Received: 19 September 2020 | Revised: 26 November 2020 | Accepted: 5 January 2021

DOI: 10.1111/biom.13426

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Gaussian graphical model-based heterogeneity analysis via penalized fusion

Mingyang Ren[1,2,3] | **Sanguo Zhang**[1,2] | **Qingzhao Zhang**[4] | **Shuangge Ma**[3]

[1] School of Mathematics Sciences, University of Chinese Academy of Sciences, Beijing, China

[2] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing, China

[3] Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA

[4] MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, Xiamen, China

Correspondence
Qingzhao Zhang, MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics, and Fujian Key Lab of Statistics, Xiamen University, Xiamen, 361005, China;
Shuangge Ma, Department of Biostatistics, Yale School of Public Health, New Haven, CT 06511, USA.
Email: zhangqingzhao@amss.ac.cn; shuangge.ma@yale.edu

Funding information
Beijing Natural Science Foundation, Grant/Award Number: Z190004; Key Program of Joint Funds of the National Natural Science Foundation of China, Grant/Award Number: U19B2040; University of Chinese Academy of Sciences, Grant/Award Number: Y95401TXX2; National Natural Science Foundation of China, Grant/Award Numbers: 11971404, 71988101; 111 Project, Grant/Award Number: B13028; NSF, Grant/Award Number: 1916251; NIH, Grant/Award Numbers: CA241699, CA196530; and a Yale Cancer Center Pilot Award

Abstract

Heterogeneity is a hallmark of cancer, diabetes, cardiovascular diseases, and many other complex diseases. This study has been partly motivated by the unsupervised heterogeneity analysis for complex diseases based on molecular and imaging data, for which, network-based analysis, by accommodating the interconnections among variables, can be more informative than that limited to mean, variance, and other simple distributional properties. In the literature, there has been very limited research on network-based heterogeneity analysis, and a common limitation shared by the existing techniques is that the number of subgroups needs to be specified a priori or in an ad hoc manner. In this article, we develop a penalized fusion approach for heterogeneity analysis based on the Gaussian graphical model. It applies penalization to the mean and precision matrix parameters to generate regularized and interpretable estimates. More importantly, a fusion penalty is imposed to "automatedly" determine the number of subgroups and generate more concise, reliable, and interpretable estimation. Consistency properties are rigorously established, and an effective computational algorithm is developed. The heterogeneity analysis of non-small-cell lung cancer based on single-cell gene expression data of the Wnt pathway and that of lung adenocarcinoma based on histopathological imaging data not only demonstrate the practical applicability of the proposed approach but also lead to interesting new findings.

KEYWORDS
Gaussian graphical model, lung cancer, penalized fusion, unsupervised heterogeneity analysis

1 | INTRODUCTION

Heterogeneity is a hallmark of many complex diseases such as cancer, diabetes, cardiovascular diseases, and others, with patients having the same disease behaving differently. Most heterogeneity analysis can be classified as unsupervised and supervised, which serve different purposes, and no one can dominate the other. In this article, we focus on unsupervised heterogeneity analysis. Two motivating examples are considered in this article. The first is the heterogeneity analysis of regulatory T cells in non-small-cell lung cancer, which has been recognized to

have substantial influence on tumors (Ward and Kemp, 2017). However, our literature review suggests that it is still unknown how many subgroups of regulatory T cells are there and what immune function each subgroup has. Another example is also related to lung cancer but based on histopathological imaging data, which is more broadly available and cost-effective. However, as cancer modeling based on high-dimensional histopathological imaging data is much more recent, heterogeneity analysis is relatively limited, and both the number and structures of subgroups remain largely unknown. Extensive additional examples are available in the literature (Gao *et al.*, 2016). For such analysis, mean, variance, and other "simple" distributional properties have been widely used. Despite significant successes, it is recognized that the existing techniques are still insufficient, and there is a strong demand for more informative approaches.

The "natural next step" is to utilize information on the interconnections among variables along with means and variances. Analyses under other contexts, for example, regression, have shown that interconnections contain rich information beyond marginal means and variances, and that incorporating such information can significantly improve modeling. The simplest way of describing interconnections is perhaps via the correlation and precision matrices, which describe unconditional and conditional dependence, respectively. In this study, we conduct analysis under the Gaussian graphical model (GGM) framework, which is perhaps the most popular for the analysis of precision matrix and conditional dependence. The "classic" GGM has been designed for normally distributed variables, which is also assumed in our analysis. Following recent developments (Xue and Zou, 2012), both the GGM and proposed analysis can be extended to accommodate nonnormal data.

In the literature, there have been extensive methodological, theoretical, and computational studies on GGM under homogeneity. In these studies, the penalization technique has been popular (Friedman *et al.*, 2008). Under heterogeneity, when the subgrouping structure is known, various integrated analysis approaches have been developed. The central scheme is to take advantage of certain similarity shared by multiple subgroups so as to improve beyond separated estimation. Accordingly, several magnitude- and sign-based shrinkage techniques have been developed. For example, Guo *et al.* (2011) develop a hierarchical penalty that targets removing common zeros in precision matrices shared by subgroups. Danaher *et al.* (2014) adopt a single fusion penalty, which is applied to pairwise differences (of elements in precision matrices) between subgroups. A more challenging scenario, as can be partly seen from our data analysis, is where the subgrouping structure is unknown. In principle, such GGM analysis can be conducted using finite mixture modeling techniques. Two representative works combine the Gaussian graphical mixture model with the truncated fusion penalty (which is applied to pairwise differences between subgroups Gao *et al.*, 2016) and the group penalty (which is applied to elements of the precision matrices across subgroups; Hao *et al.*, 2018). In the existing finite mixture modelings, including those for simpler settings such as regression, an important and challenging task is to determine the number of subgroups. Some studies assume that it is known a priori, although it is usually difficult to justify such a choice. Some other studies adopt ad hoc approaches without rigorous justifications. As such, there is still a strong demand for more effective methods.

There are more remotely related studies that have examined the number of subgroups issue. For example, a family of convex clustering adopts the adaptive fused penalty on clustering center parameters and data dependently determines the number of clusters based on center differences and achieves a unique global solution (Wang *et al.*, 2018). However, such studies do not analyze precision matrices and conditional dependence. Another family of methods adopts Bayesian techniques and can infer the number of subgroups. A representative example is the Dirichlet process mixture models (DPM) (Wang and Dunson, 2011). DPMs critically rely on two parameters: the concentration parameter and base distribution. However, the implications of particular choices are difficult to assess (Görür and Rasmussen, 2010).

Motivated by the importance of unsupervised heterogeneity analysis for complex diseases and insufficiency of the existing approaches, our overarching goal is to take the interconnections among variables into consideration and conduct more informative heterogeneity analysis under the GGM framework. This study is built on but significantly advances from the existing GGM studies by accommodating heterogeneity. It shares some penalization and shrinkage strategies with the existing heterogeneity studies that assume known subgrouping structures, but significantly advances by assuming that the subgrouping structure is unknown and needs to be determined fully data dependently. The assumed data and model structures share some similar spirit with the existing finite mixture modelings. However, this study significantly advances from the literature by not assuming a known number of subgroups and determining it using a rigorous approach. Compared to those limited to numerical analysis, this study advances by also conducting rigorous theoretical investigation. Last but not least, two heterogeneity analyses on lung cancer are conducted, built on single-cell gene expression data and histopathological imaging data, respectively. This not only demonstrates the applicability of the proposed approach but also provides a

biologically sensible alternative of dissecting lung cancer heterogeneity. Overall, with methodological, theoretical, and numerical advancements, this study is warranted beyond the existing literature.

2 | METHODS

Denote n as the number of independent subjects. For subject $i(=1,\ldots,n)$, p-dimensional measurement $\boldsymbol{x}_i$ is available. Further assume that the n subjects belong to K_0 subgroups, where the value of K_0 is unknown. For the lth subgroup, assume the Gaussian distribution:

$$f_l\left(\boldsymbol{x};\boldsymbol{\mu}_l^*,\boldsymbol{\Sigma}_l^*\right)=(2\pi)^{-p/2}\left|\boldsymbol{\Sigma}_l^*\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_l^*\right)^\top(\boldsymbol{\Sigma}_l^*)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_l^*\right)\right\},$$

where the mean and covariance matrix are unknown. Overall, $\boldsymbol{x}_i$s satisfy distribution:

$$f(\boldsymbol{x})=\sum_{l=1}^{K_0}\pi_l^* f_l\left(\boldsymbol{x};\boldsymbol{\mu}_l^*,\boldsymbol{\Sigma}_l^*\right),$$

where the mixture probabilities π_l^*s are also unknown.

The data setting here is a "combination" of that under the homogeneous GGM and that under finite mixture modeling. Different from Guo *et al.* (2011) and Danaher *et al.* (2014), the subgrouping memberships are unknown. Different from Gao *et al.* (2016) and Hao *et al.* (2018), here K_0 is for notational purposes, and its value is not assumed to be known. As in "standard" mixture modeling, the percentages of subgroups are not assumed to be known. We note again that the normality assumption can be dropped. As our goal is to conduct more effective heterogeneity analysis, this normality assumption will not be further discussed.

2.1 | Penalized estimation

As in "standard" GGM studies, we are interested in estimating the means and precision matrices. In addition, with heterogeneity, we are also interested in K_0 and the mixture probabilities. With their estimates, the posterior probability of a subject belonging to a specific subgroup can be calculated using the Bayesian rule.

We propose the penalized objective function:

$$\mathcal{L}(\boldsymbol{\Omega},\boldsymbol{\pi}|\boldsymbol{X}):=\frac{1}{n}\sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k f_k\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k^{-1}\right)\right)-\mathcal{P}(\boldsymbol{\Omega}),\tag{1}$$

where $\boldsymbol{X}$ denotes the collection of observed data, $\boldsymbol{\Omega}=(\boldsymbol{\Omega}_1^\top,\ldots,\boldsymbol{\Omega}_K^\top)^\top$, $\boldsymbol{\Omega}_k=\mathrm{vec}(\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k)=(\mu_{k1},\ldots,\mu_{kp},\theta_{k11},\ldots,\theta_{kp1},\ldots,\theta_{k1p},\ldots,\theta_{kpp})\in\mathbb{R}^{p^2+p}$, $\boldsymbol{\Theta}_k=\boldsymbol{\Sigma}_k^{-1}$ is the kth precision matrix with the ijth entry θ_{kij}, $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)^\top$,

$$\mathcal{P}(\boldsymbol{\Omega})=\sum_{k=1}^{K}\sum_{j=1}^{p}p(|\mu_{kj}|,\lambda_1)+\sum_{k=1}^{K}\sum_{i\neq j}p(\left|\theta_{kij}\right|,\lambda_2)+\sum_{k<k'}p\left(\left(\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k'}\|_2^2+\|\boldsymbol{\Theta}_k-\boldsymbol{\Theta}_{k'}\|_F^2\right)^{1/2},\lambda_3\right),\tag{2}$$

$\|\cdot\|_F$ is the Frobenius norm and $p(\cdot,\lambda)$ is a concave penalty function with tuning parameter $\lambda>0$. In our numerical study, we adopt minimax concave penalty (MCP) (Zhang, 2010) and note that smoothly clipped absolute deviation (SCAD) and some other penalties are also applicable. K is a known constant that satisfies $K>K_0$. Consider:

$$(\widehat{\boldsymbol{\Omega}},\widehat{\boldsymbol{\pi}})=\underset{\boldsymbol{\Omega},\boldsymbol{\pi}}{\mathrm{argmax}}\,\mathcal{L}(\boldsymbol{\Omega},\boldsymbol{\pi}|\boldsymbol{X}).$$

Denote $\{\widehat{\boldsymbol{\Upsilon}}_1,\ldots,\widehat{\boldsymbol{\Upsilon}}_{\widehat{K}_0}\}$ as the distinct values of $\widehat{\boldsymbol{\Omega}}$, that is, $\{k:\widehat{\boldsymbol{\Omega}}_k\equiv\widehat{\boldsymbol{\Upsilon}}_l,k=1,\ldots,K\}_{l=1,\ldots,\widehat{K}_0}$ constitutes a partition of $\{1,\ldots,K\}$. Then there are $\widehat{K}_0$ subgroups with estimated mean and precision parameters in $\widehat{\boldsymbol{\Omega}}$. The mixture probabilities can be extracted from $\widehat{\boldsymbol{\pi}}$.

2.1.1 | Rationale

We consider high-dimensional settings, which include low-dimensional settings as special cases and can be more challenging and practically more useful. To data dependently identify the number of subgroups, we effectively "search" between 1 and K using the penalized fusion technique, which differs significantly from, for example, comparing the Bayesian information criterion (BIC) and other criteria under different subgroup numbers (whose statistical properties are often very challenging to establish). In practice, although it is hard to know K_0 a priori, its "upper bound" can be easy to specify, either based on some biological knowledge or the maximum number of subgroups desirable. To be cautious, as shown in our numerical analysis, K can be set as a relatively large number. In $\mathcal{P}(\boldsymbol{\Omega})$, the first two are sparsity penalties. In GGM analysis, centralization is often conducted. For a mean component, if the same value is shared by all subgroups, then centralization leads to mean zero, demanding sparsity. As such, sparsity penalization on means has been routinely applied in the literature (Hao *et al.*, 2018; Gao *et al.*, 2016). On the other hand, it is noted that, unlike under homogeneity, it is

impossible to achieve equal means using centralization. As such, it is necessary to include mean parameters in analysis. The second penalty is imposed on parameters of the precision matrices, which is routine. Both sparsity penalties can also regularize estimation. The most significant advancement is the third, fusion penalty. With this penalty, we shrink differences among the K subgroups and encourage equality (a smaller number of subgroups). As the subgroups are defined by both means and variances of the normal distributions, the fusion penalty is applied to both mean and precision matrix parameters, which intuitively can be more effective than simply being applied to means or precision matrices. Although there have been extensive developments of the fusion technique (Danaher *et al.*, 2014; Gao *et al.*, 2016), the proposed penalty is significantly different by simultaneously applying to both means and precision matrices, which is necessary for the heterogeneous GGM. Similar to the existing penalized fusion studies, the subgrouping structures can be inferred by examining the distinct estimates. And similar to the existing GGM studies, the conditional dependence relationships for each subgroup can be obtained by examining the nonzero estimates of the precision matrices. An "additional" product is the estimated means for the subgroups. It is noted that the number of precision matrix parameters is $O(p^2)$, which may dominate the number of mean parameters $O(p)$. In our theoretical development described in the next section, we establish sufficient conditions that lead to consistency. Our numerical study also shows satisfactory numerical results. When it is desirable to measure "average effects," it is possible to normalize the precision matrix penalty term. With minor changes in the assumed conditions, we also expect consistency. As similar problems have also been encountered in published literature (e.g., group penalization), we will not discuss further.

2.2 | Statistical properties

Denote the true values of parameters as $\boldsymbol{\Upsilon}^* = (\boldsymbol{\Upsilon}_1^{*\top}, \ldots, \boldsymbol{\Upsilon}_{K_0}^{*\top})^\top$ and $\boldsymbol{\Upsilon}_l^* = \mathrm{vec}(\boldsymbol{\mu}_l^*, \boldsymbol{\Theta}_l^*)$ for $l = 1, \ldots, K_0$. Let $b = \min_{1 \leqslant l_1 \neq l_2 \leqslant K_0} \|\boldsymbol{\Upsilon}_{l_1}^* - \boldsymbol{\Upsilon}_{l_2}^*\|_2$ be the minimal difference of the common parameters between two groups. Define $\mathcal{S}_l = \{(i,j) : \theta_{lij}^* \neq 0, 1 \leqslant i \neq j \leqslant p\}$, $\mathcal{D}_l = \{j : \mu_{lj}^* \neq 0, 1 \leqslant j \leqslant p\}$. Let $|\mathcal{S}|$ be the cardinality of the set $\mathcal{S}$, and denote the sparsity parameters $d = \max\{|\mathcal{D}_l|, l = 1, \ldots, K_0\}$, $s = \max\{|\mathcal{S}_l|, l = 1, \ldots, K_0\}$. In the Supporting Information, we describe the assumed conditions and their interpretations. With these conditions, we can establish the following consistency results.

Theorem 1. *Suppose that Conditions B.1–B.5 (in the Supporting Information) hold. Assume that* $K^2K_0^4(s+p)\log p/n = o(1)$, $K_0^2\sqrt{\frac{(s+p)\log p}{n}} = O(\lambda_1) = O(\lambda_2)$, *the minimal signal in* $\boldsymbol{\Upsilon}^*$ *is larger than* $a \cdot \max\{\lambda_1, \lambda_2\}$, *where* a *is the regularization parameter of the concave penalty defined in Condition B.4,* $\lambda_3 \gg KK_0^2\sqrt{\frac{(s+p)\log p}{n}}$, *and* $b > a\lambda_3$. *Then we have*

1. (Consistency of $\widehat{K}_0$*)* $P(\widehat{K}_0 = K_0) \to 1$.

2. (Rate of convergence) There exists a local maximum of (1) that satisfies:

$$\sum_{l=1}^{\widehat{K}_0} \left(\|\widehat{\boldsymbol{\mu}}_l - \boldsymbol{\mu}_l^*\|_2 + \|\widehat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_l^*\|_F\right) = O_p\left(K_0^2\left(\sqrt{d\log p/n} + \sqrt{(s+p)\log p/n}\right)\right).$$

3. (Sparsistency) Denote the set of the nonzero elements of $\widehat{\boldsymbol{\mu}}_l$ *as* $\widehat{\mathcal{D}}_l = \{j : \widehat{\mu}_{lj} \neq 0\}$, *and the set of the nonzero off-diagonal elements of* $\widehat{\boldsymbol{\Theta}}_l$ *as* $\widehat{\mathcal{S}}_l = \{(i,j) : i \neq j, \widehat{\theta}_{lij} \neq 0\}$. *Then with probability tending to 1,* $\widehat{\mathcal{S}}_l = \mathcal{S}_l$, *and* $\widehat{\mathcal{D}}_l = \mathcal{D}_l$ *for any* $l = 1, \ldots, \widehat{K}_0$.

It is noted that, as in other heterogeneity analyses, there is a "reshuffling" of estimates issue, which will not be further discussed. This theorem suggests that the proposed approach has multiple much-desired consistency properties. In particular, it can accurately estimate the number of subgroups with a high probability. Such a property has not been established for many of the existing mixture modelings. In addition, the estimates are consistent under high-dimensional settings. In particular, the rate of convergence $O(\sqrt{(s+p)\log p/n})$ (if K and K_0 are bounded) is optimal for estimating sparse precision matrices under the Frobenius norm (Cai *et al.*, 2016). With respect to sparsistency, the advantage of the concave penalty is that the precision matrices can be less sparse compared to under the L_1 penalty (under which the order of s is not larger than $O(p)$). The proof of Theorem 1 and more discussions on the sparsistency are provided in the Supporting Information.

Similar to in the existing high-dimensional GGM studies (Hao *et al.*, 2018), only the local convergence properties are established. We recognize that there are some recent works on global convergence. However, they are often limited to very stringent conditions and/or low-dimensional cases. For example, Daskalakis *et al.* (2017) characterize the global convergence behaviors of an Expectation Maximization (EM) algorithm for fitting two equally weighted Gaussian subgroups. Zhao *et al.* (2020) assume that the mixture weights are known for GGMs with an arbitrary number of subgroups. Overall, global convergence for more general high-dimensional GGMs demands significant future research.

2.3 | Computation

We develop an effective EM+Alternating Direction Method of Multipliers (ADMM) algorithm for optimizing objective function (1). Some developments may be specific to MCP, and optimizing with other penalties may demand minor modifications. Denote $\boldsymbol{\gamma} = (\gamma_{ik})_{n\times K}$, where γ_{ik} is the latent indicator variable showing the membership of the ith observation in the mixture. If γ_{ik}s were available, then the penalized log-likelihood function (1) for the complete data could be written as

$$\mathcal{L}(\boldsymbol{\Omega}|\boldsymbol{X},\boldsymbol{\gamma}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik} \left\{\log\pi_k + \log f_k\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k^{-1}\right)\right\} - \mathcal{P}(\boldsymbol{\Omega}). \quad (3)$$

In the tth expectation step, the conditional expectation of (3) is

$$E_{\boldsymbol{\gamma}|\boldsymbol{X},\boldsymbol{\Omega}^{(t-1)}}[\mathcal{L}(\boldsymbol{\Omega}|\boldsymbol{X},\boldsymbol{\gamma})] = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}^{(t)} \left\{\log\pi_k + \log f_k\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Theta}_k^{-1}\right)\right\} - \mathcal{P}(\boldsymbol{\Omega}), \quad (4)$$

where $\gamma_{ik}^{(t)}$ can be computed as

$$\gamma_{ik}^{(t)} = \frac{\pi_k^{(t-1)} f_k\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k^{(t-1)},\left(\boldsymbol{\Theta}_k^{(t-1)}\right)^{-1}\right)}{\sum_{k=1}^{K}\pi_k^{(t-1)} f_k\left(\boldsymbol{x}_i;\boldsymbol{\mu}_k^{(t-1)},\left(\boldsymbol{\Theta}_k^{(t-1)}\right)^{-1}\right)}. \quad (5)$$

In the tth maximization step, maximizing (4) with respect to π_k leads to the estimate:

$$\pi_k^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{ik}^{(t)}. \quad (6)$$

For $\boldsymbol{\mu}_k$, maximizing (4) with respect to $\{\boldsymbol{\mu}\} = \boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K$ is equivalent to solving

$$\{\boldsymbol{\mu}^{(t)}\} = \underset{\{\boldsymbol{\mu}\}}{\operatorname{argmin}}\left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}^{(t)} \left\{(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Theta}_k^{(t-1)}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\} + \mathcal{P}(\boldsymbol{\Omega})\right). \quad (7)$$

For this problem, the local quadratic approximation can be adopted, which can lead to an explicit solution at each iteration. Details are provided in the Supporting Information.

An update for $\{\boldsymbol{\Theta}\} = \boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_K$ can be developed by noting the connection with the joint graphical lasso estimation (Danaher *et al.*, 2014). More specifically, maximizing (4) with respect to $\{\boldsymbol{\Theta}\}$ is equivalent to solving

$$\{\boldsymbol{\Theta}_k^{(t)}, k=1,\ldots,K\} = \underset{\{\boldsymbol{\Theta}\}}{\operatorname{argmax}}\left(\sum_{k=1}^{K} n_k \left[\log\{\det(\boldsymbol{\Theta}_k)\} - \operatorname{tr}\left(\widetilde{\boldsymbol{S}}_k\boldsymbol{\Theta}_k\right)\right] - \mathcal{P}(\{\boldsymbol{\Theta}\})\right), \quad (8)$$

where $n_k = \sum_{i=1}^{n}\gamma_{ik}^{(t)}$, $\widetilde{\boldsymbol{S}}_k = \frac{\sum_{i=1}^{n}\gamma_{ik}^{(t)}(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(t)})(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{(t)})^\top}{\sum_{i=1}^{n}\gamma_{ik}^{(t)}}$ is the pseudosample covariance matrix, and $\mathcal{P}(\{\boldsymbol{\Theta}\}) = \sum_{k=1}^{K}\sum_{i\neq j}p(|\theta_{kij}|,\lambda_2) + \sum_{k<k'}p((\|\boldsymbol{\mu}_k^{(t)}-\boldsymbol{\mu}_{k'}^{(t)}\|_2^2 + \|\boldsymbol{\Theta}_k-\boldsymbol{\Theta}_{k'}\|_F^2)^{1/2},\lambda_3)$.

The solution to (8) can be effectively obtained using the ADMM technique. Compared to that in Danaher *et al.* (2014) and some other studies, our optimization faces more challenges with the fusion penalty. To tackle this problem, we resort to the effective sparse alternating minimization algorithm (Wang *et al.*, 2018). More details are provided in the Supporting Information.

The overall algorithm is achieved by iterating (5), (6), (7), and (8). It is summarized in Algorithm S1 of the Supporting Information. The initialization is implemented using the K-means method, and we note that some other sensible clustering methods can also generate satisfactory initial values. To improve accuracy, refitting can be conducted afterwards. Convergence properties of this algorithm can be established following Hao *et al.* (2018), and convergence is achieved in all of our numerical studies.

2.3.1 | Tuning parameter selection

The proposed approach involves three tuning parameters $(\lambda_1,\lambda_2,\lambda_3)$. Recent studies have shown that optimization with three tunings is computationally manageable. This is especially true with the simple updates as described above. For tuning parameter selection, we conduct a grid search and optimize the adaptive BIC-type criterion: $-2\sum_{i=1}^{n}\log\{\sum_{l=1}^{\widehat{K}_0}\widehat{\pi}_l f_l(\boldsymbol{x}_i;\widehat{\boldsymbol{\mu}}_l,(\widehat{\boldsymbol{\Theta}}_l)^{-1})\} + \sum_{l=1}^{\widehat{K}_0}(\log n \cdot s_{1l} + 2s_{2l})$, where $s_{1l} = |\{j : \widehat{\mu}_{lj}\neq 0, 1\leqslant j\leqslant p\}|$, $s_{2l} = |\{(i,j) : \widehat{\theta}_{lij}\neq 0, 1\leqslant i<j\leqslant p\}|$, for $l = 1,\ldots,\widehat{K}_0$. The degrees of freedom of the precision matrices are associated with a smaller weight, following the recommendation of Danaher *et al.* (2014).

3 | SIMULATION

We consider a three-class ($K_0 = 3$) problem. The number of variables is set as $p = 100$. With respect to

sample sizes, we consider both balanced (under which all subgroups have sample size 200) and imbalanced (under which the subgroups have sample sizes 150, 200, and 250) scenarios. The observations are generated as follows. First, the subgroup membership y_is are randomly sampled from $\{1, 2, 3\}$. Then, $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}(y_i), \boldsymbol{\Sigma}(y_i))$. The first eight components of the subgroup means are $(\mu\mathbf{1}_4^\top, -\mu\mathbf{1}_4^\top)^\top I(y_i = 1) + \mu\mathbf{1}_8 I(y_i = 2) + (-\mu)\mathbf{1}_8 I(y_i = 3)$, where $\mathbf{1}_4$ is a length-four vector of all ones. The rest $p - 8$ components of the means are set as zero. For μ, we consider 1.5 and 2.0.

We simulate three types of network structures. Under the first type, all subgroups have tridiagonal precision matrices. Here the diagonal elements are all equal to one, and the nonzero off-diagonal elements are $0.2I(y_i = 1) + 0.3I(y_i = 2) + 0.4I(y_i = 3)$. Under the second and third types, each network consists of 10 equally sized disjoint subnetworks (modules). Among the 10 subnetworks, eight are shared by the three subgroups. In addition, subgroups 1 and 2 share one subnetwork. The same is true for subgroups 1 and 3. Subgroups 2 and 3 also have a unique subnetwork of their own. For the structures of the subnetworks, we consider the following two popular choices. (1) *Power-law network*: For the precision matrix of the first subgroup, 10 power-law subnetworks are all generated with two edges added in each step. The initial 10-block precision matrix $(\theta_{1ij})_{p\times p}$ is generated by

$$\theta_{1ij} = \begin{cases} 1 & i = j; \\ 0 & i \neq j, i \nsim j; \\ \text{Unif}([-0.4, -0.1] \cup [0.1, 0.4]) & i \neq j, i \sim j; \end{cases}$$

where $i \sim j$ means that there is an edge between nodes i and j and $i \nsim j$ means otherwise. To ensure positive definiteness, we set $\theta_{1jj} = \sum_{i\neq j} |\theta_{1ij}| + 0.1$. Subgroup-specific subnetworks are generated similarly and independently. (2) *Nearest-neighbor network:* Consider the precision matrix of the first subgroup. For each subnetwork, we generate $p/10$ points randomly on a unit square, calculate all $p/10 \times (p/10 - 1)/2$ pairwise distances, and find the m nearest neighbors of each point. The nearest-neighbor network is obtained by linking any two points that are among the m-nearest neighbors of each other. The integer m controls the degree of sparsity, and we set $m = 3$ in our simulation.

With the proposed approach, we set $K = 6$. We have experimented with larger values and observed similar results. The following relevant competitors are considered: (a) K-means clustering. The number of subgroups is selected using the gap statistic. This method is implemented using R package *NbClust*. (b) A two-step approach. The method (a) is used to obtain the subgroups, and then the joint graphical lasso (Danaher *et al.*, 2014) is used to estimate the precision matrices. This approach is implemented using R package *JGL*. We also consider a variation of this approach, under which the top principle components, as opposed to the original measurements, are used for clustering analysis. (c) The simultaneous clustering and estimation of heterogeneous graphical models (SCAN) method (Hao *et al.*, 2018) with the correct number of subgroups. This approach applies the group lasso penalty to the same elements of all precision matrices and encourages common structures. We note that setting the number of subgroups as the true leads to favorable performance and is not realistic in data analysis. (d) The SCAN method with wrong numbers of subgroups ($K = 2, 4$, and 6). We note that although there have been extensive works on GGM, heterogeneity analysis, and mixture modeling *separately*, heterogeneity analysis under the GGM framework is quite limited. The SCAN method is perhaps the most relevant.

To gauge performance of the proposed and alternative approaches, we consider the following measures: (a) mean and sd of $\widehat{K}_0$, (b) percentage of $\widehat{K}_0$ equal to K_0 (denoted as "per"), (c) subgrouping error defined by

$$\text{CE}(\widehat{\varphi}, \varphi) := \binom{n}{2}^{-1} |\{(i, j) : I\big(\widehat{\varphi}(\boldsymbol{x}_i) = \widehat{\varphi}\big(\boldsymbol{x}_j\big)\big) \neq I\big(\varphi(\boldsymbol{x}_i) = \varphi\big(\boldsymbol{x}_j\big)\big); i < j\}|,$$

where $\widehat{\varphi}$ and φ are the estimated and true subgrouping memberships, respectively, (d) mean squared error (MSE) for $\{\boldsymbol{\mu}\}$ and $\{\boldsymbol{\Theta}\}$. When $\widehat{K}_0 = K_0$, the MSEs are defined as $\sum_{k=1}^{K_0} \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^*\|_2/K_0$ and $\sum_{k=1}^{K_0} \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k^*\|_2/K_0$. When $\widehat{K}_0 \neq K_0$, they are defined as

$$\text{MSE}(\boldsymbol{\mu}) = \frac{1}{\widehat{K}_0} \sum_{k=1}^{\widehat{K}_0} \sum_{l'=1}^{K_0} \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_{l'}^*\|_2 \cdot I$$
$$(l' = \underset{l}{\text{argmin}}\{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_l^*\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_l^*\|_F^2\}),$$

$$\text{MSE}(\boldsymbol{\Theta}) = \frac{1}{\widehat{K}_0} \sum_{k=1}^{\widehat{K}_0} \sum_{l'=1}^{K_0} \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_{l'}^*\|_2 \cdot I$$
$$(l' = \underset{l}{\text{argmin}}\{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_l^*\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_l^*\|_F^2\});$$

and (e) true and false positive rates (TPR and FPR) for the off-diagonal elements of the precision matrices. When $\widehat{K}_0 = K_0$, they are defined as

$$\text{TPR} = \frac{1}{K_0} \sum_{k=1}^{K_0} \frac{\sum_{i<j} I(\theta_{kij} \neq 0, \widehat{\theta}_{kij} \neq 0)}{\sum_{i<j} I\big(\theta_{kij} \neq 0\big)},$$

$$\text{FPR} = \frac{1}{K_0} \sum_{k=1}^{K_0} \frac{\sum_{i<j} I(\theta_{kij} = 0, \widehat{\theta}_{kij} \neq 0)}{\sum_{i<j} I\big(\theta_{kij} = 0\big)}.$$

TABLE 1 Simulation results under the balanced design and $\mu = 1.5$: mean (sd)

Network	Method	per	$\widehat{K}_0$	CE	MSE(μ)	MSE(Θ)	TPR	FPR
Tridiagonal	K-means	0	2(0)	0.297(0.102)	1.603(0.024)	-	-	-
	K-means+JGL	0	2(0)	0.297(0.102)	1.603(0.024)	1.522(0.041)	0.999(0.002)	0.122(0.015)
	SCAN($K = 2$)	0	2(0)	0.246(0.166)	1.548(0.009)	1.524(0.030)	0.998(0.003)	0.113(0.019)
	SCAN($K = 3$)	1	3(0)	0.001(0.001)	0.452(0.025)	1.403(0.042)	0.940(0.021)	0.092(0.028)
	SCAN($K = 4$)	0	4(0)	0.100(0.019)	0.688(0.044)	1.392(0.017)	0.952(0.010)	0.132(0.003)
	SCAN($K = 6$)	0	6(0)	0.189(0.076)	0.828(0.057)	3.394(0.076)	0.869(0.032)	0.183(0.036)
	FGGM	0.98	2.98(0.14)	0.018(0.125)	0.365(0.022)	1.301(0.070)	0.967(0.013)	0.115(0.017)
Nearest-neighbor	K-means	0	2(0)	0.260(0.131)	1.401(0.048)	-	-	-
	K-means+JGL	0	2(0)	0.260(0.131)	1.401(0.048)	2.494(0.063)	0.879(0.018)	0.141(0.022)
	SCAN($K = 2$)	0	2(0)	0.240(0.107)	1.954(0.147)	2.529(0.095)	0.867(0.022)	0.125(0.027)
	SCAN($K = 3$)	1	3(0)	0.029(0.012)	0.529(0.069)	2.061(0.092)	0.851(0.023)	0.130(0.029)
	SCAN($K = 4$)	0	4(0)	0.111(0.029)	0.946(0.055)	2.000(0.047)	0.802(0.026)	0.127(0.019)
	SCAN($K = 6$)	0	6(0)	0.246(0.023)	0.702(0.039)	1.772(0.014)	0.737(0.018)	0.123(0.003)
	FGGM	0.96	3.04(0.35)	0.043(0.105)	0.185(0.033)	1.766(0.083)	0.870(0.024)	0.139(0.022)
Power-law	K-means	0	2(0)	0.277(0.134)	1.494(0.063)	-	-	-
	K-means+JGL	0	2(0)	0.277(0.134)	1.494(0.063)	2.565(0.090)	0.892(0.018)	0.166(0.024)
	SCAN($K = 2$)	0	2(0)	0.244(0.145)	1.781(0.142)	2.575(0.124)	0.887(0.018)	0.152(0.031)
	SCAN($K = 3$)	1	3(0)	0.018(0.009)	0.587(0.046)	2.092(0.033)	0.872(0.012)	0.156(0.009)
	SCAN($K = 4$)	0	4(0)	0.105(0.030)	0.946(0.050)	2.042(0.020)	0.808(0.015)	0.147(0.003)
	SCAN($K = 6$)	0	6(0)	0.243(0.029)	0.623(0.057)	1.935(0.016)	0.656(0.014)	0.087(0.003)
	FGGM	0.96	3.02(0.20)	0.028(0.081)	0.181(0.021)	1.812(0.095)	0.870(0.018)	0.136(0.021)

When $\widehat{K}_0 \neq K_0$, their definitions are revised similarly as for the MSEs.

For analyzing one replicate, the proposed approach takes about 15 min on a laptop with regular configurations. The first two alternatives have low computational cost, taking about 1 and 5 min, respectively. The computational cost of SCAN highly depends on K. For example, with $K = 6$, the analysis of one replicate takes about 50 min, and with $K = 3$, it takes less 10 min. With 100 replicates, summary results for the balanced and imbalanced designs with $\mu = 1.5$ are shown in Tables 1 and 2, respectively. Results for $\mu = 2.0$ are shown in Tables S1 and S2 of the Supporting Information. The variation of K-means+JGL, using top principle components that contain 80%, 90%, and 95% of information, does not improve over K-means+JGL (Tables S3 of the Supporting Information) and hence is not additionally considered. Observations made under different settings are similar. The two K-means-based approaches cannot properly identify the number of subgroups. SCAN has satisfactory performance when the number of subgroups is correctly specified. Here it is noted that in published studies, there is a lack of rigorous proposal for determining subgroup number. Compared to SCAN with $K = 3$, the proposed approach, with fully data-driven $\widehat{K}$, has advantageous estimation and similar identification results. Consider, for example, the nearest-neighbor structure in Table 1. SCAN and the proposed approach have CE 0.029 and 0.043, MSE($\boldsymbol{\mu}$) 0.529 and 0.185, MSE($\boldsymbol{\Theta}$) 2.061 and 1.766, TPR 0.851 and 0.870, and FPR 0.130 and 0.139, respectively. We conjecture that the estimation advantage of the proposed approach may be partly attributable to the advantage of MCP. SCAN has highly unsatisfactory performance when K is mis-specified.

4 | DATA ANALYSIS

4.1 | Heterogeneity analysis of regulatory T cells in non-small-cell lung cancer

There are two major classes of T cells: CD4+ and CD8+, which function in different and complementary ways in cell-mediated immune reactions. CD4+ T cells can be further divided into T helper cells (CD4+ $CD25^{low/int}$) and regulatory T cells (CD4+ $CD25^{high}$), according to the cell surface markers. Tumor-infiltrating regulatory T cells (Tregs) have been associated with poor survival in many cancers including melanoma and lung cancer. In recent studies, it has been suggested that the removal of Tregs can evoke and enhance antitumor immune response (Tanaka and Sakaguchi, 2017), and the heterogeneity of Tregs may have a direct impact on tumors (Ward and Kemp, 2017).

TABLE 2 Simulation results under the imbalanced design and $\mu = 1.5$: mean (sd)

Network	Method	per	$\widehat{K}_0$	CE	MSE(μ)	MSE(Θ)	TPR	FPR
Tridiagonal	K-means	0	2(0)	0.314(0.051)	1.353(0.028)	-	-	-
	K-means+JGL	0	2(0)	0.314(0.051)	1.353(0.028)	1.788(0.054)	0.999(0.002)	0.112(0.015)
	SCAN($K = 2$)	0	2(0)	0.304(0.024)	1.534(0.005)	1.464(0.032)	0.999(0.002)	0.107(0.014)
	SCAN($K = 3$)	1	3(0)	0.001(0.002)	0.495(0.027)	1.440(0.026)	0.920(0.020)	0.082(0.018)
	SCAN($K = 4$)	0	4(0)	0.118(0.018)	0.630(0.035)	1.370(0.058)	0.935(0.025)	0.115(0.023)
	SCAN($K = 6$)	0	6(0)	0.155(0.056)	0.907(0.065)	3.405(0.075)	0.897(0.037)	0.183(0.076)
	FGGM	0.97	3.05(0.44)	0.029(0.159)	0.421(0.035)	1.469(0.035)	0.947(0.038)	0.126(0.036)
Nearest-neighbor	K-means	0	2(0)	0.315(0.054)	1.124(0.061)	-	-	-
	K-means+JGL	0	2(0)	0.315(0.054)	1.124(0.061)	2.472(0.064)	0.881(0.016)	0.138(0.021)
	SCAN($K = 2$)	0	2(0)	0.305(0.062)	1.926(0.156)	2.459(0.099)	0.877(0.021)	0.132(0.032)
	SCAN($K = 3$)	1	3(0)	0.022(0.010)	0.612(0.052)	2.104(0.100)	0.843(0.027)	0.126(0.032)
	SCAN($K = 4$)	0	4(0)	0.120(0.032)	0.850(0.045)	1.957(0.039)	0.811(0.014)	0.133(0.020)
	SCAN($K = 6$)	0	6(0)	0.240(0.026)	0.751(0.058)	1.792(0.017)	0.732(0.010)	0.125(0.002)
	FGGM	0.92	3.08(0.39)	0.057(0.141)	0.216(0.051)	1.843(0.082)	0.865(0.025)	0.145(0.020)
Power-law	K-means	0	2(0)	0.315(0.045)	1.219(0.083)	-	-	-
	K-means+JGL	0	2(0)	0.315(0.045)	1.219(0.083)	2.526(0.080)	0.894(0.016)	0.162(0.021)
	SCAN($K = 2$)	0	2(0)	0.304(0.059)	1.753(0.168)	2.518(0.116)	0.891(0.018)	0.152(0.028)
	SCAN($K = 3$)	1	3(0)	0.018(0.008)	0.593(0.037)	2.109(0.069)	0.870(0.015)	0.159(0.019)
	SCAN($K = 4$)	0	4(0)	0.129(0.029)	0.796(0.041)	2.006(0.017)	0.821(0.014)	0.142(0.003)
	SCAN($K = 6$)	0	6(0)	0.237(0.039)	0.658(0.056)	1.903(0.062)	0.682(0.047)	0.104(0.024)
	FGGM	0.94	3.04(0.24)	0.042(0.124)	0.191(0.042)	1.860(0.071)	0.867(0.020)	0.139(0.024)

With the fast development of single-cell sequencing technologies, there have been extensive studies on the heterogeneity of T cells. Despite significant progresses, it is still recognized as an open problem (Zielinski, 2017; Guo *et al.*, 2018). It is noted that in the existing analyses, relatively simpler data distributional properties have been examined. Advancing from these studies, we conduct network-based heterogeneity analysis to further account for the interconnections among gene expressions.

Specifically, we focus on the heterogeneity analysis of tumor-infiltrating Tregs in non-small-cell lung cancer (Guo *et al.*, 2018). In recent studies (Kared *et al.*, 2020), it has been suggested that the Wnt signaling pathway is associated with the heterogeneity of CD4+ T cells. As such, we focus on genes within this pathway. With a limited sample size, focusing on the most promising pathway can generate more reliable results than conducting genome-wide analysis. The analyzed single-cell gene expression data are downloaded from Gene Expression Omnibus (GEO, 2018). After removing genes with missingness and with standard deviations less than 0.25, analysis is conducted on 902 samples and 73 genes. We recognize that raw single-cell gene expression data may not have multivariate normal distributions. However, with proper processing, normal distributions can provide a sensible working approximation.

Published studies suggest a small to moderate number of subgroups. We set $K = 10$ in our analysis, which should be large enough. Three subgroups are identified, with sizes 409, 316, and 177, respectively. In Figure 1 (this figure appears in color in the electronic version of this article), we show their network structures. More detailed estimation results, including on the subgroup means, are available from the authors. It is easy to observe that the three network structures are significantly different—the pairwise comparisons using the test of Li and Chen (2012) give p-values <0.001. They have 51, 87, and 276 edges, respectively. In addition, properties of the individual nodes are significantly different. For example, consider genes VANGL1 and EP300, which have been established as playing important roles in lung cancer. They are isolated nodes in the first network, have a small number of edges in the second network and are key hub nodes in the third network.

To gain more insights into the three subgroups, in Figure 1, we show the normalized expressions of several key immune-related genes, including CTLA4, FOXP3, TIGIT, and others. These genes have been reported to promote the differentiation of T cells into Tregs with an immunosuppressive function (Lucca and Dominguez, 2020). Brief discussions on their biological functionalities, especially related to lung cancer, are provided in the Supporting Information. It is noted that these genes are not in the

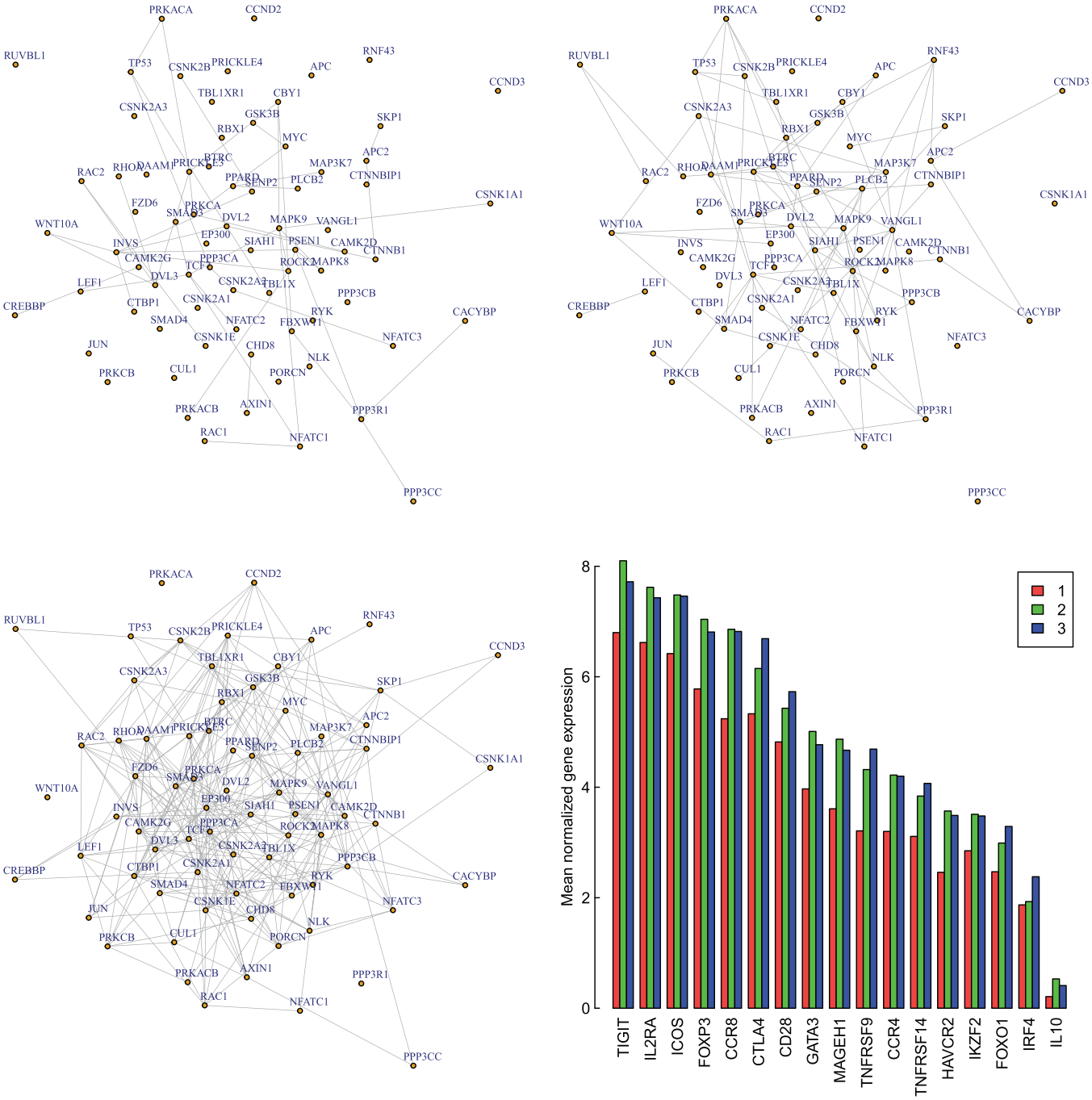

FIGURE 1 Analysis of regulatory T cells: identified subgrouping network structures and mean normalized expressions of 17 immune-related genes in the subgroups. This figure appears in color in the electronic version of this article

Wnt signaling pathway—as such, there is not an "overfitting" problem. It is easy to see that the expressions of these genes differ significantly across the three subgroups (p-values from ANOVA <0.05). This can support the biological sensibility of our analysis to a certain extent.

Data are also analyzed using the alternatives considered in simulation by setting $K = 3$ (for a higher level of comparability). Summary results and comparisons with those described above are presented in the Supporting Information. Significant differences are observed. There are 120, 75, and 90 edges, respectively, for the three subgroups identified by the K-means+JGL method, among which 8, 2, and 14 edges overlap with those in the three networks identified by the proposed approach (Figure S1). With the SCAN method, the three identified subgroups have 80, 109, and 111 edges, respectively, among which 5, 5, and 25 overlap with those identified using the proposed approach (Figure S2). The sample subgrouping results using different methods are significantly different (Table S4).

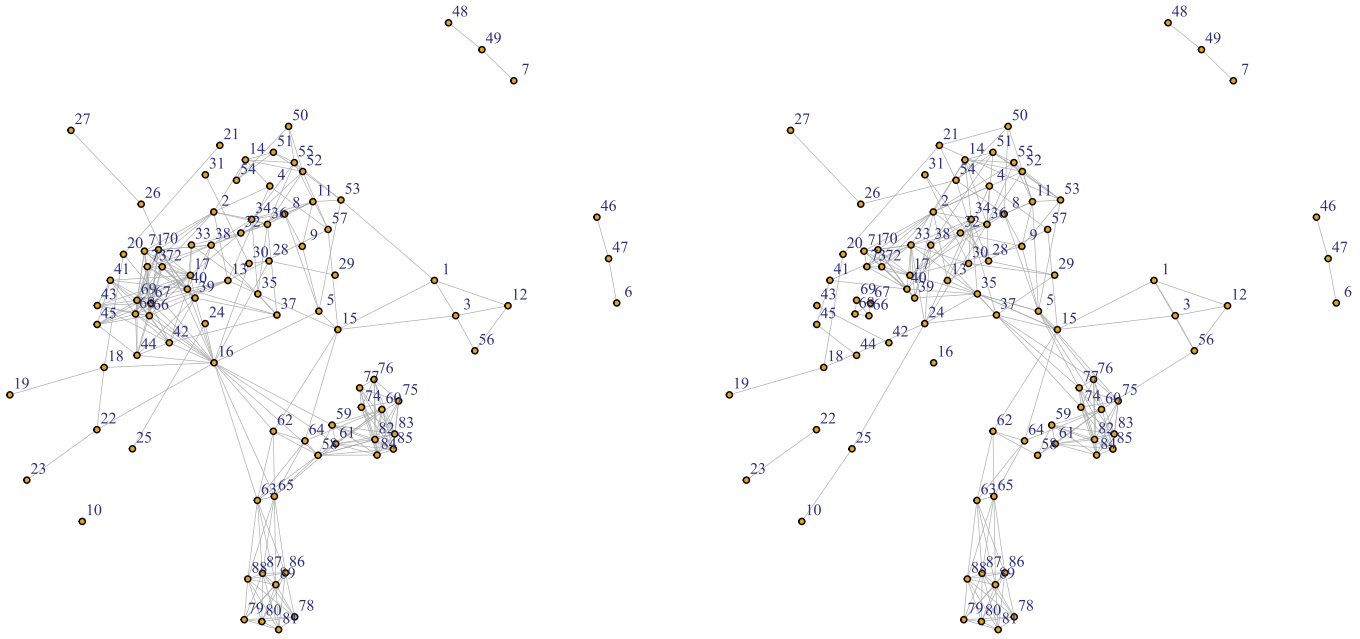

FIGURE 2 Analysis of lung adenocarcinoma (LUAD) histopathological imaging data: identified subgrouping network structures

4.2 | LUAD heterogeneity analysis using histopathological imaging data

Histopathological images, which are generated in the biopsy procedure, have been long used for diagnosis and staging. In the past, heterogeneity analysis, specific to the identification of subtypes and other subcategories, was mostly realized by pathologists via looking at the biopsy slides under microscope. In the past few years, histopathological imaging features extracted by digital imaging processing software have been established as fast, cost-effective, and effective. Here we analyze the imaging features extracted using *cellprofiler* from the histopathological slides of the TCGA LUAD patients, with data downloaded from the TCGA portal (TCGA, 2021). The feature-processing pipeline has been described in Zhang *et al.* (2020). We further remove features with standard deviations less than 1 to focus on the "more interesting" features. The working dataset contains 307 LUAD subjects and 89 imaging features.

With $K = 10$, the proposed approach identifies two sample subgroups, with sizes 246 and 61, respectively. The two network structures are shown in Figure 2. Summary results using the two alternatives (Table S5, Figure S3, and Figure S4) and detailed information on the imaging features (Table S6) are provided in the Supporting Information. The two networks share some common characteristics, for example, the connections among nodes (6, 46, 47), (7, 48, 49), (78, 81), and (86, 89). However, they overall differ significantly (p-value <0.001). They contain 301 and 250 edges, respectively. Other node properties also differ significantly between the networks. For example, node 16 is an important hub in the first network but an isolated node in the second network. As another example, the module consisting of nodes 66 and 69 is isolated in the second network but has many linked edges in the first network. Different from molecular data, the biological implications of high-dimensional imaging features extracted using automated imaging processing software are not as clear. As such, we do not further examine biological interpretations.

In Table 3, we compare clinical features between the two subgroups. It is noted that they are not used in subgrouping, and hence there is no overfitting. The two subgroups differ significantly in tumor status (which is the state or condition of an individual's neoplasm at a particular point of time). The differences in Per(prebroncholiator) and overall survival are borderline significant. The other three variables have p-values<0.1. We note that our heterogeneity analysis is unsupervised and purely based on the imaging features, and such heterogeneity may or may not correspond to certain clinical differences. The differences observed in Table 3 suggest that our analysis actually has clinical implications.

5 | DISCUSSION

In this article, we have conducted unsupervised sample heterogeneity analysis based on the GGM. Methodological advancements include using distributional properties

TABLE 3 Analysis of LUAD data: summary of clinical features for the two subgroups

Clinical variable (sample size)		Subgroup 1	Subgroup 2	*p* value
[a] Tumor status (258)	Tumor free	170	25	
	With tumor	45	18	
	Incidence	0.209	0.419	0.006
[a] Kras indicator (170)	Negative	106	33	
	Positive	28	3	
	Positive rate	0.209	0.083	0.094
[a] ICD 10 (307)	C34.1	149	31	
	C34.2	14	1	
	C34.3	73	23	
	Others	10	6	0.095
[b] Overall survival status (307)	Deceased	76	30	
	Living	170	31	0.067
[c] Per(prebroncholiator) (126)		78.74(20.91)	92.09(27.89)	0.052
[c] Per(postbroncholiator) (67)		77.11(25.57)	95.83(27.37)	0.094

aCategorical variables. *p* values from Fisher's exact test. ICD 10 C34.1: malignant neoplasm of bronchus or lung (upper lobe), C34.2: middle lobe, and C34.3: lower lobe.

bOverall survival status: *p* value from the log-rank test.

cContinuous variables. Mean (sd) and *p* values from ANOVA. Per(pre-/postbroncholiator): The percentage comparison to a normal value reference range of the volume of air that a patient can forcibly exhale from the lungs in 1 s pre-/postbronchodilator.

more complicated/informative than mean and variance for heterogeneity analysis, fully data dependently determining the number of subgroups, and estimating means along with precision matrices. As can be partly seen from the proof (Supporting Information), such methodological innovations have brought significant challenges, and our theoretical developments can also shed light into other high-dimensional heterogeneity and network analyses. Simulation has demonstrated satisfactory computational properties and accurate subgrouping and identification. The two data examples have shown that the proposed analysis can provide a new view of disease heterogeneity, and the validity of analysis has some indirect support.

This study can be potentially extended in multiple directions. The penalized fusion technique for heterogeneity analysis can be coupled with other distributional properties to data dependently determine the number and structure of subgroups. As briefly mentioned above, the proposed approach can be applied to nonnormal data by replacing the simple correlation with the rank correlation and other robust measures. In some of the existing joint estimations of multiple networks, promoting certain similarity (in magnitude or sparsity structure) has been considered. It may be of interest to further extend the proposed approach to accommodate such similarity. It is also of interest to further explore the implications of the two lung cancer heterogeneity analysis results.

ACKNOWLEDGEMENTS

We thank the editor and reviewers for careful review and insightful comments. This work was supported by the Beijing Natural Science Foundation (Z190004), Key Program of Joint Funds of the National Natural Science Foundation of China (U19B2040), University of Chinese Academy of Sciences (Y95401TXX2), National Natural Science Foundation of China (11971404, 71988101), 111 Project (B13028), NSF (1916251), NIH (CA241699, CA196530), and a Yale Cancer Center Pilot Award.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are openly available in GEO (Gene Expression Omnibus) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99254 and in TCGA (The Cancer Genome Atlas) at https://portal.gdc.cancer.gov/projects/TCGA-LUAD.

ORCID

Shuangge Ma https://orcid.org/0000-0001-9001-4999

REFERENCES

Cai, T., Liu, W. and Zhou, H. (2016) Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44, 455–488.

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

Daskalakis, C., Tzamos, C. and Zampetakis, M. (2017) Ten steps of EM suffice for mixtures of two Gaussians. *Proceedings of the 2017 Conference on Learning Theory*, PMLR 65, 704–710, 2017.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.

Gao, C., Zhu, Y., Shen, X. and Pan, W. (2016) Estimation of multiple networks in Gaussian mixture models. *Electronic Journal of Statistics*, 10, 1133–1154.

GEO (Gene Expression Omnibus) T cell landscape of non-small cell lung cancer revealed by deep single-cell RNA sequencing. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99254 [Accessed January 8, 2021].

Görür, D. and Rasmussen, C.E. (2010) Dirichlet process Gaussian mixture models: choice of the base distribution. *Journal of Computer Science and Technology*, 25, 653–664.

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

Guo, X., Zhang, Y., Zheng, L., Zheng, C. and Zhang, Z. (2018) Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nature Medicine*, 24, 978–985.

Hao, B., Sun, W., Liu, Y. and Cheng, G. (2018) Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18, 7981–8038.

Kared, H., Tan, S.W., Lau, M.C., Chevrier, M., Tan, C., How, W., et al. (2020) Immunological history governs human stem cell memory CD4 heterogeneity via the Wnt signaling pathway. *Nature Communications*, 11, 1–17.

Li, J. and Chen, S. (2012) Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40, 908–940.

Lucca, L.E. and Dominguez-Villar, M. (2020) Modulation of regulatory T cell function and stability by co-inhibitory receptors. *Nature Reviews Immunology*, 20(11), 680–693.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009) Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104, 735–746.

Tanaka, A. and Sakaguchi, S. (2017) Regulatory T cells in cancer immunotherapy. *Cell Research*, 27, 109–118.

TCGA. (The Cancer Genome Atlas). Available at https://portal.gdc.cancer.gov/projects/TCGA-LUAD. [Accessed January 8, 2021].

Wang, B., Zhang, Y., Sun, W.W. and Fang, Y. (2018) Sparse convex clustering. *Journal of Computational and Graphical Statistics*, 27, 393–403.

Wang, L. and Dunson, D. (2011) Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20, 196–216.

Ward, H.K.A. and Kemp, R.A. (2017) Regulatory T-cell heterogeneity and the cancer immune response. *Clinical & Translational Immunology*, 6, e154.

Xue, L. and Zou, H. (2012) Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40, 2541–2571.

Zhang, C. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, S., Fan, Y., Zhong, T. and Ma, S. (2020) Histopathological imaging features-versus molecular measurements-based cancer prognosis modeling. *Scientific Reports*, 10, 15030.

Zhao, R., Li, Y. and Sun, Y. (2020) Statistical convergence of the EM algorithm on Gaussian mixture models. *Electronic Journal of Statistics*, 14, 632–660.

Zielinski, C.E. (2017) Human T cell immune surveillance: phenotypic, functional and migratory heterogeneity for tailored immune responses. *Immunology Letters*, 190, 125–129.

SUPPORTING INFORMATION

Details of the computational algorithms (referenced in Section 2.3), conditions and proofs (referenced in Sections 2.2 and 5) , and additional numerical results (referenced in Sections 3 and 4) are available at the Biometrics website on Wiley Online Library. R programs implementing the proposed method are available at www.github.com/shuanggema.

How to cite this article: Ren M, Zhang S, Zhang Q, Ma S. Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*. 2021;1–12. https://doi.org/10.1111/biom.13426