Statistica Sinica Preprint No: SS-2023-0109	
Title	Regulation-Incorporated Gene Expression Network-Based Heterogeneity Analysis
Manuscript ID	SS-2023-0109
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0109
Complete List of Authors	Rong Li, Qingzhao Zhang and Shuangge Ma
Corresponding Authors	Shuangge Ma
E-mails	shuangge.ma@yale.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

REGULATION-INCORPORATED GENE EXPRESSION NETWORK-BASED HETEROGENEITY ANALYSIS

Rong Li[1], Qingzhao Zhang[2], Shuangge Ma[1]

[1] *Yale University*

[2] *Xiamen University*

Abstract: Gene expression-based heterogeneity analysis has been extensively conducted. In recent studies, it has been shown that network-based analysis, which takes a system perspective and accommodates the interconnections among genes, can be more informative than that based on simpler statistics. Gene expressions are regulated. Incorporating regulations in analysis can better delineate the "sources" of gene expression effects. Although conditional network analysis can somewhat serve this purpose, it does not render enough attention to the regulation relationships. In this article, significantly advancing from the existing heterogeneity analyses based only on gene expression networks, conditional gene expression network analyses, and regression-based heterogeneity analyses, we propose heterogeneity analysis based on gene expression networks (after accounting for or "removing" regulation effects) as well as regulations of gene expressions. A high-dimensional penalized fusion approach is proposed, which can determine the number of sample groups and parameter values in a single step. An effective computational algorithm is proposed. It is rigorously proved that the proposed approach enjoys the estimation, selection, and grouping consistency properties. Extensive simulations demonstrate its practical superiority over closely related alternatives.

In the analysis of two breast cancer datasets, the proposed approach identifies heterogeneity and gene network structures different from the alternatives and with sound biological implications.

Key words and phrases: Heterogeneity analysis, Gene expression network, Regulation, Penalization.

1. Introduction

Many complex diseases are intrinsically heterogeneous, with samples having the same disease diagnosis behaving differently. In early studies, heterogeneity analysis is often based on low-dimensional clinical and demographic measurements. With the development of high-throughput profiling, omics measurements, which may more informatively capture disease biology, have been increasingly used in heterogeneity analysis (Lee et al., 2021). Among the various omics measurements, gene expressions have drawn special attention because of important biological implications, broad availability of data, and promising empirical results. Through a series of studies (Church et al., 2019; Pio et al., 2022), gene expression-based heterogeneity analysis has demonstrated significant successes. It can be supervised and unsupervised, and the two types of analysis serve different purposes. In this study, we conduct unsupervised heterogeneity analysis, under which different sample groups have different gene expression properties.

Some gene expression-based heterogeneity analyses, especially some early ones (Leek and Storey, 2007; Church et al., 2019), are based on simple data characteristics such as mean and variance. In recent studies, it has been shown that gene expression network (graph)-based analysis can take a system perspective and lead to more informative heterogeneity structures (Tang et al., 2018; Pio et al., 2022). Here it is noted that network-based heterogeneity analysis can also accommodate information on mean and variance. The existing network-based heterogeneity analysis studies are mostly based on the Gaussian Graphical Model (GGM) technique, and there have been two main families of approaches. The first family is based on the finite mixture model technique (Hao et al., 2018), and a common challenge is how to determine the number of sample groups. The second family is based on the fusion technique (Radchenko and Mukherjee, 2017), which may provide a more "straightforward" way of determining the number of groups.

Gene expressions are regulated by multiple types of regulators (methylation, microRNAs, etc.). Published studies have suggested that the interconnections among gene expressions, as reflected in networks, can be attributed to regulators as well as "net connections" (Kagohara et al., 2018). As schematically presented in Figure 1, gene expression networks without accounting for regulators (left two plots) can be denser and hence less lucid than those accounting for regulatory effects (right two plots). With the growing popularity of multiomics studies (that collect data

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

on gene expressions and their regulators), multiple strategies/approaches have been developed for the collective analysis of gene expression and regulator data. Examples include pooling multiple types of data and jointly modeling (Lee et al., 2021), analyzing regulation relationships for example using regression (Seal et al., 2020), and others. In the context of network analysis, conditional approaches, for example conditional GGM, have been developed for studying gene expression interconnections with account for regulators (Yin and Li, 2011; Sohn and Kim, 2012; Cai et al., 2013; Wang, 2015). In conditional analysis, especially under the context of heterogeneity analysis, regulation relationships often serve as a "middle step" and do not play an important role (Huang et al., 2018; Lartigue et al., 2021).

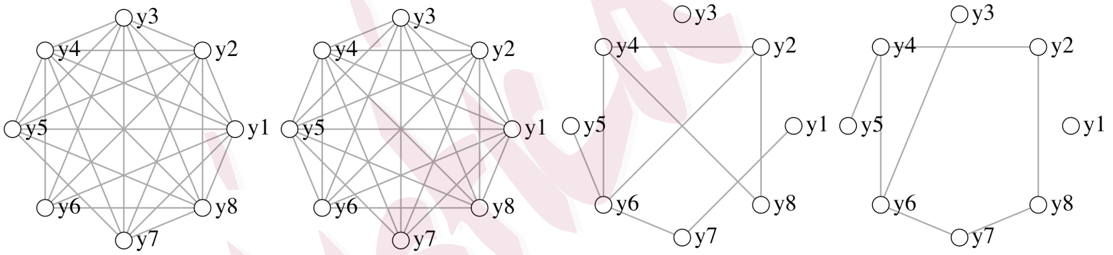

Figure 1: Schematic example: gene expression networks for two sample groups before (left two) and after (right two) accounting for regulators.

In this article, our goal is to further advance gene expression network-based heterogeneity analysis by developing a new approach that can more effectively accommodate regulator data. This has been made possible by the increasing availability of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

multiomics data and motivated by the successes of existing gene expression network-based heterogeneity analyses as well as their limitation in accounting for regulation relationships. This study has a solid ground. Specifically, it belongs to the family of network-based heterogeneity analysis techniques and can enjoy similar merits as Danaher et al. (2014) and Hao et al. (2018). It is built on the GGM technique – note that, following Cai et al. (2013), the normality assumption can be relaxed to make the proposed analysis more broadly applicable. Similar to Ren et al. (2022), it is built on the sparse penalized fusion technique (Ma and Huang, 2017), can "automatically" determine the number of sample groups, and has advantages over the finite mixture modeling approaches.

On the other hand, this study also advances from the existing ones in multiple important aspects. First, it considers gene expression network interconnections after accounting for regulator effects. As schematically shown in published studies (Wytock and Kolter, 2013) and the right two plots of Figure 1, such interconnections can be sparser and easier to interpret. Additionally, they may reflect more essential gene relationships (Sohn and Kim, 2012). Second, significantly different from most conditional analyses, we more explicitly model the gene expression-regulator relationships and, more importantly, include such relationships in defining the heterogeneity structure. This has been motivated by the findings that such regulations have important implications for exploring the "hidden" sources of variations in complex diseases,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

as dysregulations can be directly associated with disease risk, progression, prognosis, etc. Published literature has also stressed that identifying heterogeneous genetic regulatory mechanisms is essential in the precision medicine era (Kagohara et al., 2018). For the example in Figure 1, some additional analysis results are reported in Section S1 (Supplementary Materials), which may further suggest the advantage and necessity of incorporating regulators. Third, with both gene expression networks and gene expression-regulator relationships, and along with heterogeneity, the proposed method differs significantly from the existing ones in both the mixture and penalized fusion structures. In the proposed analysis, each component of the mixture corresponds to a conditional GGM, and both the networks and regulation relationships are subject to fusion penalization, which introduces additional technical challenges. The proposed method can simplify to some existing ones, for example, by not accommodating regulators, by focusing on heterogeneous distribution means, and by assuming a known number of groups. A brief comparison of the different methods is provided in Table S2 (Supplementary Materials). Computational and theoretical developments, although may share some similar spirit with the existing studies, can be more complicated and demand careful investigations. Last but not least, as demonstrated in our data analysis, this study may deliver a practically useful new approach and findings for deciphering heterogeneity of complex diseases. It is noted that, although developed in the context of gene expressions and their regulators, the proposed analysis can have much broader applications. For example, for some other types of omics data (e.g., proteins), heterogeneity analysis and network-based analysis have also been conducted, and there are also upstream measurements with regulatory relationships. Another example may be human disease network analysis, where demographic variables, environmental factors, lifestyle, and others can be viewed as “regulators”.

The rest of this article is organized as follows. In Section 2, we introduce the proposed method and present its rationale, computation, and theoretical properties. Simulation study is conducted in Section 3 to gauge performance and compare with alternatives. Data analysis is presented in Section 4. Concluding remarks are presented in Section 5. Additional computational, theoretical, and numerical results are provided in Supplementary Materials.

2. Methods

Suppose that the observations $(\boldsymbol{y}_i, \boldsymbol{x}_i), i = 1, \ldots, n$ are independent, where $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{ip})$ and $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{i,q+1})$. In our analysis, $\boldsymbol{y}_i$'s are gene expressions, and $\boldsymbol{x}_i$'s are regulators such as microRNAs, copy number variations, and DNA methylation, all of which can significantly affect gene expression levels. As noted in conditional analysis and other analyses (Tabor et al., 2002), the collection of regulators does not need to be complete, in the sense that $\boldsymbol{x}_i$ does not need to include

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

all or a specific set of regulators. Let $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)$ denote the deterministic design matrix. Assume that the n subjects belong to K_0 groups defined based on gene expression networks and/or gene expression-regulator relationships, where the group memberships and value of K_0 are unknown. For the l-th group, consider the regulation model:

$$\boldsymbol{y} = \boldsymbol{\Gamma}_l \boldsymbol{x} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\Gamma}_l$ is the $p \times (q+1)$ coefficient matrix, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_l$. Let $\boldsymbol{\Theta}_l = \boldsymbol{\Sigma}_l^{-1}$ be the l-th precision matrix and $\boldsymbol{\Omega}_l = \text{vec}(\boldsymbol{\Gamma}_l, \boldsymbol{\Theta}_l) = (\gamma_{11,l}, \ldots, \gamma_{1(q+1),l}, \ldots, \gamma_{p(q+1),l}, \theta_{11,l}, \ldots, \theta_{1p,l}, \ldots, \theta_{pp,l})^T$ be its vectorized representation. Conditional on $\boldsymbol{x}$, assume that $\boldsymbol{y}$ follows a multivariate normal distribution $N(\boldsymbol{\Gamma}_l \boldsymbol{x}, \boldsymbol{\Theta}_l^{-1})$, that is,

$$f_l(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Omega}_l) = (2\pi)^{-p/2} |\boldsymbol{\Theta}_l|^{1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\Gamma}_l \boldsymbol{x})^T \boldsymbol{\Theta}_l (\boldsymbol{y} - \boldsymbol{\Gamma}_l \boldsymbol{x}) \right\}.$$

Here samples in the same group share the same precision matrix and the same coefficient matrix.

Although it is difficult to know K_0, it is often easy to specify an "upper bound" $K > K_0$. To be cautious, K can be taken as a relatively large number. With K groups, we denote $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \cdots, \boldsymbol{\Omega}_K)^T$ and consider the mixture distribution:

$$f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Omega}) = \sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Omega}_l),$$

where π_l is the mixture probability of the l-th group. Denote $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, which is also unknown.

For parameter estimation and determination of the heterogeneity structure, we propose the penalized objective function:

$$\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi} | \boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \sum_{l=1}^{K} \pi_l f_l(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_l) \right\} - \mathcal{P}(\boldsymbol{\Omega}). \tag{2.1}$$

Here, the penalty is proposed as:

$$\begin{aligned} \mathcal{P}(\boldsymbol{\Omega}) = & \sum_{l=1}^{K} \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) + \sum_{l=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{q+1} p(|\gamma_{jm,l}|, \lambda_2) \\ & + \sum_{l<l'} p \left\{ (\|\boldsymbol{\Theta}_l - \boldsymbol{\Theta}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l - \boldsymbol{\Gamma}_{l'}\|_F^2)^{1/2}, \lambda_3 \right\}. \end{aligned} \tag{2.2}$$

$\theta_{jm,l}$ is the jm-th entry of the l-th precision matrix $\boldsymbol{\Theta}_l$. $\gamma_{jm,l}$ is the jm-th entry of the l-th coefficient matrix. $\| \cdot \|_F$ is the Frobenius norm. $p(\cdot, \lambda)$ is a penalty function with regularization parameter $\lambda > 0$. Convenient choices are MCP and SCAD. The proposed estimator $(\hat{\boldsymbol{\Omega}}, \hat{\boldsymbol{\pi}})$ is defined as the maximizer of $\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{\pi} | \boldsymbol{Y}, \boldsymbol{X})$. Denote $\hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_{\hat{K}_0}$ as the unique values of $\hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_K$. Then it is concluded that there are $\hat{K}_0$ groups, with corresponding parameter values $\hat{\boldsymbol{\Omega}}_1, \ldots, \hat{\boldsymbol{\Omega}}_{\hat{K}_0}$. The sparsity patterns of the precision matrix estimates directly correspond to the structures of the networks. Specifically, if and only if the (j, m)-th entry of the estimate for $\boldsymbol{\Theta}_l$ is zero, the corresponding two genes are not connected conditional on the other genes after removing the shared effects of the regulators in the l-th sample group. The sparsity of the estimate of $\boldsymbol{\Gamma}_l$ describes the sparse regulations of the regulators on the gene expressions in the l-th sample group. Note that the regulations can flexibly vary across genes, which corresponds to the different sparsity patterns across the rows of $\mathbf{\Gamma}_l$. This can be partly seen from the sparsity structure of the coefficient matrices in the real data analysis. The estimated mixture probabilities can be obtained accordingly. The proposed approach is characterized by the simultaneous estimation of the precision matrices, coefficient matrices, and group memberships. Its brief flowchart is presented in Figure S1 (Supplementary Materials).

Rationale The proposed modeling has two components: gene expression network and gene expression-regulator relationship. For the first component, we adopt the GGM technique as in multiple published studies. For the second component, we adopt linear regression. Although nonlinear regulations have been proposed, linear regression can be preferred considering the high dimensionality of gene expressions and regulators. It has also been shown to have satisfactory performance (Yin and Li, 2011; Cai et al., 2013). We adopt penalization for regularized estimation and selection. In (2.2), the first two sparse penalties have been commonly adopted (Rothman et al., 2010; Yin and Li, 2011). With the third penalty term, we start with $K(> K_0)$ sample groups and examine if two groups can be shrunk together. By examining the final estimates, we can directly obtain the estimated number of groups as well as model parameters for all groups. The fusion strategy has been adopted in multiple recent heterogeneity analyses and shown to be advantageous over multiple alternatives. Significantly different from the existing heterogeneity analysis (Ren et al., 2022), in (2.2), the regression coefficient matrices are considered along with the precision matrices – that is, the heterogeneity structure is jointly defined by the gene expression networks and regulation relationships.

2.1 Computation

For optimization, we develop an EM + Altering Direction Method of Multipliers (ADMM) algorithm. The complete data log-likelihood function is:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\omega_{il}\left\{\log\pi_k+\log f_l(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Omega}_l)\right\},$$

where ω_{il} is the latent indicator variable showing the group membership of the ith sample in the mixture. The EM algorithm maximizes the objective function composed of the above complete data log-likelihood function and penalty function in (2.2) iteratively in the following two steps.

Expectation step: here, we compute the conditional expectation of the complete data log-likelihood function with respect to ω_{il}, given the observed data $(\boldsymbol{y}_i,\boldsymbol{x}_i)$'s and current estimates from the $(t-1)$-th step. The conditional expectation is:

$$\mathbb{E}_{\boldsymbol{L}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Omega}^{(t-1)}}\{\mathcal{L}(\boldsymbol{\Omega},\boldsymbol{\pi}|\boldsymbol{Y},\boldsymbol{X})\}=\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}L_{il}^{(t)}\left\{\log\pi_l+\log f_l\left(\boldsymbol{y}_i;\boldsymbol{x}_i,\boldsymbol{\Omega}_l\right)\right\}-\mathcal{P}(\boldsymbol{\Omega}), \tag{2.3}$$

where $L_{il}^{(t)}$ is the conditional expectation of ω_{il}, which depends on the estimates from the $(t-1)$-th step and can be computed as:

$$L_{il}^{(t)} = \frac{\pi_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_l^{(t-1)}\right)}{\sum_{l=1}^K \pi_l^{(t-1)} f_l\left(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Omega}_l^{(t-1)}\right)}. \tag{2.4}$$

Maximization step: we maximize (2.3) with respect to $(\boldsymbol{\Omega}, \boldsymbol{\pi})$. For $\boldsymbol{\pi}$, we have:

$$\pi_l^{(t)} = \frac{1}{n}\sum_{i=1}^n L_{il}^{(t)}. \tag{2.5}$$

For $\boldsymbol{\Omega}$, we update $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_K)^T$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_K)^T$ separately. For $\boldsymbol{\Gamma}_l,\ l = 1, \dots, K$, maximizing (2.3) is equivalent to solving:

$$\begin{aligned}\left\{\boldsymbol{\Gamma}^{(t)}\right\} = \arg\min_{\boldsymbol{\Gamma}} \Bigg\{ & \frac{1}{2n}\sum_{i=1}^n \sum_{l=1}^K L_{il}^{(t)} \left\{(\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)^T \boldsymbol{\Theta}_l^{(t-1)} (\boldsymbol{y}_i - \boldsymbol{\Gamma}_l \boldsymbol{x}_i)\right\} \\ & + \sum_{l=1}^K \sum_{j=1}^p \sum_{m=1}^{q+1} p(|\gamma_{jm,l}|, \lambda_2) + \sum_{l<l\prime} p\left\{(\|\boldsymbol{\Theta}_l - \boldsymbol{\Theta}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l - \boldsymbol{\Gamma}_{l'}\|_F^2)^{1/2}, \lambda_3\right\} \Bigg\}.\end{aligned} \tag{2.6}$$

We adopt the local quadratic approximation technique. Details are provided in Supplementary Materials. For $\boldsymbol{\Theta}_l,\ l = 1, \dots, K$, maximizing (2.3) is equivalent to solving:

$$\begin{aligned}\left\{\boldsymbol{\Theta}^{(t)}\right\} = \arg\min_{\boldsymbol{\Theta}} \Bigg\{ & \sum_{l=1}^K n_l^{(t)} \left\{-\log\det(\boldsymbol{\Theta}_l) + \operatorname{tr}(\mathbf{S}_{\Gamma l}^{(t)} \boldsymbol{\Theta}_l)\right\} \\ & + \sum_{l=1}^K \sum_{j \neq m} p(|\theta_{jm,l}|, \lambda_1) + \sum_{l<l'} p\left\{(\|\boldsymbol{\Theta}_l - \boldsymbol{\Theta}_{l'}\|_F^2 + \|\boldsymbol{\Gamma}_l - \boldsymbol{\Gamma}_{l'}\|_F^2)^{1/2}, \lambda_3\right\} \Bigg\},\end{aligned} \tag{2.7}$$

where $\mathbf{S}_{\Gamma l}^{(t)} = \mathbf{C}_{yl}^{(t)} - \mathbf{C}_{yx,l}^{(t)} \boldsymbol{\Gamma}_l^{(t)T} - \boldsymbol{\Gamma}_l^{(t)} \mathbf{C}_{yx,l}^{(t)T} + \boldsymbol{\Gamma}_l^{(t)} \mathbf{C}_{xl}^{(t)} \boldsymbol{\Gamma}_l^{(t)T}$, $n_l^{(t)} = \sum_{i=1}^n L_{il}^{(t)}$, and $\mathbf{C}_{yl}$,

$\mathbf{C}_{xl}$, and $\mathbf{C}_{yx,l}$ are weighted covariance matrices:

$$\mathbf{C}^{(t)}_{yl} = \sum_{i=1}^{n} L^{(t)}_{il}\boldsymbol{y}_i\boldsymbol{y}_i^T / \sum_{i=1}^{n} L^{(t)}_{il}, \mathbf{C}^{(t)}_{yx,l} = \sum_{i=1}^{n} L^{(t)}_{il}\boldsymbol{y}_i\boldsymbol{x}_i^T / \sum_{i=1}^{n} L^{(t)}_{il}, \mathbf{C}^{(t)}_{xl} = \sum_{i=1}^{n} L^{(t)}_{il}\boldsymbol{x}_i\boldsymbol{x}_i^T / \sum_{i=1}^{n} L^{(t)}_{il}.$$

This optimization is achieved using the ADMM technique (Supplementary Materials). Overall, the algorithm contains iterating (2.4), (2.5), (2.6) and (2.7). The iteration is concluded when the difference between the estimates from two consecutive steps is smaller than a prefixed constant. Satisfactory convergence is observed in all of our numerical studies. For initial values, we resort to the nonparametric mixture approach (Chauveau and Hoang, 2016) and observe satisfactory performance. With the additional complexity of the proposed approach, the computation is inevitably more complicated than some of the existing ones. However, it is still feasible and affordable. Additional information on computational cost is provided in Table S10 (Supplementary Materials).

Tuning parameter selection For selecting the optimal λ_1, λ_2 and λ_3, we conduct a grid search and minimize the Hannan-Quinn information criterion (HQC):

$$-2\sum_{i=1}^{n}\log\left\{\sum_{k=1}^{\hat{K}_0}\hat{\pi}_k f_k(\boldsymbol{y}_i;\boldsymbol{x}_i,\hat{\boldsymbol{\Gamma}}_k,\hat{\boldsymbol{\Theta}}_k)\right\} + \sum_{k=1}^{\hat{K}_0}\log\{\log(n)\}df_k, \tag{2.8}$$

where df_k is the total number of nonzero parameters in $\hat{\boldsymbol{\Gamma}}_k$ and $\hat{\boldsymbol{\Theta}}_k, k = 1, \ldots, \hat{K}_0$. The HQC criterion has been shown to have satisfactory performance in the literature. In our limited numerical studies, it is found to have comparable or better performance than some other criteria. A systematic investigation of the tuning parameter selection criterion is beyond our scope.

2.2 Theoretical properties

Denote the true parameter values as $\boldsymbol{\Upsilon}^* = (\boldsymbol{\Upsilon}_1^*, \ldots, \boldsymbol{\Upsilon}_{K_0}^*)^T$ and $\boldsymbol{\Upsilon}_k^* = \mathrm{vec}(\boldsymbol{\Gamma}_k^*, \boldsymbol{\Theta}_k^*)$ for $k = 1, \ldots, K_0$. Define $\mathcal{S}_k = \{(j,m) : \theta_{jm,k}^* \neq 0,\ 1 \leq j \neq m \leq p\}$ and the sparsity parameter $s = \max\{|\mathcal{S}_k|,\ k = 1, \ldots, K_0\}$. Similarly, define $\mathcal{D}_k = \{(j,m) : \gamma_{jm,k}^* \neq 0,\ 1 \leq j \leq p,\ 1 \leq m \leq q+1\}$ and $d = \max\{|\mathcal{D}_k|,\ k = 1, \ldots, K_0\}$. The following conditions are assumed.

(C1) For some positive constants β_1 and β_2, $0 < \beta_1 < \min_{1\leq k\leq K_0} \lambda_{\min}(\boldsymbol{\Theta}_k^*) < \max_{1\leq k\leq K_0} \lambda_{\max}(\boldsymbol{\Theta}_k^*) < \beta_2$, where $\lambda_{\min}(\boldsymbol{\Theta}_k^*)$ and $\lambda_{\max}(\boldsymbol{\Theta}_k^*)$ are the smallest and largest eigenvalues of $\boldsymbol{\Theta}_k^*$, respectively.

(C2) $\|\boldsymbol{\Theta}^*\|_\infty = \max_{k=1,\ldots,K_0} \|\boldsymbol{\Theta}_k^*\|_\infty$ and $\|\boldsymbol{\Gamma}^*\|_\infty = \max_{k=1,\ldots,K_0} \|\boldsymbol{\Gamma}_k^*\|_\infty$ are bounded.

(C3) The design matrix $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_{q+1})$ satisfies $\max_j \|\mathbf{X}_j\|_2 = O(\sqrt{n})$, $j = 1, \ldots, q+1$. For each $k = 1, \ldots, K_0$, let $\mathbf{X}_{\mathcal{D}_k}$ be the sub-matrix of $\mathbf{X}$ with the support of coefficient matrix $\mathcal{D}_k$, and $\mathbf{X}_{\mathcal{D}_k^C}$ is the corresponding complement. Define $L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}^*) = \pi_k f_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}_k^*) / \sum_{k=1}^K \pi_k f_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}_k^*)$, $\mathbb{E}\{L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}^*)\} = \int L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}^*) f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Upsilon}_k^*) d\boldsymbol{y}$ and $\mathbf{G}_k = \mathrm{diag}\left[\mathbb{E}\{L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*)\}\right]$ is a $n \times n$ diagonal matrix with $\mathbb{E}\{L_k(\boldsymbol{y}_i; \boldsymbol{x}_i, \boldsymbol{\Upsilon}^*)\}$ as its elements. Denote $\|\mathbf{B}\|_{2,\infty} = \max_{\|v\|_2} \|\mathbf{B}\mathbf{v}\|_\infty$.

For a positive constant C_0 and $\alpha_1 \in [0, 1/2)$,

$$\lambda_{\min}(\mathbf{X}_{\mathcal{D}_k}^T \mathbf{G}_k \mathbf{X}_{\mathcal{D}_k}/n) \geq C_0, \qquad \left\| (\mathbf{X}_{\mathcal{D}_k^C}^T \mathbf{G}_k \mathbf{X}_{\mathcal{D}_k})(\mathbf{X}_{\mathcal{D}_k}^T \mathbf{G}_k \mathbf{X}_{\mathcal{D}_k})^{-1} \right\|_{2,\infty} \leq O(n^{\alpha_1}).$$

(C4) Minimal signal condition:

$$\min\left\{ \{|\gamma_{jm,k}^*| : (j,m) \in \mathcal{D}_k, k = 1, \ldots, K_0\}, \{|\theta_{jm,k}^*| : (j,m) \in \mathcal{S}_k, k = 1, \ldots, K_0\} \right\}$$
$$> (a + 0.5) \cdot \max\{\lambda_1, \lambda_2\}.$$

Denote $b = \min_{1 \leq k \neq k' \leq K_0} \|\boldsymbol{\Upsilon}_k^* - \boldsymbol{\Upsilon}_{k'}^*\|_2$. Then, $b > (a + 0.5)\lambda_3$.

(C5) $\lambda_1 \gg \sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}$, $\lambda_2 \gg \sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}$, and $\lambda_3 \gg \sqrt{\frac{(d+s+p)(\log p + \log q)}{n}}$.

(C6) The K_0 clusters are sufficiently separable such that, with a small $\gamma > 0$,

$$L_k(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}^*) \cdot L_j(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\Upsilon}^*) \leq \frac{\gamma}{24(K_0 - 1)\sqrt{\max\{W, W', W''\}}},$$

for each pair $\{(j,k), 1 \leq j \neq k \leq K_0\}$. Here, $W = \max_{1 \leq k \leq K_0} W_k$, $W' = \max_{1 \leq k \leq K_0} W_k'$, and $W'' = \max_{1 \leq k \leq K_0} W_k''$, and for each $k = 1, \ldots, K_0$,

$$W_k = \sup_{t \in [0,1]} \mathbb{E}\left\{ \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y})^T \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y}) \|\boldsymbol{\Theta}_k^*(\boldsymbol{y} - \boldsymbol{\Gamma}_k^* \boldsymbol{x})\boldsymbol{x}\|_F^2 \right\},$$

$$W_k' = \sup_{t \in [0,1]} \mathbb{E}\left\{ \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y})^T \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y}) \|\boldsymbol{\Theta}_k^{*-1}\|_F^2 \right\},$$

$$W_k'' = \sup_{t \in [0,1]} \mathbb{E}\left\{ \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y})^T \delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y}) \|(\boldsymbol{y} - \boldsymbol{\Gamma}_k^* \boldsymbol{x})(\boldsymbol{y} - \boldsymbol{\Gamma}_k^* \boldsymbol{x})^T\|_F^2 \right\}.$$

Define $\tilde{\boldsymbol{\Upsilon}}_t = \boldsymbol{\Upsilon}^* + t(\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^*)$, $\tilde{\boldsymbol{\Upsilon}}_t = (\tilde{\boldsymbol{\Upsilon}}_{t,1}, \ldots, \tilde{\boldsymbol{\Upsilon}}_{t,K_0})$, $\tilde{\boldsymbol{\Upsilon}}_{t,k} = \mathrm{vec}(\tilde{\boldsymbol{\Gamma}}_{t,k}, \tilde{\boldsymbol{\Theta}}_{t,k})$ with $t \in [0,1]$, and for any $\boldsymbol{\Upsilon} \in \mathcal{B}_{\alpha_0}(\boldsymbol{\Upsilon}^*) = \{\boldsymbol{\Upsilon} : \|\boldsymbol{\Upsilon} - \boldsymbol{\Upsilon}^*\|_2 \leq \alpha_0\}$:

$$\delta_{\boldsymbol{\Upsilon}_{t,k}}(\boldsymbol{y}) = \begin{pmatrix} \mathrm{vec}\left\{ \tilde{\boldsymbol{\Theta}}_{t,k}(\boldsymbol{y} - \tilde{\boldsymbol{\Gamma}}_{t,k}\boldsymbol{x})\boldsymbol{x} \right\} \\ \frac{1}{2}\mathrm{vec}\left\{ \tilde{\boldsymbol{\Theta}}_{t,k}^{-1} - (\boldsymbol{y} - \tilde{\boldsymbol{\Gamma}}_{t,k}\boldsymbol{x})(\boldsymbol{y} - \tilde{\boldsymbol{\Gamma}}_{t,k}\boldsymbol{x})^T \right\} \end{pmatrix}.$$

(C7) $\rho(t) = \lambda^{-1}p(t, \lambda)$ is concave in $t \in [0, +\infty)$ with a continuous derivative $\rho'(t)$ satisfying $\rho(0+) = 1$, and $\rho'(0+)$ is independent of λ. There exists a constant $0 < a < \infty$ such that $\rho(t)$ is constant for all $t \geq a\lambda$.

Conditions (C1) and (C2) have been commonly assumed in the GGM literature particularly including those on heterogeneity analysis (Hao et al., 2018). The boundedness condition on the coefficients is also common for high-dimensional regression. Condition (C3) is on the design matrix and controls the correlations between variables as well as the correlations between unimportant and important variables in each sample group. It is similar to Condition 4 in Fan and Lv (2011). Condition (C4) specifies the minimal signals and minimal differences across the sample groups. Condition (C5) specifies the orders of the tuning parameters. Condition (C6) is the sufficiently separable condition and requires that a sample belongs to a group with a probability close to either zero or one. Relevant discussions can be found in Hao et al. (2018). This condition is similar to Condition 3 in Cai et al. (2021). It has been shown that, under a simpler two-component mixture model, the separable condition reduces to the commonly adopted signal-to-noise condition (Balakrishnan et al., 2017). Condition (C7) has been commonly assumed for penalized estimation/selection and is satisfied by SCAD and MCP.

Theorem 1: Suppose that Conditions (C1)-(C7) hold. If additionally $(d + s + p)(\log p + \log q)/n = o(1)$, then there exists a local maximizer of (2.1) such that,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

with probability tending to 1:

1. $\hat{K}_0 = K_0$.

2. $\sum_{k=1}^{\hat{K}_0}\left(\|\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k^*\|_F + \|\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}_k^*\|_F\right) = O_p\left(\sqrt{\frac{(d+s+p)(\log p+\log q)}{n}}\right)$.

3. Define $\hat{\mathcal{D}}_k = \{(j,m) : \hat{\gamma}_{jm,k} \neq 0\}$ and $\hat{\mathcal{S}}_k = \{(j,m) : \hat{\theta}_{jm,k} \neq 0\}$. Then $\hat{\mathcal{D}}_k = \mathcal{D}_k$ and $\hat{\mathcal{S}}_k = \mathcal{S}_k$ for $k = 1, \ldots, \hat{K}_0$.

This theorem shows that the proposed approach has the well-desired consistency properties. Specifically, it can consistently identify the number of sample groups, which has not been established in quite a few existing heterogeneity analysis studies. Additionally, it has estimation and variable selection consistency. Here we note that, as in many published studies, only local convergence is established. Global convergence will demand additional conditions and investigations. Although the obtained results are not "surprising" and somewhat similar to those in the existing literature, the present data and model settings are much more complicated and include the existing ones (for example, GGM-based heterogeneity analysis (Hao et al., 2018) and under homogeneity, and high-dimensional regression-based heterogeneity analysis (Sun et al., 2022)) as special cases, and the theoretical developments are not trivial. The proof is provided in Supplementary Materials.

3. Simulation

Simulation is conducted to gauge performance of the proposed approach and compare against relevant alternatives. We set $K_0 = 3$ and consider dimensions $p = q = 50$ and $p = q = 100$. For sample sizes, we consider three cases: a balanced case with all groups having sample sizes 200, a balanced case with all groups having sample sizes 500, and an imbalanced case with the three groups having sample sizes 150, 200 and 250. Additionally, the following three settings are considered.

(S1) $\boldsymbol{x}_i$ has the first element being 1, and the other elements follow a normal distribution $N(\mathbf{0}, \mathbf{I}_q)$. The coefficient matrices $\Gamma_1 \neq \Gamma_2 \neq \Gamma_3$. 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 $\mathrm{Unif}(-1.5, -1) \bigcup \mathrm{Unif}(1, 1.5)$. All sample groups have tridiagonal precision matrices with the diagonal elements equal to 1 and the nonzero off-diagonal elements equal to 0.2, 0.3, and 0.4 for the three sample groups, respectively.

(S2) 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 groups. Additionally, the first group shares one module with the second group and another one with the third group. The second group and the third group also have a unique module

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

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 $\text{Unif}(-0.4,-0.1)\bigcup\text{Unif}(0.1,0.4)$. The diagonal elements are all set to 1. The other settings are the same as S1.

(S3) $\boldsymbol{x}_i$'s have categorical distributions. Specifically, x_{ij} is generated randomly from $\{0,1,2\}$ with equal probabilities. The other settings are the same as S1.

The sample size and dimensionality settings are comparable to those in the literature. With the presence of both networks and high-dimensional regressions under heterogeneity, our simulation can be considerably more challenging. It is noted that, although p and q may not seem large, with the precision and regression coefficient matrices for multiple sample groups, the number of unknown parameters is considerably larger than the sample size. Both continuous and categorical regulators are simulated, mimicking, for example, methylation and copy number variation. Two types of network structures are considered, both of which are popular in the literature. When implementing the proposed method, we set $K=6$ – we have also

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0109

experimented with a few other values and found similar results. To gauge its performance, we consider the following close alternatives. It is noted that there can be other alternatives. However, the following can be more relevant.

(a) The strategy is to first conduct clustering and generate sample groups. Then estimation is conducted for each group separately. This can be the most natural choice with the existing tools. Specifically, we use a nonparametric mixture approach (Chauveau and Hoang, 2016) for clustering, which outperforms K-means and many other clustering methods. The clustering is based on Y. It is found that it outperforms that based on (X, Y), which can be caused by the high dimensionality and additional noises. The number of groups is set as $K = 3, 4, 6$, as there is not a simple way for determining its value. For estimation with each group, we apply the conditional Gaussian graphical approach with Lasso penalization (CGLasso) (Yin and Li, 2011). Tuning parameter selection is conducted using BIC as proposed in the literature.

(b) This approach is similar to (a), except that the sparse multivariate regression with covariance estimation (MRCE) approach (Rothman et al., 2010) is applied for estimation after clustering.

(c) The mixture of conditional Gaussian graphical model (MCGGM) approach (Lartigue et al., 2021) is applied. With a given number of clusters, it can achieve simultaneous clustering and estimation of the precision matrices as well as the correlation matrices (between gene expressions and regulators). Note that this approach estimates the mutual correlation matrices between $\boldsymbol{x}$ and $\boldsymbol{y}$, not the regression coefficient matrices Γ's.

(d) The heterogeneous Gaussian graphical model via penalized fusion (HeteroGGM) approach (Ren et al., 2022) is applied. It can simultaneously achieve clustering and precision matrix estimation. The number of groups is automatically determined in a way similar to the proposed approach. However, it cannot accommodate the regulations of $\boldsymbol{x}$ on $\boldsymbol{y}$.

To evaluate performance, we consider the following measures. For grouping accuracy, we consider $\hat{K}_0$ and adjusted Rand index (RI), which measures the similarity between the estimated and true grouping structures. For estimation accuracy, we consider root mean square error (RMSE). Specifically, for the precision matrices,

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

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

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

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

The above measures are defined accordingly for the coefficient matrices.

To get some intuition into performance of the proposed and alternative approaches, in Figure 2, for one simulation replicate, we compare grouping performance of the different approaches. It is clear that, for this specific replicate, the proposed approach has higher grouping accuracy. Further, in Figures S2 and S3 (Supplementary Materials), we consider one simulation replicate under S1 and S2, respectively. Additionally, we consider two sample size settings. It is observed that the proposed approach generates significantly different network estimations for different sample groups. Under S1 with a relatively simpler structure, performance is already satisfactory under the smaller sample size setting. Under S2, we observe a significant improvement in identification accuracy when sample size increases.

More definitive results are based on 100 replicates for each setting. The summary results for setting S1 and $p = q = 50$ are provided in Table 1. The results for the other settings are presented in Tables S3-S7 in Supplementary Materials. The proposed approach is observed to have competitive performance across the whole spectrum of simulation. As a representative example, we consider Table 1, the setting with group sizes (150, 200, 250). The proposed approach is able to accurately identify the

Table 1: Simulation results under S1 with $p = q = 50$. In each cell, mean (sd).

n	Method		RMSE	TPR	FPR	RI	$\hat{K}_0$
(200,200,200)	Proposed	$\boldsymbol{\Theta}$	1.740(0.613)	0.943(0.047)	0.057(0.014)		
		$\boldsymbol{\Gamma}$	1.531(1.206)	0.961(0.045)	0.008(0.015)	0.994(0.027)	3.05(0.22)
	HeteroGGM	$\boldsymbol{\Theta}$	4.146(0.175)	0.980(0.011)	0.894(0.009)		
		$\boldsymbol{\Gamma}$	-	-	-	0.569(0.250)	4.85(1.27)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	3.632(0.114)	0.861(0.017)	0.266(0.011)		
		$\boldsymbol{\Gamma}$	6.083(0.344)	0.962(0.021)	0.108(0.006)	0.249(0.096)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	3.581(0.254)	0.915(0.012)	0.209(0.027)		
		$\boldsymbol{\Gamma}$	4.666(0.855)	0.989(0.013)	0.074(0.011)	0.443(0.137)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	3.638(0.493)	0.938(0.018)	0.173(0.048)		
		$\boldsymbol{\Gamma}$	4.491(1.591)	0.989(0.011)	0.056(0.017)	0.502(0.176)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	3.870(0.122)	0.905(0.021)	0.406(0.024)		
		$\boldsymbol{\Gamma}$	6.876(0.347)	0.803(0.054)	0.114(0.016)	0.249(0.096)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	3.412(0.373)	0.951(0.021)	0.292(0.034)		
		$\boldsymbol{\Gamma}$	4.677(1.452)	0.907(0.034)	0.156(0.024)	0.443(0.137)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	3.197(0.871)	0.971(0.014)	0.278(0.027)		
		$\boldsymbol{\Gamma}$	4.166(2.313)	0.987(0.015)	0.118(0.011)	0.502(0.176)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	3.952(0.614)	0.794(0.202)	0.118(0.094)		
		$\boldsymbol{\Gamma}$	-	-	-	0.395(0.192))	3(0)
(150,200,250)	Proposed	$\boldsymbol{\Theta}$	1.457(0.130)	0.950(0.015)	0.058(0.006)		
		$\boldsymbol{\Gamma}$	0.937(0.512)	0.990(0.021)	0.003(0.002)	1.000(0.000)	3.00(0.00)
	HeteroGGM	$\boldsymbol{\Theta}$	4.152(0.142)	0.986(0.010)	0.899(0.013)		
		$\boldsymbol{\Gamma}$	-	-	-	0.547(0.244)	4.50(1.50)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	3.681(0.158)	0.889(0.020)	0.263(0.018)		
		$\boldsymbol{\Gamma}$	5.919(0.699)	0.967(0.025)	0.105(0.010)	0.212(0.072)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	3.458(0.282)	0.926(0.018)	0.200(0.027)		
		$\boldsymbol{\Gamma}$	4.464(0.813)	0.991(0.019)	0.068(0.011)	0.417(0.119)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	3.645(0.533)	0.948(0.016)	0.176(0.050)		
		$\boldsymbol{\Gamma}$	4.619(1.766)	0.987(0.021)	0.057(0.019)	0.431(0.232)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	3.929(0.623)	0.905(0.028)	0.343(0.040)		
		$\boldsymbol{\Gamma}$	6.535(0.709)	0.839(0.085)	0.146(0.022)	0.212(0.072)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	3.196(0.324)	0.957(0.011)	0.311(0.045)		
		$\boldsymbol{\Gamma}$	4.485(0.870)	0.977(0.042)	0.160(0.018)	0.417(0.119)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	3.346(0.650)	0.973(0.012)	0.284(0.046)		
		$\boldsymbol{\Gamma}$	4.557(1.894)	0.973(0.055)	0.145(0.019)	0.431(0.232)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	4.355(1.579)	0.768(0.195)	0.120(0.097)		
		$\boldsymbol{\Gamma}$	-	-	-	0.394(0.166)	3(0)
(500,500,500)	Proposed	$\boldsymbol{\Theta}$	0.754(0.035)	0.999(0.002)	0.034(0.004)		
		$\boldsymbol{\Gamma}$	0.327(0.023)	1.000(0.000)	0.001(0.000)	1.000(0.000)	3.00(0.00)
	HeteroGGM	$\boldsymbol{\Theta}$	4.049(0.105)	0.991(0.004)	0.879(0.006)		
		$\boldsymbol{\Gamma}$	-	-	-	0.660(0.087)	5.80(0.70)
	CGLasso($K = 6$)	$\boldsymbol{\Theta}$	3.090(0.165)	0.934(0.024)	0.110(0.030)		
		$\boldsymbol{\Gamma}$	3.389(0.355)	0.998(0.004)	0.056(0.020)	0.470(0.049)	6(0)
	CGLasso($K = 4$)	$\boldsymbol{\Theta}$	2.924(0.145)	0.963(0.021)	0.089(0.034)		
		$\boldsymbol{\Gamma}$	2.707(0.258)	0.999(0.002)	0.042(0.009)	0.677(0.025)	4(0)
	CGLasso($K = 3$)	$\boldsymbol{\Theta}$	2.618(0.166)	0.968(0.021)	0.046(0.018)		
		$\boldsymbol{\Gamma}$	1.813(0.067)	1.000(0.000)	0.024(0.001)	0.809(0.020)	3(0)
	MRCE($K = 6$)	$\boldsymbol{\Theta}$	2.671(0.128)	0.985(0.009)	0.266(0.026)		
		$\boldsymbol{\Gamma}$	3.235(0.353)	0.995(0.009)	0.131(0.012)	0.470(0.049)	6(0)
	MRCE($K = 4$)	$\boldsymbol{\Theta}$	2.481(0.161)	0.997(0.003)	0.221(0.027)		
		$\boldsymbol{\Gamma}$	2.519(0.287)	0.999(0.003)	0.096(0.008)	0.677(0.025)	4(0)
	MRCE($K = 3$)	$\boldsymbol{\Theta}$	2.220(0.113)	0.995(0.011)	0.163(0.012)		
		$\boldsymbol{\Gamma}$	1.621(0.092)	1.000(0.000)	0.058(0.003)	0.809(0.020)	3(0)
	MCGGM($K = 3$)	$\boldsymbol{\Theta}$	4.256(0.479)	0.797(0.231)	0.071(0.048)		
		$\boldsymbol{\Gamma}$	-	-	-	0.405(0.190)	3(0)

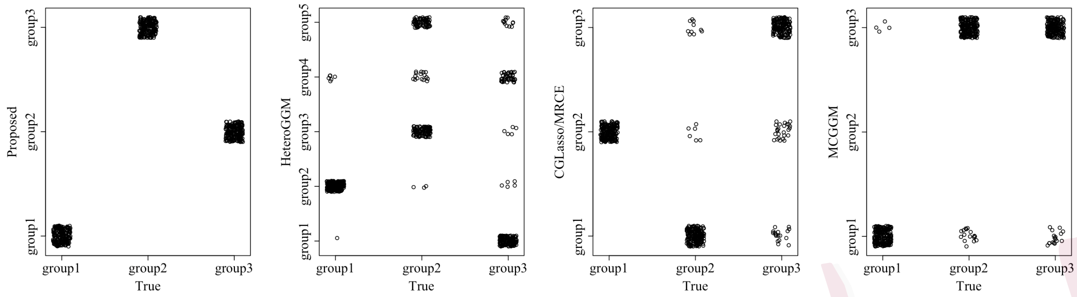

Figure 2: Analysis of one simulation replicate generated under S1 with group sizes (200, 200, 200). From left to right: Proposed method, HeteroGGM, nonparametric clustering for CGLasso and MRCE, and MCGGM.

number of sample groups, while HeteroGGM, without accounting for the regulations, over-estimates with a mean of 4.5. HeteroGGM has a satisfactory TPR value for the precision matrices, however, much inferior RMSE and FPR values. The other alternatives all have much inferior estimation performance with much larger RMSEs. They can have acceptable identification performance, especially when the number of sample groups is correctly specified – this can be very difficult in practice. In general, their identification accuracy is worse than the proposed. The alternatives fail to accurately identify the grouping structures. For this specific setting, the proposed approach has an average RI value of 1, HeteroGGM has an average RI of 0.547, and the other alternatives all have RI values below 0.5.

To further examine whether the proposed analysis can scale up, we consider the more challenging settings with $p = q = 200$ and $K_0 = 10$. The results are shown in

Table S8 and S9 (Supplementary Materials). It is again observed that the proposed approach outperforms the alternatives.

4. Data analysis

4.1 Analysis of the METABRIC data

Breast cancer has one of the highest incidence rates, and extensive profiling studies have been conducted on breast cancer. Gene expression data has been collected and analyzed in quite a few studies, among which some are multiomics (Tang et al., 2018; Lin et al., 2020). We analyze data collected in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study (Pereira et al., 2016) and refer to the published literature (Curtis et al., 2012) for information on sample and data collection and processing.

Gene expression and copy number alteration (CNA) measurements are available for 1,758 samples. In principle, the proposed analysis can be conducted with all gene expression and CNA measurements. Considering the limited sample size and large number of unknown parameters, we conduct a “candidate gene” analysis (Tabor et al., 2002) and focus on genes in the KEGG hsa05224 pathway. This pathway is named as “breast cancer” and contains well-known breast cancer related genes such as ESR, MYC, WTN, EGFR, KRAS, HRAS, NRAS, MAPK, and NOTCH. It has been examined in quite a few published studies (Dai et al., 2016), although it is noted that the perspectives taken in the published studies are significantly different from the proposed. A total of 147 genes belong to this pathway. Among them, two are not measured in the METABRIC study. As such, a total of 145 gene expressions and their corresponding CNAs are available for analysis. We refer to Curtis et al. (2012); Pereira et al. (2016) for the preprocessing of gene expression and CNA measurements.

When implementing the proposed approach, we set $K = 10$. A total of six sample groups are identified, with sizes 201, 387, 356, 303, 248, and 263. Detailed sample grouping information is available from the authors. For those six groups, the estimated gene expression networks are presented in Figure 3. The six networks have 684, 676, 432, 638, 380, and 652 edges, and Table S11 (Supplementary Materials) suggests that they have small to moderate numbers of overlapping edges. Genes with the highest degrees are presented in Figure S5 (Supplementary Materials), and significant differences are observed across the sample groups. For example, gene PIK3CD, which plays a critical role in some solid tumors including breast cancer, is an isolated node in the first network but a key hub node in the other networks, especially the sixth one. Other genes that behave differently in different sample groups include ESR1, DVL3, PGR, RPS6KB1, EGFR, FZD7 and FGFR1. There are also genes that behave similarly in all sample groups, such as CSNK1A1, E2F3 and MYC – they are established breast cancer markers and have high degrees in all the networks. The estimated coefficient matrices are presented in Figure 4, where we observe notable differences. In addition, it is observed that the cis regulations are usually the strongest, which is as expected. The regulation relationships are sparse, and there are a few trans regulations. In general, different genes have different sets of regulators. Some genes are co-regulated by certain regulators, for example, as observed in the case of SHC1 and FZD2.

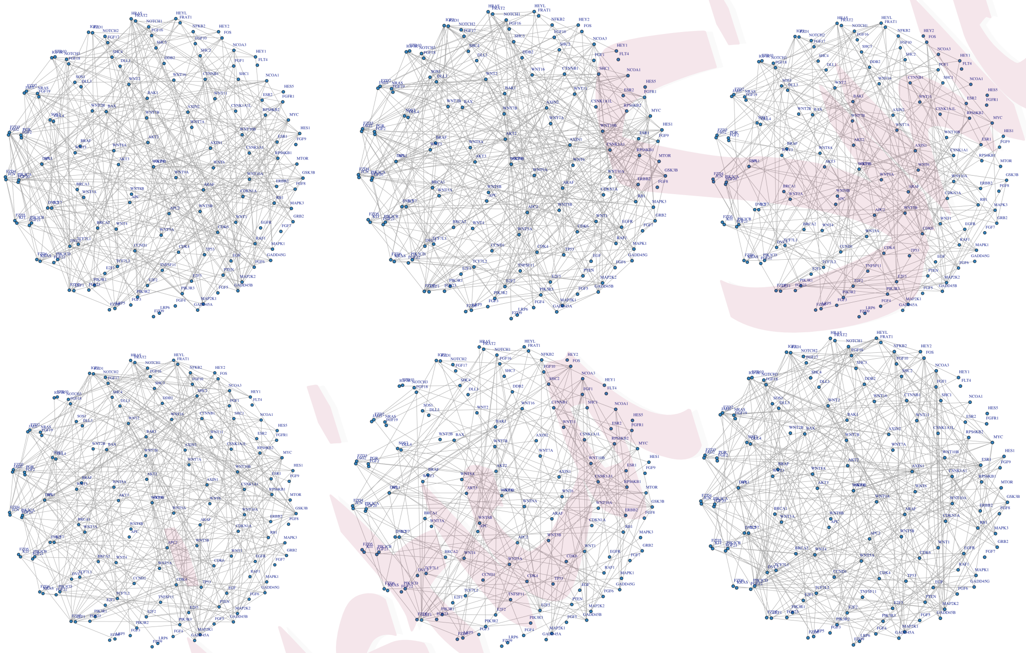

Figure 3: Analysis of METABRIC data: network structures for the six sample groups.

The proposed analysis is unsupervised. A review of relevant literature suggests that there is a lack of way for determining whether the identified sample groups and their differences (in gene expression network and regulation relationship) are clinically sensible. Here, to provide "indirect support", we compare key clinical features

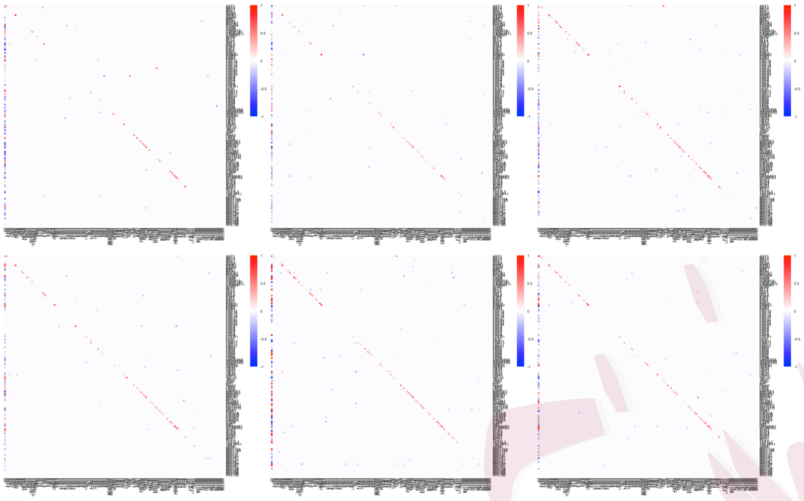

Figure 4: Analysis of METABRIC data: heatmaps of the estimated coefficient matrices for the six sample groups.

across the identified groups. In Table S12 (Supplementary Materials), we report the analysis of variance results for tumor size, mutation count, and tumor burden, all of which have significant clinical implications. In Figure S6 (Supplementary Materials), we further compare overall survival and relapse free survival. Significant differences are observed, suggesting that the six sample groups have notable clinical differences. Breast cancer can be classified as luminal A, luminal B, HER2-enriched, basal-like, and Claudin-low (Prat et al., 2015). In Table S13 (Supplementary Materials), we compare the identified six groups against these five subtypes. The Rand index between these two types of grouping is 0.736, suggesting certain consistency. For example, the basal-like ones are mostly in Group 5 identified by the proposed approach, and the HER2-enriched ones are mostly in Group 6. On the other hand, it is recognized that these two groupings also have notable differences. For example, the Claudin-low ones are almost equally presented in Group 2 and Group 5. Here it is noted that comparing with the Claudin subtypes, similar to the above analysis of clinical outcomes/phenotypes, is meant to provide additional insight into the clustering results. The Claudin subtyping is defined based on specific biomarkers, has a strategy/approach significantly different from the proposed, and cannot be used to evaluate clustering accuracy of the proposed approach.

Data is also analyzed using the alternative approaches. With HeteroGGM, the number of sample groups is data-dependently selected to be six. For the other alternatives, we fix the number of groups as five for better comparability. Here it is noted that MCGGM generates two empty groups, leading to three nontrivial ones. The heterogeneity analysis comparisons are summarized in Table S14 and Figure S7 (Supplementary Materials). It is observed that different approaches lead to significantly different groupings. Specifically, CGLasso and HeteroGGM have stronger overlappings with the proposed approach, while MCGGM generates highly imbalanced groups. The Rand index values between the five Claudin subtypes and the alternative approaches are 0.726 (CGLasso and MRCE), 0.730 (HeteroGGM), and 0.485 (MCGGM). The five networks generated by CGLasso have 590, 132, 96, 142 and 154 edges. Those generated by HeteroGGM have 594, 616, 662, 446, 752 and 3340 edges. And those generated by MCGGM have 238, 70 and 472 edges. MRCE fails in the first sample group due to its extremely small size and identifies 625, 393, 905, and 579 edges in the other sample groups. It is apparent that the network structures are also significantly different. More detailed results are available from the authors.

4.2 Analysis of the TCGA data

In Supplementary Materials, we analyze The Cancer Genome Atlas (TCGA) data on breast cancer. The sample size is 1,048, and we also analyze the 147 gene expressions and their corresponding CNAs in the KEGG hsa05224 (breast cancer) pathway. The proposed approach identifies three sample groups, which have significantly different gene expression network structures, regulation relationships, and clinical features. Additionally, it is found that its findings are significantly different from those of the alternatives. It is noted that the METABRIC and TCGA data have significant differences and cannot be pooled for analysis or directly compared.

5. Discussion

Using gene expression and regulator data, we have developed a new heterogeneity analysis approach that is based on high-dimensional conditional relationships as well as high-dimensional regulations. This analysis/approach includes multiple existing ones as special cases and can be more comprehensive/informative. Theoretical developments not only provide a solid foundation for the proposed approach but also may advance complex high-dimensional statistics – it is noted that there have been limited developments that collectively conduct conditional network and regression analysis, especially in the challenging context of heterogeneity analysis. We have convincingly demonstrated the practical effectiveness of the proposed approach. Multiple aspects of this study may demand additional research. For example, it can be of interest to accommodate other network constructions, other regulation models, and other types of data. As in the literature, we have stacked multiple types of regulators in a single vector. With proper data preprocessing, this has been shown as effective. It is noted that, as in some other analyses, when the dimension of gene expressions and regulators is extremly high, there is a risk of overfitting. For this study and beyond, it can be of interest to examine how to more effectively merge multiple types of regulators. A theoretical challenge, which has also been encountered by many published studies, is the asymptotic validity of the proposed tuning parameter selection. Additionally, as in many published studies, establishing the global maximization of the penalized likelihood without concavity is challenging. The proposed approach does not demand all relevant regulators. In practical data analysis, it can be of interest to examine the impact of missing regulators.

Supplementary Material

Online supplementary materials contain additional computational, theoretical, and numerical results.

Acknowledgments

This study is supported by the National Bureau of Statistics of China 2022LZ34, National Natural Science Foundation of China 11971404, NIH CA204120 and NSF 2209685. We thank the editor and reviewers for their careful review and insightful comments.

References

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics* (1), 77–120.

Cai, B., J. Zhang, and W. W. Sun (2021). Jointly modeling and clustering tensors in high dimensions. *arXiv preprint arXiv:2104.07773*.

Cai, T. T., H. Li, W. Liu, and J. Xie (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika 100*(1), 139–156.

Chauveau, D. and V. T. L. Hoang (2016). Nonparametric mixture models with conditionally independent multivariate component densities. *Computational Statistics & Data Analysis 103*, 1–16.

Church, B. V., H. T. Williams, and J. C. Mar (2019). Investigating skewness to understand gene expression heterogeneity in large patient cohorts. *BMC Bioinformatics 20*(24), 1–14.

Curtis, C., S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486*(7403), 346–352.

Dai, X., L. Xiang, T. Li, and Z. Bai (2016). Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *Journal of Cancer 7*(10), 1281–1294.

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

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*(8), 5467–5484.

Hao, B., W. W. Sun, Y. Liu, and G. Cheng (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research 18*, 1–58.

Huang, F., S. Chen, and S.-J. Huang (2018). Joint estimation of multiple conditional gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems 29*(7), 3034–3046.

Kagohara, L. T., G. L. Stein-O'Brien, D. Kelley, E. Flam, H. C. Wick, L. V. Danilova, H. Easwaran, A. V. Favorov, J. Qian, D. A. Gaykalova, et al. (2018). Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. *Briefings in Functional Genomics 17*(1), 49–63.

Lartigue, T., S. Durrleman, and S. Allassonnière (2021). Mixture of conditional gaussian graphical models for unlabelled heterogeneous populations in the presence of co-factors. *SN Computer Science 2*(6), 1–20.

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

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics 3*(9), e161.

Lin, Y., W. Zhang, H. Cao, G. Li, and W. Du (2020). Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes 11*(8), 888.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Pereira, B., S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut, et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications 7*(1), 11479.

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

Prat, A., E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and

M. Muñoz (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast 24*, S26–S35.

Radchenko, P. and G. Mukherjee (2017). Convex clustering via l_1 fusion penalization. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 79*(5), 1527–1546.

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

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

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

Sohn, K.-A. and S. Kim (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Artificial Intelligence and Statistics*, Volume 22, pp. 1081–1089. PMLR.

Sun, Y., Z. Luo, and X. Fan (2022). Robust structured heterogeneity analysis approach for high-dimensional data. *Statistics in Medicine 41*(17), 3229–3259.

Tabor, H. K., N. J. Risch, and R. M. Myers (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics 3*(5), 391–397.

Tang, J., D. Kong, Q. Cui, K. Wang, D. Zhang, Y. Gong, and G. Wu (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Frontiers in Oncology 8*, 374.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica 25*(3), 831–851.

Wytock, M. and Z. Kolter (2013). Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International Conference on Machine Learning*, pp. 1265–1273. PMLR.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*(4), 2630–2650.

Rong Li, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA.

E-mail: rong.li.rl946@yale.edu

Qingzhao Zhang, Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, Xiamen, China.

E-mail: zhangqingzhao@amss.ac.cn

Shuangge Ma, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA.

E-mail: shuangge.ma@yale.edu

Zhang and Ma are joint corresponding authors.