer prognosis than that within subgroup 4. This suggests that the proposed analysis may lead to clinically meaningful findings that can complement the existing subtyping. An interesting finding is that compared to the Basal-like subtype within subgroup 2, the Basal-like subtype within subgroup 4 exhibits enriched expression in growth factor signaling, particularly involving EGFR, MET, BRAF, and CTNNB1, as illustrated in Supplementary Fig. S5 (Supplementary Materials). This observation is consistent with the known characteristics of BL2 (a subtype of Basal-like breast cancer) documented in the literature (Hubalek et al. 2017).

This dataset is also analyzed using the alternative approaches. We fix the number of subgroups as $K = 4$ for better comparability. The Rand index results are presented in Supplementary Table S13 (Supplementary Materials), which suggests that different approaches lead to significantly different subgrouping structures. For each approach, we compare survival across the identified subgroups and present the results in Supplementary Table S14 (Supplementary Materials). It is observed that the proposed approach can better separate the subjects into subgroups with more distinct survival, which can provide "indirect support" to the validity of the proposed approach.

5. DISCUSSION

In this study, we have developed a novel heterogeneity analysis approach that is based on both gene expression networks and gene expression-regulator relationships. Advancing from the existing literature, we have proposed a way to effectively and flexibly incorporate prior information contained in vast publications. Simulation and the analysis of a breast cancer dataset have demonstrated the practical utility of the proposed approach.

This study can be extended in multiple ways. The proposed analysis is not limited to "gene expressions + regulators." For example, it is directly applicable to "protein expressions + gene expressions" with ligand-receptor pairs and protein-protein interaction (PPI) networks as the prior information. When there are multiple types of regulators, there have been a few developments in more subtly integrating them (as opposed to directly stacking them together). It is also possible to refine text mining and extract higher-quality prior information. For example, the convolutional neural network (CNN) technique developed in Wang et al. (2023) can be a viable choice. Some domain-specific language representation models, such as BioBERT pre-trained on large-scale biomedical corpora, can be further applied to enhance the extraction of prior information (Lee et al. 2020). Additionally, integrating some natural language processing techniques with web-based tools, such as GeneDive (Previde et al. 2018), can also facilitate the exploration of gene interactions. Theoretical exploration, such as the identifiability of the mixture model and consistency properties, may also be of interest (Ho and Nguyen 2016; Balakrishnan et al. 2017). The proposed strategy for incorporating prior information has been motivated by several recent successes. It is possible to develop other information-incorporating strategies. It will also be of interest to develop more data analysis.

ACKNOWLEDGMENTS

We thank the reviewers, associate editor, and editor for their careful review and insightful comments.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Biostatistics Journal* online.

FUNDING

This work was supported in part by the MOE Project of Key Research Institute of Humanities and Social Sciences [22JJD910001], National Natural Science Foundation of China [72271237], NIH [CA204120], and NSF [2209685].

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

Balakrishnan S, Wainwright MJ, Yu B. Statistical guarantees for the EM algorithm: from population to sample-based analysis. Ann Stat. 2017:45(1):77–120.

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016:17(2):167–177.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn. 2011:3(1):1–122.

Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Hodgson JG, Weinrich S, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. J Pathol. 2013:231(1):63–76.

Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. 2013:501(7467):338–345.

Chauveau D, Hoang VTL. Nonparametric mixture models with conditionally independent multivariate component densities. Comput Stat Data Anal. 2016:103:1–16.

Church BV, Williams HT, Mar JC. Investigating skewness to understand gene expression heterogeneity in large patient cohorts. BMC Bioinformatics. 2019:20(24):1–14.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012:486(7403):346–352.

Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan J-B, Zhang K, Chun J, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016:13(3):241–244.

Hao B, Sun WW, Liu Y, Cheng G. Simultaneous clustering and estimation of heterogeneous graphical models. J Mach Learn Res. 2018:18(217):1–58.

Henao JD, Lauber M, Azevedo M, Grekova A, Theis F, List M, Ogris C, Schubert B. Multi-omics regulatory network inference in the presence of missing data. Brief Bioinf. 2023:24(5):bbad309.

Ho N, Nguyen X. On strong identifiability and convergence rates of parameter estimation in finite mixtures. Electronic J Stat. 2016:10(1):271–307.

Hubalek M, Czech T, Müller H. Biological subtypes of triple-negative breast cancer. Breast Care. 2017: 12(1):8–14.

Jiang Y, He Y, Zhang H. Variable selection with prior information for generalized linear models via the prior lasso method. J Am Stat Assoc. 2016:111(513):355–376.

Kagohara LT, Stein-O'Brien GL, Kelley D, Flam E, Wick HC, Danilova LV, Easwaran H, Favorov AV, Qian J, Gaykalova DA, et al. Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. Brief Funct Genomics. 2018:17(1):49–63.

Kang T, Moore R, Li Y, Sontag E, Bleris L. Discriminating direct and indirect connectivities in biological networks. Proc Natl Acad Sci USA, 2015:112(41):12893–12898.

Lee D, Park Y, Kim S. Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. Brief Bioinf. 2021:22(3):1–19.

Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020:36(4):234–1240.

Li R, Zhang Q, Ma S. Regulation-incorporated gene expression network-based heterogeneity analysis, arXiv, arXiv:2308.03946, 2023, preprint: not peer reviewed.

Li Y, Xu S, Ma S, Wu M. Network-based cancer heterogeneity analysis incorporating multi-view of prior information. Bioinformatics. 2022:38(10):2855–2862.

Meeks JJ, Al-Ahmadie H, Faltas BM, Taylor III JA, Flaig TW, DeGraff DJ, Christensen E, Woolbright BL, McConkey DJ, Dyrskjøt L. Genomic heterogeneity in bladder cancer: challenges and possible solutions to improve outcomes. Nat Rev Urol. 2020:17(5):259–270.

Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, Riggs M, Eberling Y, Troge T, Grubor V, et al. Inferring tumor progression from genomic heterogeneity. Genome Res. 2010:20(1):68–80.

Pereira B, Chin S-F, Rueda OM, Vollan H-Km, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut S-J, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun. 2016:7(1):1–16.

Pio G, Mignone P, Magazzù G, Zampieri G, Ceci M, Angione C. Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction. Bioinformatics. 2022:38(2): 487–493.

Previde P, Thomas B, Wong M, Mallory EK, Petkovic D, Altman RB, Kulkarni A. Genedive: a gene interaction search and visualization tool to facilitate precision medicine. Pacific Symposium on Biocomputing 2018. Singapore: World Scientific, 2018, pp. 590–601.

Ren M, Zhang S, Zhang Q, Ma S. Gaussian graphical model-based heterogeneity analysis via penalized fusion. Biometrics. 2022:78(2):524–535.

Rueda OM, Sammut S-J, Seoane JA, Chin S-F, Caswell-Jin J-L, Callari M, Batra R, Pereira B, Bruna A, Ali AR, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019:567(7748):399–404.

Seal DB, Das V, Goswami S, De RK. Estimating gene expression from dna methylation and copy number variation: a deep learning regression model for multi-omics integration. Genomics. 2020:112(4):2833–2841.

Tang J, Kong D, Cui Q, Wang K, Zhang D, Gong Y, Wu G. Prognostic genes of breast cancer identified by gene co-expression network analysis. Front Oncol. 2018:8:374.

Tantardini M, Ieva F, Tajoli L, Piccardi C. Comparing methods for comparing networks. Sci Rep. 2019:9(1):17557.

Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. Nat Comput Sci. 2021:1(6):395–402.

Wang F, Liang D, Li Y, Ma S. Prior information-assisted integrative analysis of multiple datasets. Bioinformatics. 2023:39(8):btad452.

Wang YR, Huang H. Review on statistical methods for gene network reconstruction using expression data. J Theor Biol. 2014:362:53–61.

Yin J, Li H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. Ann Appl Stat. 2011:5(4):2630–2650.

Zhao Y, Zhu H, Lu Z, Knickmeyer RC, Zou F. Structured genome-wide association studies with Bayesian hierarchical variable selection. Genetics. 2019:212(2):397–415.