Statistica Sinica Preprint No: SS-2024-0258	
Title	Grouped Heterogeneous Gaussian Graphical Models for High-Dimensional Clustered Data
Manuscript ID	SS-2024-0258
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0258
Complete List of Authors	Xin Zeng, Shuangge Ma and Qingzhao Zhang
Corresponding Authors	Qingzhao Zhang
E-mails	zhangqingzhao@amss.ac.cn
Notice: Accepted author version.	

Grouped Heterogeneous Gaussian Graphical Models for High-Dimensional Clustered Data

Xin Zeng[1], Shuangge Ma[2] and Qingzhao Zhang*[1]

[1]Xiamen University and [2]Yale University

Abstract: Clustered data-based analysis has been extensively conducted in various studies. Recent research has demonstrated that a network-based heterogeneity analysis, which adopts a system perspective and incorporates the interconnections among variables while considering heterogeneity between components, can provide more informative results compared to approaches based on simpler statistics. Moreover, incorporating grouping strategies in analysis can better delineate the sources of heterogeneity and enable more flexible modeling for clustered data. In this article, we introduce a novel approach called the grouped heterogeneous Gaussian graphical models (Grouped-HGGM) for network analysis of high-dimensional clustered data. Our approach assumes that clusters can be divided into distinct groups, and any heterogeneity across clusters is captured through the cluster-wise mixture probabilities. Unlike most previous approaches that assume that the number of components is known in advance, an appealing feature of our method is the automatic determination of the number of components and sparse estimation using a fusion technique. Consistency properties are

*Corresponding author.

rigorously established, and an effective computational algorithm is developed. Extensive simulations demonstrate the practical superiority of the proposed approach over closely related alternatives. In the analysis of breast cancer data, the proposed approach identifies heterogeneity structures different from the alternatives.

Key words and phrases: Clustered data, Gaussian graphical models, Grouping strategies, Heterogeneity analysis.

1. Introduction

Consider data that can be divided into m known clusters, which can be based on postal regions, institutions, species, and others. This type of data is referred to as clustered data, which frequently appears in various areas of research, such as neuroscience (Galbraith, Daniel and Vissel (2010)), personalized medicine (Fokkema et al. (2018)), and labor economics (Pereda-Fernandez (2021)). For a given cluster i ($i = 1, \ldots, m$), heterogeneity often manifests in real data, indicating the presence of subgroup structures. As a popular analytical strategy, the finite mixture model (McLachlan and Peel (2000)) could be useful to flexibly capture heterogeneity. In the left panel of Figure 1, we show a dataset consisting of m clusters, and each cluster is

treated as a finite Gaussian mixture comprising K_0 subgroups/components:

$$f_i(\boldsymbol{x}) = \sum_{l=1}^{K_0} \pi_{il}^* f_l(\boldsymbol{x}; \boldsymbol{\mu}_l^*, \boldsymbol{\Sigma}_l^*), \quad i = 1, \ldots, m, \tag{1.1}$$

where $f_l(\boldsymbol{x}; \boldsymbol{\mu}_l^*, \boldsymbol{\Sigma}_l^*)$ is a Gaussian distribution with mean vector $\boldsymbol{\mu}_l^*$ and covariance matrix $\boldsymbol{\Sigma}_l^*$. For m clusters from the model (1.1), one of the primary objectives in the statistical analysis would be accounting for heterogeneity across clusters, prompting us to consider the scenario that the cluster-wise distributions share common latent distributions f_l. The heterogeneity across clusters is captured through the cluster-wise mixture probabilities $(\pi_{i1}^*, \ldots, \pi_{iK_0}^*)$ in the model (1.1). Specifically, we assume that all the clusters are divided into a finite number of groups by a grouping strategy, and the clusters within the same group share the same mixture probability. For the sake of clarity of hierarchy, let us provide a concise example. Cancer data stemming from diverse regions or laboratories (clusters) fall within the purview of clustered data. Given the existence of cancer subtypes (components), the investigation necessitates conducting component-specific network analysis. Moreover, the groups are derived through the mentioned grouping strategy. The idea of assuming shared latent distributions across clusters is not new and can be traced back to earlier literature (Teh et al. (2006), Rodriguez, Dunson and Gelfand (2008), Sugasawa (2021)). Our study advances by further considering a useful and popular setting. Specif-

ically, we take into account the interrelationships among variables in clustered data, which is both promising and challenging. Under the Gaussian Graphical Model (GGM) framework, these interrelationships can be recovered from the inverse covariance matrix $(\mathbf{\Sigma}_l^*)^{-1}$. As depicted in the middle panel of Figure 1, learning the component-specific network structures is also one of the objectives of our study.

The GGM framework stands out as particularly attractive for analyzing conditional dependencies, primarily due to its clear interpretations, favorable statistical properties, and computational advantages. The problem of estimating multiple networks in the presence of sample heterogeneity under GGM has been extensively studied. Two main scenarios have been explored: one is where it is unknown which samples belong to which distributions (Gao et al. (2016), Ren et al. (2022)), and the other is where this information is known (Guo et al. (2011), Danaher, Wang and Witten (2014)). In this study, we consider the former scenario as it offers more flexibility. While heterogeneous GGMs have been widely employed for heterogeneity analysis across various disciplines, this article can be unique in conducting unsupervised heterogeneity analysis for high-dimensional clustered data. The GGM assumes that observations follow a multivariate normal distribution, wherein the conditional independence of two nodes

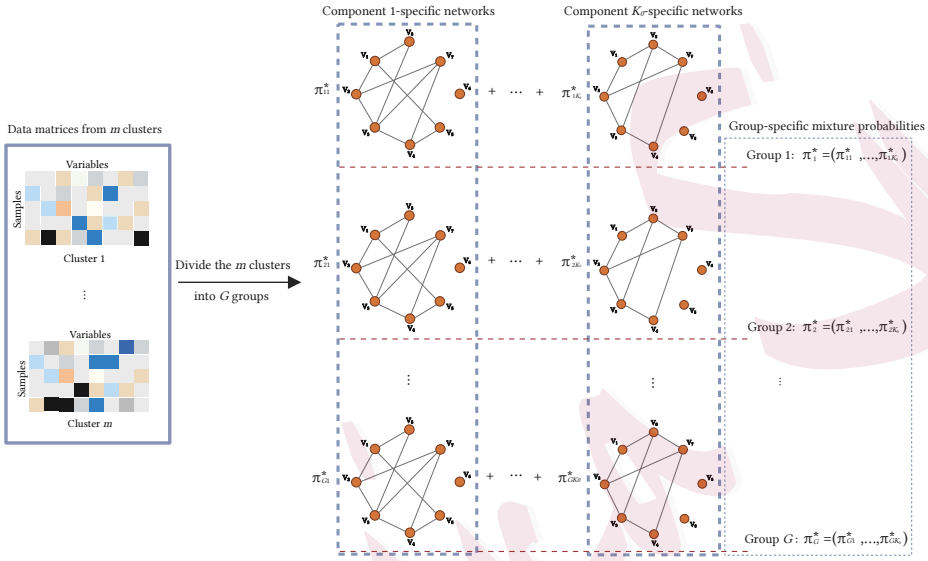

Figure 1: Heterogeneity network analysis of clustered data using Grouped-HGGM, including data structures (left), component-specific networks analysis by GGM (middle), and group assignment of m clusters into G groups (right).

(variables) is equivalent to the corresponding element in the precision matrix having a zero value. Consequently, determining the network structure involves a sparse estimation of the precision matrix, for which various techniques, including penalization, have been employed. Notable contributions include joint graphical Lasso (Danaher, Wang and Witten (2014)) and its subsequent refinements (Gao et al. (2016)). Additionally, a more encompassing approach known as SCAN (Hao et al. (2018)) enables sparsity in both cluster means and precision matrices. However, the aforementioned studies typically necessitate pre-specifying K_0, which is usually difficult to justify such a choice. Ren et al. (2022) proposed a fusion penalty to identify number of components. Although related research has also conducted high-dimensional unsupervised heterogeneity analysis, our work, which involves a key difference in the incorporation of a broader data framework, aims to investigate not only component heterogeneity but also heterogeneity among clusters, this is theoretically and computationally challenging. The aforementioned studies can be viewed as special cases with $m = 1$. Currently, the available literature addressing these specific issues remains very limited. Consequently, there is a strong demand for more effective methods in this domain.

Our contributions in this article are twofold. First, on the method-

ological front, we propose a novel approach called the grouped heterogeneous Gaussian graphical model (Grouped-HGGM) for network analysis of high-dimensional clustered data. Beyond following the GGM framework to establish the network (graph) of variables, the significant advances include incorporating heterogeneity and grouping strategies. Under the framework of heterogeneity analysis, the proposed analysis identifies the number of components as well as their respective network structures. Furthermore, it seeks to group datasets from multiple clusters to explore the heterogeneity among these clusters using cluster-wise mixture probabilities. It belongs to the family of network-based heterogeneity analysis and may exhibit similar merits as those demonstrated in Gao et al. (2016), Hao et al. (2018), and Ren et al. (2022). On the theoretical front, we provide a non-asymptotic statistical analysis of the output generated directly from the ECM algorithm. When simultaneously identifying the number of components and accommodating the heterogeneity among clusters through the determination of the number of groups, our analysis focus on additional technical challenges in both theory and computation. The theoretical and computational developments in this study, although sharing some similarities with the existing studies, are notably more intricate and demand additional investigations. With advancements in methodology, theory, and numerical analysis, this

study offers a significant contribution beyond the existing literature.

The remainder of the article is organized as follows. Section 2 presents a novel method for analyzing heterogeneity in clustered data with a network structure, and an ECM algorithm is developed to implement it. The theoretical properties of the proposed estimators are established in Section 3. In Section 4, a comparative study is conducted to assess the performance of the proposed method against alternatives. Section 5 illustrates the proposed method using a real data example. Further discussions are provided in Section 6. Additional numerical results, detailed algorithms, and proofs of the theorems are provided in the Supplementary Material.

2. Methods

Let $\boldsymbol{x}_{ij}$ denote the j-th p-dimensional measurement in the i-th cluster for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where m is the number of clusters and n_i is the number of within-cluster samples or cluster size which can be unequal across clusters. Assume further that the the n_i subjects belong to K_0 components defined based on cancer subtypes, where K_0 and component memberships are unknown. This is a reasonable assumption as the data within each cluster is of the same type. For instance, the number K_0 of cancer subtypes remains consistent across all clusters. In model (1.1), the cluster-wise share

common latent distributions, that is, the mean $\boldsymbol{\mu}_l^*$ and covariance matrix $\boldsymbol{\Sigma}_l^*$ of l-th component are same for all m clusters. The parameters $\boldsymbol{\mu}_l^*$, $\boldsymbol{\Sigma}_l^*$ and cluster-wise mixture probabilities $\boldsymbol{\pi}_i^* = (\pi_{i1}^*, \ldots, \pi_{iK_0}^*)$ are unknown. To model the heterogeneity across clusters, we divide the m clusters into G groups, i.e.,

$$f_i(\boldsymbol{x}_{ij}) = \sum_{l=1}^{K_0} \pi_{g_i l}^* f_l(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_l^*, \boldsymbol{\Sigma}_l^*), \tag{2.2}$$

where $g_i \in \{1, \ldots, G\}$ and $\boldsymbol{\pi}_{g_i}^* = (\pi_{g_i 1}^*, \ldots, \pi_{g_i K_0}^*)$ are unknown grouping parameters and group-specific mixture probabilities, respectively. Thus the clusters within the same group share the same mixture probabilities, which leads to the same mixture distributions in a given group (the right panel of Figure 1). It is worth noting that, an unknown number of components K_0 complicates the problem, a finite mixture model with too few numbers of components may fail to adequately capture heterogeneity, leading to biased estimates. Conversely, an excessive number of components can result in data overfitting and diminish interpretability. Many existing studies (Danaher, Wang and Witten (2014), Gao et al. (2016), Hao et al. (2018), Sugasawa (2021) and Li et al. (2022)) require a prior specification of the value of K_0 in model (2.2) or select it using information criteria. It is naturally unreasonable to specify K_0 directly. Recent literature (Pei et al. (2022), Chen et al. (2023)) on mixture modeling highlights the difficulty of

consistently selecting the K_0 using information criteria. In our study, we do not require a prior specification of the value of K_0 in the model (2.2) or select it using information criteria.

2.1 Penalized estimation

While it is challenging to determine K_0 a priori, specifying its "upper bound" $K > K_0$ can be relatively straightforward. To be cautious, K can be taken as a relatively large number and bounded. This can be based on some contexts about the data or simply by selecting a relatively appropriate number. With this K, we consider the following grouped heterogeneous Gaussian graphical models:

$$f_i(\boldsymbol{x}_{ij}) = \sum_{k=1}^{K} \pi_{g_i k} f_k(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k^{-1}), \quad i = 1, \ldots, m, j = 1, \ldots, n_i, \tag{2.3}$$

where $\sum_{k=1}^{K} \pi_{g_i k} = 1$, $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kp})^\top$ is the mean vector, and $\boldsymbol{\Theta}_k = \boldsymbol{\Sigma}_k^{-1}$ is the precision matrix for the k-th component with the (i, j)-th entry θ_{kij}. Define $\boldsymbol{g}$ and $\boldsymbol{\pi}$ as the collections of g_i and π_{gk}, respectively, which are also unknown. For parameter estimation and determination of the heterogeneity structure, we propose the penalized objective function:

$$\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X}) = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \left(\sum_{k=1}^{K} \pi_{g_i k} f_k(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k^{-1}) \right) - P(\boldsymbol{\Omega}), \tag{2.4}$$

where $N = \sum_{i=1}^{m} n_i$ and $\boldsymbol{X}$ stand for the total sample size and the collection of observed data, respectively. $\boldsymbol{\Omega}_k = \text{vec}(\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) = (\mu_{k1}, \ldots, \mu_{kp}, \theta_{k11}, \ldots, \theta_{kp1}, \ldots, \theta_{k1p}, \ldots, \theta_{kpp})^\top \in \mathbb{R}^{p+p^2}$, and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)^\top \in \mathbb{R}^{K(p+p^2)}$. Here, the penalty is proposed as

$$P(\boldsymbol{\Omega}) = \sum_{k=1}^{K} \sum_{r=1}^{p} p_{\lambda_1}(|\mu_{kr}|) + \sum_{k=1}^{K} \sum_{i \neq j} p_{\lambda_2}(|\theta_{kij}|) + \sum_{k<k'} p_{\lambda_3}(\tau_{k,k'}), \qquad (2.5)$$

where $\tau_{k,k'} = (||\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}||_2^2 + ||\boldsymbol{\Theta}_k - \boldsymbol{\Theta}_{k'}||_F^2)^{\frac{1}{2}}$, $||\cdot||_F$ is the Frobenius norm, and $p_\lambda(\cdot)$ is a concave penalty function with tuning parameter $\lambda > 0$. Consider $(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\pi}}) = \underset{\boldsymbol{\Omega}, \boldsymbol{g}, \boldsymbol{\pi}}{\arg\max}\, \mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X})$. Define $\{\widehat{\Upsilon}_1, \ldots, \widehat{\Upsilon}_{\widehat{K}_0}\}$ as the distinct values of $\widehat{\boldsymbol{\Omega}}$, that is, $\{k : \widehat{\boldsymbol{\Omega}}_k \equiv \widehat{\Upsilon}_l, k = 1, \ldots, K\}_{l=1,\ldots,\widehat{K}_0}$ constitutes a partition of $\{1, \ldots, K\}$. Then there are $\widehat{K}_0$ components with the estimated mean and precision parameters in $\widehat{\boldsymbol{\Omega}}$. The sparsity patterns of the precision matrix estimates directly correspond to the structures of the networks. Specifically, if and only if the (i, j)-th entry of the estimate for $\boldsymbol{\Theta}_k$ is zero, the corresponding two variables are not connected conditional on the other variables.

Rationale The proposed modeling approach considers the heterogeneity across clusters by incorporating grouping parameters in (2.4) and obtains the number of components $\widehat{K}_0$ using the penalized fusion technique. In (2.5), the first two penalties are imposed on the parameters of the mean and precision matrices, respectively, which are common practices and aim

to promote sparsity. The most notable advancement in our work lies in the third, the fusion penalty. This penalty serves to shrink the differences among the K components and encourage equality, ultimately resulting in a smaller number of components. In this article, the cluster-wise distributions in the model (2.2) share common Gaussian distributions f_l, which are defined by both mean $\boldsymbol{\mu}_l^*$, covariance matrix $\boldsymbol{\Sigma}_l^*$. Consequently, the fusion penalty is applied to both the mean and precision matrix parameters, which intuitively leads to a more effective penalty than solely applying it to a single parameter.

2.2 Computation

An Expectation-Conditional-Maximization (ECM) algorithm facilitated by the Alternating Direction Method of Multipliers (ADMM) algorithm is developed for optimizing objective function (2.4). The ECM algorithm performs a conditional maximization in each M-step, that is, each parameter π_{gk}, $\boldsymbol{\mu}_k$, $\boldsymbol{\Theta}_k$, and g_i is maximized separately by fixing other parameters. We estimate the grouping parameter $g_i (i = 1, \ldots, m)$ and other unknown parameters simultaneously via the maximum likelihood method. Here it is assumed that the number of groups G is known, and its estimation will be discussed later. To develop the ECM algorithm, we introduce latent vari-

ables $z_{ij} \in \{1, \ldots, K\}$ and consider a hierarchical expression of model (2.3) given by $\boldsymbol{x}_{ij}|(z_{ij} = k) \sim f_k(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k^{-1})$ and $Pr(z_{ij} = k) = \pi_{g_i k}$. Then the penalized log-likelihood function for the complete data can be written as:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X}) &= \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \mathrm{I}(z_{ij} = k) \left\{ \log f_k(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k^{-1}) + \log \pi_{g_i k} \right\} \\ &\quad - P(\boldsymbol{\Omega}). \end{aligned} \tag{2.6}$$

In the E-step, we compute the conditional probabilities of $z_{ij} = k$ for $k = 1, \ldots, K$, given the data and current parameter values $\boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}$. The expectation of the complete log-likelihood is given by:

$$\begin{aligned} &Q(\boldsymbol{\Omega}, \boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X}, \boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) \\ &= \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{K} \omega_{ijk}^{(t)} \log f_k(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k^{-1}) + \frac{1}{N} \sum_{i=1}^{m} \sum_{k=1}^{K} \log \pi_{g_i k} \sum_{j=1}^{n_i} \omega_{ijk}^{(t)} - P(\boldsymbol{\Omega}) \\ &\equiv Q_1(\boldsymbol{\Omega} \mid \boldsymbol{X}, \boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) + Q_2(\boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X}, \boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}) - P(\boldsymbol{\Omega}), \end{aligned} \tag{2.7}$$

where $\omega_{ijk}^{(t)}$ can be computed as:

$$\omega_{ijk}^{(t)} = \frac{\pi_{g_i k}^{(t-1)} f_k\left(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k^{(t-1)}, \left(\boldsymbol{\Theta}_k^{(t-1)}\right)^{-1}\right)}{\sum_{k=1}^{K} \pi_{g_i k}^{(t-1)} f_k\left(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_k^{(t-1)}, \left(\boldsymbol{\Theta}_k^{(t-1)}\right)^{-1}\right)}, i = 1, \ldots, m, j = 1, \ldots, n_i. \tag{2.8}$$

The maximization of $Q_1(\boldsymbol{\Omega} \mid \boldsymbol{X}, \boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)})$ in (2.7) with respect to $\boldsymbol{\Omega}$ can be divided into K maximization problems, and $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K$ can be separately updated. In addition, maximizing $Q_2(\boldsymbol{g}, \boldsymbol{\pi} \mid \boldsymbol{X}, \boldsymbol{\Omega}^{(t-1)}, \boldsymbol{g}^{(t-1)}, \boldsymbol{\pi}^{(t-1)})$

includes discrete optimization of $\boldsymbol{g}$ in the space $\{1,\ldots,G\}^m$, and such a search is computationally burdensome. Alternatively, we first maximize $Q_2(\boldsymbol{g}^{(t-1)},\boldsymbol{\pi}\mid \boldsymbol{X},\boldsymbol{\Omega}^{(t-1)},\boldsymbol{g}^{(t-1)},\boldsymbol{\pi}^{(t-1)})$ with respect to $\boldsymbol{\pi}$, denote $\boldsymbol{\pi}^{(t)}$ as the maximizer, and then maximize $Q_2(\boldsymbol{g},\boldsymbol{\pi}^{(t)}\mid \boldsymbol{X},\boldsymbol{\Omega}^{(t-1)},\boldsymbol{g}^{(t-1)},\boldsymbol{\pi}^{(t-1)})$ with respect to $\boldsymbol{g}$ to obtain $\boldsymbol{g}^{(t)}$. This allows a separate updating for each element of $\boldsymbol{g}$.

In the t-th maximization step, maximizing (2.7) with respect to π_{gk} leads to the estimate:

$$\pi_{gk}^{(t)}=\frac{1}{\sum\limits_{i:g_i^{(t)}=g}\sum\limits_{j=1}^{n_i}\sum\limits_{k=1}^{K}\omega_{ijk}^{(t)}}\sum_{i:g_i^{(t)}=g}\sum_{j=1}^{n_i}\omega_{ijk}^{(t)}. \tag{2.9}$$

For $\boldsymbol{\mu}_k$, maximizing (2.7) with respect to $\{\boldsymbol{\mu}\}=\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K$ is equivalent to solving:

$$\{\boldsymbol{\mu}^{(t)}\}=\arg\min_{\{\boldsymbol{\mu}\}}\left\{\frac{1}{2N}\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\omega_{ijk}^{(t)}\left[(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Theta}_k^{(t-1)}(\boldsymbol{x}_{ij}-\boldsymbol{\mu}_k)\right]+P(\{\boldsymbol{\mu}\})\right\}, \tag{2.10}$$

where

$$P(\{\boldsymbol{\mu}\})=\sum_{k=1}^{K}\sum_{r=1}^{p}p_{\lambda_1}(|\mu_{kr}|)+\sum_{k<k'}p_{\lambda_3}\left(\left(||\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k'}||_2^2+||\boldsymbol{\Theta}_k^{(t-1)}-\boldsymbol{\Theta}_{k'}^{(t-1)}||_F^2\right)^{\frac{1}{2}}\right).$$

Here the local quadratic approximation can be adopted and lead to an explicit solution at each iteration. Details are provided in the Supplementary Material.

For $\boldsymbol{\Theta}_k$, maximizing (2.7) with respect to $\{\boldsymbol{\Theta}\} = \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K$ is equivalent to solving:

$$\{\boldsymbol{\Theta}_k^{(t)}\} = \arg\max_{\{\boldsymbol{\Theta}\}} \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \omega_{ijk}^{(t)} \left[\log\{\det(\boldsymbol{\Theta}_{\boldsymbol{k}})\} - \mathrm{tr}(\widetilde{\boldsymbol{S}}_k \boldsymbol{\Theta}_{\boldsymbol{k}}) \right] - P(\{\boldsymbol{\Theta}\}) \right\}, \tag{2.11}$$

where $\widetilde{\boldsymbol{S}}_k$ is a pseudo sample covariance matrix defined as

$$\widetilde{\boldsymbol{S}}_k = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \omega_{ijk}^{(t)} (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_k^{(t)})^\top}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \omega_{ijk}^{(t)}},$$

and

$$P(\{\boldsymbol{\Theta}\}) = \sum_{k=1}^{K} \sum_{i \neq j} p_{\lambda_2}(|\theta_{kij}|) + \sum_{k<k'} p_{\lambda_3} \left(\left(||\boldsymbol{\mu}_k^{(t)} - \boldsymbol{\mu}_{k'}^{(t)}||_2^2 + ||\boldsymbol{\Theta}_k - \boldsymbol{\Theta}_{k'}||_F^2 \right)^{\frac{1}{2}} \right).$$

To effectively solve (2.11), the ADMM technique is employed. Different from the existing literature, the addition of the fusion penalty complicates the optimization process, rendering it more challenging. To address this issue, the efficient sparse alternating minimization algorithm (S-AMA)(Wang et al. (2018)) is adopted.

For updating g_i, we compute all the values of the objective function for $g = 1, \ldots, G$ and select the maximizer. That is, update g_i by solving:

$$g_i^{(t)} = \arg\max_{g=1,\ldots,G} \left\{ \frac{1}{N} \sum_{k=1}^{K} \log \pi_{gk}^{(t)} \sum_{j=1}^{n_i} \omega_{ijk}^{(t)} \right\}. \tag{2.12}$$

The ECM algorithm can be sensitive to the choice of initial estimators. In our numerical study, the initialization is implemented using the

K-means method, where the precision matrices are obtained by directly inverting the covariance matrices. We note that some other sensible clustering methods may also generate satisfactory initial values. Denote the obtained cluster-wise mixture probabilities as $\boldsymbol{\pi}_i^{(0)} = (\pi_{i1}^{(0)}, \ldots, \pi_{iK}^{(0)})$ and parameters $\boldsymbol{\Omega}_{ik}^{(0)}$. Then, the standard K-means algorithm with G groups is applied to $\boldsymbol{\pi}_1^{(0)}, \ldots, \boldsymbol{\pi}_m^{(0)}$, and we adopt the estimated grouping and centroids as the initial estimators of $\boldsymbol{g}$ and $\boldsymbol{\pi}$, respectively. The initial estimators of $\boldsymbol{\Omega}_k$ can be obtained via calculating the element-wise medians or means of $\{\boldsymbol{\Omega}_{1k}^{(0)}, \ldots, \boldsymbol{\Omega}_{mk}^{(0)}\}$ for $k = 1, \ldots, K$. The algorithm of the update process is summarized in Algorithm S1 of the Supplementary Material.

The proposed approach requires tuning $(\lambda_1, \lambda_2, \lambda_3)$ and G. To determine the optimal values of the tuning parameters, we minimize the following BIC-type criterion:

$$\text{BIC}_{G,\lambda} = -2Q(\widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\pi}})|_{K=\widehat{K}_0} + \log N \left\{ G(\widehat{K}_0 - 1) + m + \sum_{l=1}^{\widehat{K}_0} s_{1l} \right\} + 2 \sum_{l=1}^{\widehat{K}_0} s_{2l}, \tag{2.13}$$

where $s_{1l} = |\{r : \hat{\mu}_{lr} \neq 0, 1 \leq r \leq p\}|$, and $s_{2l} = |\{(i,j) : \hat{\theta}_{lij} \neq 0, 1 \leq i < j \leq p\}|$ for $l = 1 \ldots, \widehat{K}_0$.

Remark 1 To the best of our knowledge, there is currently no existing literature that provides a guarantee for the global convergence of the ECM algorithm in a general case. However, with a reasonably good initialization,

the ECM algorithm can be ensured to converge to a solution within a small neighborhood of the true solution, in terms of statistical accuracy, even if it may get trapped in a local optimum after a sufficient number of iterations. Specifically, we require the initial value to be within a ball of constant radius around the true solution. By conducting enough iterations, we can ensure that the optimization error is dominated by the statistical error, and as a result, the overall error bound is of the same order as the statistical error. Further details can be found in the Supplementary Material, specifically in the proof of Result 1.

Remark 2 The BIC-type criterion (2.13) is analogous to that of Sugasawa (2021) and Hao et al. (2018), which have demonstrated great performance in their studies. The limitation of our work lies in not establishing a rigorous theoretical guarantee for $\widehat{G}$. To our knowledge, providing such theoretical proof in high-dimensional settings poses challenges, and such limitations about information criterion are shared by many existing studies (Göbler et al. (2024), Ren et al. (2022), Hao et al. (2018)).

3. Theoretical properties

Denote the true parameter values as $\Upsilon^* = (\Upsilon_1^{*\top}, \ldots, \Upsilon_{K_0}^{*\top})^\top$ and $\Upsilon_l^* = \text{vec}(\boldsymbol{\mu}_l^*, \boldsymbol{\Theta}_l^*)$ for $l = 1, \ldots, K_0$. Define $\mathcal{S}_l = \{(i,j) : \theta_{lij}^* \neq 0, 1 \leq i \neq j \leq p\}$

as the index set of non-zero elements in the l-th precision matrix and the sparsity parameter $s = \max\{|\mathcal{S}_l|, l = 1, \ldots, K_0\}$. Similarly, define $\mathcal{D}_l = \{r : \mu^*_{lr} \neq 0, 1 \leq r \leq p\}$ as the non-zero elements in the l-th mean vector and the sparsity parameter $d = \max\{|\mathcal{D}_l|, l = 1, \ldots, K_0\}$. Without loss of generality, suppose that the cluster labels are assigned such that the cluster size increases with the cluster index, that is, $n_1 \leq n_2 \leq \cdots \leq n_m$. The following conditions are needed.

Condition 1. For some positive constants β_1, β_2, $0 < \beta_1 < \min_{l=1,\ldots,K_0} \psi_{\min}(\boldsymbol{\Theta}^*_l) < \max_{l=1,\ldots,K_0} \psi_{\max}(\boldsymbol{\Theta}^*_l) < \beta_2 < \infty$, where $\psi_{\min}(\boldsymbol{\Theta}^*_l)$ and $\psi_{\max}(\boldsymbol{\Theta}^*_l)$ are the smallest and largest eigenvalues of $\boldsymbol{\Theta}^*_l$, respectively.

Condition 2. $||\boldsymbol{\mu}^*||_\infty = \max_{l=1,\ldots,K_0} ||\boldsymbol{\mu}^*_l||_\infty$ and $||\boldsymbol{\Theta}^*||_\infty = \max_{l=1,\ldots,K_0} ||\boldsymbol{\Theta}^*_l||_\infty$ are bounded, where $||\boldsymbol{\Theta}^*_l||_\infty = \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\theta^*_{lij}|$ and θ^*_{lij} is the (i,j)-th entry of $\boldsymbol{\Theta}^*_l$.

Condition 3. The K_0 components are sufficiently separable such that each pair $\{(l, l'), 1 \leq l \neq l' \leq K_0\}$,

$$\mathrm{pr}(\boldsymbol{x} \in \mathcal{A}_l)\mathrm{pr}(\boldsymbol{x} \in \mathcal{A}_{l'}) \leq \frac{\varrho}{24(K_0 - 1)\sqrt{\max\{W, W', W''\}}},$$

where $\mathcal{A}_l$ and $\mathcal{A}_{l'}$ are the l-th and l'-th components, respectively, $\varrho = c \min\{\beta_1, \frac{1}{2}(\beta_2 + 2\alpha_0)^{-2}\}$ for a constant c. Here, $W = \max_l W_l$, $W' =$

$\max_l W_l'$, and $W'' = \max_l W_l''$ with

$$\begin{aligned} W_l &= \sup_{t\in[0,1]} \mathbb{E}\left[\left[\delta_{\Upsilon_{tl}}(\boldsymbol{x})\right]^\top \delta_{\Upsilon_{tl}}(\boldsymbol{x}) \left\|\boldsymbol{\Theta}_{l'}^*(\boldsymbol{x}-\boldsymbol{\mu}_{l'}^*)\right\|_2^2\right], \\ W_l' &= \sup_{t\in[0,1]} \mathbb{E}\left[\left[\delta_{\Upsilon_{tl}}(\boldsymbol{x})\right]^\top \delta_{\Upsilon_{tl}}(\boldsymbol{x}) \left\|\boldsymbol{\Theta}_{l'}^{*-1}\right\|_F^2\right], \\ W_l'' &= \sup_{t\in[0,1]} \mathbb{E}\left[\left[\delta_{\Upsilon_{tl}}(\boldsymbol{x})\right]^\top \delta_{\Upsilon_{tl}}(\boldsymbol{x}) \left\|(\boldsymbol{x}-\boldsymbol{\mu}_{l'}^*)(\boldsymbol{x}-\boldsymbol{\mu}_{l'}^*)^\top\right\|_F^2\right], \end{aligned}$$

Define $\tilde{\Upsilon}_t = \Upsilon^* + t(\Upsilon - \Upsilon^*)$, $\tilde{\Upsilon}_t = (\tilde{\Upsilon}_{t1}, \ldots, \tilde{\Upsilon}_{tK_0})$, and $\tilde{\Upsilon}_{tl} = \mathrm{vec}(\tilde{\boldsymbol{\mu}}_{tl}, \tilde{\boldsymbol{\Theta}}_{tl})$ with $t \in [0,1]$. For any $\Upsilon \in \mathcal{B}_{\alpha_0}(\Upsilon^*) = \{\Upsilon : \|\Upsilon - \Upsilon^*\|_2 \le \alpha_0\}$:

$$\delta_{\Upsilon_{tl}}(\boldsymbol{x}) = \begin{pmatrix} \tilde{\boldsymbol{\Theta}}_{tl}(\boldsymbol{x} - \tilde{\boldsymbol{\mu}}_{tl}) \\ \frac{1}{2}\mathrm{vec}\{\tilde{\boldsymbol{\Theta}}_{tl}^{-1} - (\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{tl})(\boldsymbol{x}-\tilde{\boldsymbol{\mu}}_{tl})^\top\} \end{pmatrix}.$$

Condition 4. $\min\left\{\{|\mu_{lr}^*| : r \in \mathcal{D}_l, l = 1, \ldots, K_0\}, \{|\theta_{lij}^*| : (i,j) \in \mathcal{S}_l, l = 1, \ldots, K_0\}\right\} > (a + \frac{1}{2}) \cdot \max\{\lambda_1, \lambda_2\}$. Let $b = \min_{1 \le l \ne l' \le K_0} ||\Upsilon_l^* - \Upsilon_{l'}^*||_2$, and therefore $b > (a + \frac{1}{2})\lambda_3$.

Condition 5. $\lambda_1, \lambda_2, \lambda_3 \gg \sqrt{(s+p)\log p / N}$.

Condition 6. $\rho(t) = \lambda^{-1} p_\lambda(t)$ is concave in $t \in [0, \infty)$ with a continuous derivative $\rho'(t)$ satisfying $\rho(0) = 0$, and $\rho'(0+) = 1$ is independent of λ. There exists a constant $0 < a < \infty$ such that $\rho(t)$ is constant for all $t \ge a\lambda$.

Condition 7. The probability density function $g(\boldsymbol{X}; \boldsymbol{g}, \Upsilon^*) \equiv \prod_{i=1}^m \prod_{j=1}^{n_i} f_i(\boldsymbol{x}_{ij}; g_i, \Upsilon^*)$ is identifiable in $(\boldsymbol{g}, \boldsymbol{\pi}, \boldsymbol{\Omega})$ up to the permutation of the component and grouping labels, where $f_i(\boldsymbol{x}_{ij}; g_i, \Upsilon^*) = \sum_{l=1}^{K_0} \pi_{g_i l}^* f_l(\boldsymbol{x}_{ij}; \boldsymbol{\mu}_l^*, \boldsymbol{\Theta}_l^{*-1})$ is a mixture distribution.

Condition 8. For the interior point Υ^0 from a compact set $\boldsymbol{S} \in \mathbb{R}^{K_0(p+p^2)}$, there exist functions $M_t(\boldsymbol{x}), t = 1, 2, 3$, such that for Υ in a neighborhood of $N(\Upsilon^0)$,

$$\frac{f_i(\boldsymbol{x}; g_i, \Upsilon)}{f_i(\boldsymbol{x}; g_i', \Upsilon)} < M_1(\boldsymbol{x}), \left|\frac{\partial \log f_i(\boldsymbol{x}; g_i', \Upsilon)}{\partial \Upsilon_s}\right| < M_2(\boldsymbol{x}), \left|\frac{\partial^2 \log f_i(\boldsymbol{x}; g_i', \Upsilon)}{\partial \Upsilon_s \partial \Upsilon_{s'}}\right| < M_3(\boldsymbol{x}),$$

for arbitrary $\boldsymbol{g}$ and $\boldsymbol{g}'$, where $\boldsymbol{g}'$ as the collection of g_i'. Υ_s and $\Upsilon_{s'}$ are different components of Υ. $E[M_1^2(\boldsymbol{x})] < \infty$, $E[M_2^2(\boldsymbol{x})] < \infty$ and $E[M_3^2(\boldsymbol{x})] < \infty$.

Conditions 1 and 2 are commonly assumed in the literature on high-dimensional heterogeneity analysis. Condition 3, also known as the sufficiently separable condition, requires that if a sample belongs to a component with a probability close to 1, then its probability of belonging to any other component should be close to 0. Further discussions on this condition can be found in Hao et al. (2018). Condition 4 specifies the minimum signals and minimal differences across sample components. Condition 5 provides the orders of the tuning parameters. Finally, Condition 6 is a common assumption and is satisfied by SCAD (Fan and Li (2001)) and MCP (Zhang (2010)). Conditions 7 and 8 are regularity conditions for establishing the consistency of grouping parameters. Note that the finite mixtures of Poisson, negative binomial, and most continuous distributions, including the Gaussian distribution, satisfy Condition 7. With these conditions and given $K > K_0$ is bounded, we can establish the following consistency results.

Theorem 1. *Suppose that Conditions 1-8 hold, assume that* $(s+p)\log p/N = o(1)$. *There exists a local maximum of (2.4) that satisfies:*

1. *(Consistency of* $\hat{g}_i$*)* $\frac{1}{m}\sum_{i=1}^{m} I(\hat{g}_i \neq g_i^*) = o_p(n_1^{-\delta})$ *for any* $\delta > 0$, *where* g_i^* *and* $\hat{g}_i$ *are the true parameter value and estimate of* g_i, *respectively.*

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

3. *(Rate of convergence)* $\sum_{l=1}^{K_0}\left(||\widehat{\boldsymbol{\mu}}_l - \boldsymbol{\mu}_l^*||_2 + ||\widehat{\boldsymbol{\Theta}}_l - \boldsymbol{\Theta}_l^*||_F\right) = O_p\left(\sqrt{\frac{(s+p)\log p}{N}}\right)$.

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

Remark 3 Theorem 1 demonstrates that the proposed approach exhibits the well-desired consistency properties, particularly in terms of accurately identifying the number of components and the weak consistency of grouping parameters. These results represent a significant advancement over much of the existing literature, where such an identification has been nontrivial. Note that K is bounded in Theorem 1, this setting is sufficient and enables our statistical rate for the precision matrix estimation under the Frobenius norm to achieve the optimal rate $O(\sqrt{(s+p)\log p/N})$ established in Cai et al. (2016). With regard to sparsistency, the concave penalty offers the advantage that precision matrices can exhibit less sparsity compared to those penalized under the L_1 penalty (where the order of s is no larger than

$O(p)$). Although the consistency of estimation and variable selection results may not be deemed "surprising" in light of the existing literature, it is important to emphasize that our study involves high-dimensional clustered data with network structures. This represents a more complex scenario compared to many existing studies, as it includes them as special cases when $m = 1$. To the best of our knowledge, there has been very limited research on these results in the model settings addressed in our study. The proof is provided in the Supplementary Material.

4. Numerical studies

In this section, we assess the finite sample performance of the proposed approach and compare it with relevant alternative methods. We consider a three-class problem ($K_0 = 3$) with $p = 100$. We set equal cluster sizes in all clusters, i.e., $n_i = n$ for $i = 1, \ldots, m$, where m denotes the number of clusters. We consider various combinations of m and n, including $(m, n) = (20, 80), (20, 160), (40, 80)$, and $(40, 160)$. The observations are generated according to the following procedure. The component labels $L_{ij}{}'$s are sampled from $\{1, 2, 3\}$ with probability $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$ for the i-th cluster, that is, each component for i-th cluster can have a different sample size. Then we generate $\boldsymbol{x}_{ij} \sim N(\boldsymbol{\mu}(L_{ij}), \boldsymbol{\Sigma}(L_{ij}))$. The first eight compo-

nents of the component means are $(\mu\mathbf{1}_4^\top, -\mu\mathbf{1}_4^\top)^\top I(L_{ij}=1) + \mu\mathbf{1}_8 I(L_{ij}=2) + (-\mu)\mathbf{1}_8 I(L_{ij}=3)$, and the rest $p-8$ components are set as 0. For μ, we consider 1.5 and 2.5. Thereby, all the clusters share the same mean and covariance, and variations of the cluster-wise distributions are determined by the cluster-wise mixture probabilities $\boldsymbol{\pi}_i, i = 1, \ldots, m$. To generate $\boldsymbol{\pi}_i$, we use a symmetric Dirichlet distribution with the concentration parameter of $\alpha_d = 2$.

For the structure of the precision matrices, we consider three popular choices, namely, tridiagonal, nearest-neighbor, and power-law. (1) *Tridiagonal network*: The diagonal elements are all equal to one, and the nonzero off-diagonal elements are $0.2I(L_{ij}=1)+0.3I(L_{ij}=2)+0.4I(L_{ij}=3)$. Under the second and third types, each network is composed of 10 equally sized and separate subnetworks. Out of the 10 subnetworks, eight are shared by the three components. For the remaining two subnetworks, pairs share one subnetwork in common, while each component has a unique subnetwork. We consider the following two commonly used methods for generating the subnetworks. (2) *Nearest-neighbor network*: For each subnetwork in the first component, $p/10$ points are generated randomly on a unit square, and all $p/10 \times (p/10-1)/2$ pairwise distances are calculated. Then we find the n' nearest neighbors of each point. The nearest-neighbor network is

obtained by linking any two points that are among the n'-nearest neighbors of each other. n' controls the degree of sparsity, and we set $n' = 3$. (3) *Power-law network*: For the precision matrix of the first component, 10 power-law subnetworks are generated with two edges added in each step. The initial 10-block precision matrix $(\theta_{1ij})_{p\times p}$ is generated via:

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

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 consider $\theta_{1jj} = \sum_{i\neq j} |\theta_{1ij}| + 0.1$. It is noted that the component-specific subnetworks are generated similarly and independently in the latter two networks.

With the proposed approach, we set $K = 6$, and G is determined using the information criterion (2.13) from the set of $\{1, 2, \ldots, 10\}$. For comparison, we combine the classical *glasso* method (with $K = 2$, 3, 4, and 6) with the grouping method introduced in this article, which we refer to as "Glasso+grouping". Similarly, we also adopt the "JGL+grouping" method. To evaluate the performance of our proposed approach, we consider the following measures: (a) The ratio of the number of times $\widehat{K}_0$ equals K_0 to the total number of runs (denoted as 'Ratio'), (b) mean and standard deviation

(sd) of $\widehat{K}_0$, (c) mean and sd of selected G, (d) component membership error defined by

$$\begin{aligned}\mathrm{CE}(\widehat{\varphi}_i, \varphi_i) &= \binom{N}{2}^{-1} \sum_{i=1}^{m} |\{(a,b): \\ &\qquad I(\widehat{\varphi}_i(\boldsymbol{x}_{ia}) = \widehat{\varphi}_i(\boldsymbol{x}_{ib})) \neq I(\varphi_i(\boldsymbol{x}_{ia}) = \varphi_i(\boldsymbol{x}_{ib})); a < b\}|,\end{aligned}$$

where $\widehat{\varphi}_i$ and φ_i are the estimated and true component membership for the i-th cluster, respectively, (e) root mean squared error (RMSE) for $\boldsymbol{\mu}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\pi}$. When $\widehat{K}_0 \neq K_0$, the RMSEs are defined as

$$\begin{aligned}\mathrm{RMSE}(\boldsymbol{\mu}) &= \frac{1}{\widehat{K}_0} \sum_{k=1}^{\widehat{K}_0} \sum_{l'=1}^{K_0} \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^*_{l'}\|_2 \\ &\qquad \cdot I(l' = \arg\min_{l} \{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^*_l\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}^*_l\|_F^2\}), \\ \mathrm{RMSE}(\boldsymbol{\Theta}) &= \frac{1}{\widehat{K}_0} \sum_{k=1}^{\widehat{K}_0} \sum_{l'=1}^{K_0} \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}^*_{l'}\|_F \\ &\qquad \cdot I(l' = \arg\min_{l} \{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^*_l\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}^*_l\|_F^2\}).\end{aligned}$$

When $\widehat{K}_0 = K_0$, the RMSEs are defined as $\mathrm{RMSE}(\boldsymbol{\mu}) = \sum_{k=1}^{K_0} \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_2 / K_0$ and $\mathrm{RMSE}(\boldsymbol{\Theta}) = \sum_{k=1}^{K_0} \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_k\|_F / K_0$. Note that $\mathrm{RMSE}(\boldsymbol{\pi})$ is only defined under $\widehat{K}_0 = K_0$: $\mathrm{RMSE}(\boldsymbol{\pi}) = \sum_{g=1}^{G} \sum_{i:g_i=g} \|\widehat{\boldsymbol{\pi}}_g - \boldsymbol{\pi}_i\|_2 / G$, (f) the true positive rate (TPR, percentage of true edges selected) and the false positive rate (FPR, percentage of false edges selected) for the off-diagonal

elements of the precision matrices. When $\widehat{K}_0 = K_0$, they are defined as

$$\begin{aligned}
\text{TPR} &= \frac{1}{K_0}\sum_{k=1}^{K_0}\frac{\sum_{i<j} I(\theta_{kij}\neq 0, \hat{\theta}_{kij}\neq 0)}{\sum_{i<j} I(\theta_{kij}\neq 0)}, \\
\text{FPR} &= \frac{1}{K_0}\sum_{k=1}^{K_0}\frac{\sum_{i<j} I(\theta_{kij}= 0, \hat{\theta}_{kij}\neq 0)}{\sum_{i<j} I(\theta_{kij}= 0)}.
\end{aligned}$$

When $\widehat{K}_0 \neq K_0$, we have definitions:

$$\begin{aligned}
\text{TPR} &= \frac{1}{\widehat{K}_0}\sum_{k=1}^{\widehat{K}_0}\sum_{l'=1}^{K_0}\frac{\sum_{i<j} I(\theta_{kij}\neq 0, \hat{\theta}_{kij}\neq 0)}{\sum_{i<j} I(\theta_{kij}\neq 0)} \\
&\quad \cdot I(l' = \arg\min_{l}\{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_l^*\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_l^*\|_F^2\}), \\
\text{FPR} &= \frac{1}{\widehat{K}_0}\sum_{k=1}^{\widehat{K}_0}\sum_{l'=1}^{K_0}\frac{\sum_{i<j} I(\theta_{kij}= 0, \hat{\theta}_{kij}\neq 0)}{\sum_{i<j} I(\theta_{kij}= 0)} \\
&\quad \cdot I(l' = \arg\min_{l}\{\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_l^*\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}_l^*\|_F^2\}).
\end{aligned}$$

With 100 replicates, the summary results for the different scenarios of (m, n) and $\mu = 1.5$ are presented in Table 1, and the additional results can be found in Tables S1-S3 of Supplementary Material. Similar observations are made across different settings. Notably, JGL+grouping demonstrates satisfactory performance when the number of components is correctly specified. Similarly, Glasso+grouping also demonstrates good performance when $K = 3$, but it tends to select more complex models compared to JGL+grouping and the proposed method. The results of Glasso+grouping may be unsatisfactory when K is incorrectly specified. The proposed approach demonstrates competitive performance across the

Table 1: Simulation results under $(m, n) = (20, 80), (20, 160)$ and $\mu = 1.5$.

In each cell, mean (sd).

(m,n)	Network	Method	CE	RMSE($\boldsymbol{\mu}$)	RMSE($\boldsymbol{\Theta}$)	RMSE($\boldsymbol{\pi}$)	TPR	FPR	Ratio	$\widehat{K}_0$	G
(20,80)	Tridiagonal	JGL+grouping(K=3)	0.002(0.001)	0.137(0.022)	1.118(0.038)	0.057(0.016)	0.996(0.003)	0.145(0.004)	1	3(0)	6.48(1.10)
		Glasso+grouping(K=2)	0.210(0.031)	2.005(0.402)	1.422(0.133)	-	0.995(0.002)	0.213(0.067)	0	2(0)	4.56(1.32)
		Glasso+grouping(K=3)	0.001(0.000)	0.135(0.025)	1.125(0.020)	0.064(0.015)	1.000(0.000)	0.201(0.005)	1	3(0)	6.50(1.28)
		Glasso+grouping(K=4)	0.059(0.015)	0.340(0.091)	1.402(0.360)	-	0.989(0.004)	0.210(0.037)	0	4(0)	5.40(1.53)
		Glasso+grouping(K=6)	0.116(0.031)	0.351(0.147)	1.435(0.402)	-	0.978(0.034)	0.212(0.021)	0	6(0)	4.35(1.50)
		Grouped-HGGM	0.004(0.014)	0.160(0.040)	1.154(0.051)	0.068(0.025)	0.965(0.011)	0.147(0.014)	0.95	3.06(0.30)	6.18(1.27)
	Nearest-neighbor	JGL+grouping(K=3)	0.002(0.002)	0.124(0.029)	1.615(0.067)	0.055(0.010)	0.961(0.006)	0.164(0.004)	1	3(0)	6.50(1.28)
		Glasso+grouping(K=2)	0.215(0.027)	1.975(0.451)	1.684(0.090)	-	0.994(0.006)	0.225(0.010)	0	2(0)	4.51(1.44)
		Glasso+grouping(K=3)	0.005(0.002)	0.137(0.031)	1.480(0.069)	0.060(0.015)	0.982(0.010)	0.215(0.008)	1	3(0)	6.67(1.30)
		Glasso+grouping(K=4)	0.054(0.014)	0.337(0.175)	1.557(0.134)	-	0.975(0.010)	0.235(0.014)	0	4(0)	5.65(1.45)
		Glasso+grouping(K=6)	0.129(0.025)	0.387(0.221)	1.811(0.750)	-	0.917(0.014)	0.223(0.044)	0	6(0)	4.40(1.58)
		Grouped-HGGM	0.004(0.018)	0.023(0.088)	1.405(0.128)	0.060(0.019)	0.923(0.018)	0.157(0.015)	0.95	3.07(0.29)	5.87(1.29)
	Power-law	JGL+grouping(K=3)	0.004(0.005)	0.189(0.038)	1.354(0.060)	0.065(0.013)	0.961(0.006)	0.166(0.006)	1	3(0)	6.49(1.30)
		Glasso+grouping(K=2)	0.213(0.034)	2.101(0.425)	1.725(0.122)	-	0.989(0.007)	0.224(0.011)	0	2(0)	4.47(1.41)
		Glasso+grouping(K=3)	0.002(0.002)	0.255(0.043)	1.393(0.071)	0.065(0.019)	0.985(0.009)	0.206(0.006)	1	3(0)	6.64(1.28)
		Glasso+grouping(K=4)	0.064(0.018)	0.311(0.167)	1.489(0.156)	-	0.976(0.008)	0.223(0.011)	0	4(0)	5.42(1.50)
		Glasso+grouping(K=6)	0.128(0.023)	0.359(0.151)	1.664(0.582)	-	0.924(0.012)	0.236(0.035)	0	6(0)	4.35(1.24)
		Grouped-HGGM	0.005(0.018)	0.024(0.028)	1.488(0.112)	0.075(0.045)	0.940(0.018)	0.158(0.009)	0.94	3.06(0.29)	5.91(1.33)
(20,160)	Tridiagonal	JGL+grouping(K=3)	0.002(0.001)	0.128(0.038)	0.834(0.034)	0.050(0.010)	0.998(0.001)	0.126(0.010)	1	3(0)	8.48(1.46)
		Glasso+grouping(K=2)	0.230(0.025)	1.868(0.388)	1.294(0.115)	-	0.999(0.000)	0.199(0.064)	0	2(0)	5.70(1.84)
		Glasso+grouping(K=3)	0.001(0.001)	0.115(0.019)	0.854(0.018)	0.042(0.011)	1.000(0.000)	0.172(0.005)	1	3(0)	8.48(1.34)
		Glasso+grouping(K=4)	0.061(0.015)	0.177(0.045)	1.104(0.122)	-	1.000(0.000)	0.217(0.026)	0	4(0)	7.36(1.85)
		Glasso+grouping(K=6)	0.119(0.027)	0.211(0.080)	1.187(0.342)	-	0.995(0.008)	0.204(0.025)	0	6(0)	5.81(1.78)
		Grouped-HGGM	0.001(0.001)	0.107(0.031)	0.797(0.049)	0.050(0.011)	1.000(0.000)	0.116(0.015)	0.98	3.04(0.15)	8.11(1.37)
	Nearest-neighbor	JGL+grouping(K=3)	0.003(0.001)	0.154(0.041)	1.115(0.034)	0.055(0.007)	0.975(0.004)	0.137(0.008)	1	3(0)	8.45(1.44)
		Glasso+grouping(K=2)	0.215(0.025)	2.017(0.454)	1.310(0.091)	-	0.985(0.004)	0.204(0.007)	0	2(0)	6.24(1.76)
		Glasso+grouping(K=3)	0.002(0.001)	0.088(0.019)	1.104(0.045)	0.046(0.012)	0.998(0.005)	0.199(0.005)	1	3(0)	8.40(1.52)
		Glasso+grouping(K=4)	0.062(0.015)	0.211(0.160)	1.178(0.145)	-	0.992(0.002)	0.216(0.014)	0	4(0)	7.45(1.55)
		Glasso+grouping(K=6)	0.135(0.028)	0.248(0.179)	1.184(0.112)	-	0.977(0.008)	0.203(0.035)	0	6(0)	5.98(1.81)
		Grouped-HGGM	0.001(0.003)	0.075(0.038)	0.933(0.075)	0.060(0.010)	0.984(0.004)	0.144(0.029)	0.97	3.04(0.17)	7.78(1.75)
	Power-law	JGL+grouping(K=3)	0.001(0.001)	0.160(0.048)	1.112(0.033)	0.059(0.010)	0.975(0.005)	0.139(0.010)	1	3(0)	8.61(1.41)
		Glasso+grouping(K=2)	0.215(0.028)	2.107(0.428)	1.433(0.131)	-	0.992(0.006)	0.216(0.009)	0	2(0)	6.22(1.91)
		Glasso+grouping(K=3)	0.002(0.003)	0.092(0.040)	1.106(0.036)	0.046(0.012)	0.991(0.003)	0.194(0.012)	1	3(0)	8.67(1.43)
		Glasso+grouping(K=4)	0.057(0.018)	0.195(0.121)	1.174(0.177)	-	0.995(0.003)	0.216(0.011)	0	4(0)	7.56(1.75)
		Glasso+grouping(K=6)	0.124(0.023)	0.253(0.118)	1.233(0.164)	-	0.972(0.006)	0.212(0.020)	0	6(0)	5.65(1.71)
		Grouped-HGGM	0.001(0.005)	0.084(0.031)	0.891(0.071)	0.049(0.015)	0.984(0.006)	0.140(0.014)	0.97	3.03(0.21)	7.92(1.79)

entire range of simulations. As an illustrative example, we examine Table 1, which corresponds to the Power-law setting with $(m, n) = (20, 160)$. Grouped-HGGM exhibits satisfactory TPR, FPR, RMSE values, and a Ratio of 0.97. In contrast, the other alternatives display significantly poorer estimation performance with considerably larger RMSEs. Additionally, it is noted that increasing sample size improves numerical performance for all measures. Here we also report the average values of the selected number of groups G. It is observed that, as the variations in sample size n for each cluster or the number of clusters m increase, the average value of G tends to increase as well. This finding is in line with that made by Sugasawa (2021). An additional insight from the Glasso+grouping method with different K values is that an incorrectly specified K can lead to different values of G. We perform an additional simulation study on μ to see how well the algorithm can detect the number of components as a function of problem hardness. The results can be found in Section S4 of Supporting Materials.

5. Analysis of TCGA data

According to the latest global burden of cancer survey, breast cancer has surpassed lung cancer as the most commonly diagnosed cancer, with an

estimated 2.3 million new cases (Sung et al. (2021)). In this section, we analyze breast invasive carcinoma data collected from The Cancer Genome Atlas (TCGA (2023)). The analysis aims to achieve two primary objectives: (i) Identify the number of components (subtypes) and their corresponding network structures to investigate gene expression patterns among different components. (ii) Group datasets from 19 source sites to explore the heterogeneity among these source sites through the group-specific mixture probability.

The mRNA expression data for tumor tissues are obtained from a diverse set of 907 samples collected from 19 reputable source sites, including renowned universities, hospitals, and cancer research centers. Due to the limitation in sample size, it is more reliable to focus on analyzing the most promising pathway rather than conducting a genome-wide analysis. In our study, a total of 147 gene expressions are selected for downstream analysis. These genes belong to the Kyoto Encyclopedia of Genes and Genomes (KEGG (2023)) hsa05224 pathway, which is named "breast cancer" and contains well-known breast cancer-related genes such as HER2 (also called ERBB2), MYC, WTN, HRAS, NRAS, BRAF, BRCA1, NOTCH, and others. Although the data has been examined in quite a few published studies, it is noted that the perspectives taken in the published studies are signifi-

cantly different from the proposed.

The methodology presented in this article is conducted in an unsupervised manner. When implementing our proposed approach, we selected a value of $K = 8$. This choice is considered sufficiently large since there is currently no evidence in existing research suggesting that the number of breast invasive carcinoma subtypes exceeds 8. Following the implementation, a total of five distinct components are identified, with sizes of 385, 205, 131, 161, and 25. Additional comprehensive information regarding these components can be obtained directly from the authors upon request. Regarding these five components, we present the corresponding estimated gene expression networks in Figure S1 (Supplementary Material). These networks exhibit 464, 783, 741, 978, and 953 edges, respectively. Furthermore, Table S4 (Supplementary Material) demonstrates that these networks display a varying degree of overlap in their edges, ranging from small to moderate. Table 2 compares the identified five components with four clinically confirmed subtypes. The Rand index (RI), measuring the degree of agreement between these two grouping approaches, yields a value of RI = 0.7442, suggesting certain consistency. The orange nodes of Figure S1 display 25 genes with the highest degrees in each component, highlighting notable distinctions among the sample groups. Our results are summarized as follows:

(i) The orange nodes of Figure S1 show that ERBB2 is found to be over-expressed exclusively in component 3, as the HER2+ ones are mostly in component 3 from Table 2. (ii) Our results in Table 2 demonstrate that the basal-like ones are predominantly identified in component 4 by the proposed approach. Additionally, JAG1, JAG2, and DLL3, known markers of triple-negative breast invasive carcinoma, exhibit high expression levels (the orange nodes) in the center-right panel of Figure S1.

Table 2: Compare the identified five components with four clinically confirmed subtypes.

	Component 1	Component 2	Component 3	Component 4	Component 5	Sum
Luminal A	318	113	20	0	20	471
Luminal B	60	92	38	0	3	193
Basal-like	0	0	7	160	0	167
HER2+	7	0	66	1	2	76
Sum	385	205	131	161	25	907

The unsupervised component/subgroup analysis demonstrates excellent performance. In addition to examining gene behavior, we also explore the optimal grouping of 19 source sites. The results of the grouping strategy are presented in Table 3. Within the same group, source sites share identical mixture probabilities (the last column), thus constituting the same mixture distribution. Variation in mixture probabilities indicates hetero-

geneity among groups, thereby facilitating a comprehensive interpretation of the characteristics of the four groups. Specifically, the mixture probabilities of the fifth component are 0.0096, 0.0845, 0.4375, and 0.0000 in groups 1 to 4, respectively, so that the fifth component can be a meaningful component to distinguish the four groups. Furthermore, we conduct an additional analysis of the data employing the Glasso+grouping and obtain RI = 0.7149. Since the alternative approach does not compress the number of components, we specify the number of components as 4. Detailed results can be found in Tables S5-S7 and Figure S2 of the Supplementary Material.

6. Discussion

In this article, we propose a new method called grouped heterogeneous Gaussian graphical model (Grouped-HGGM), which provides a comprehensive and flexible framework for unsupervised heterogeneity analysis in high-dimensional clustered data with network structures. The main strength of Grouped-HGGM lies in the expression of cluster-wise conditional distributions as finite mixtures of latent distributions that are shared across all clusters. Moreover, the between-cluster heterogeneity is captured through a cluster-wise mixture probability that follows a carefully designed group-

Table 3: The grouping strategy of the 19 source sites using proposed approach.

Group label	Source site	TSS code	Group-specific mixture probability
Group 1	Indivumed	A8	$\pi_1 = (0.4540, 0.2172, 0.1345, 0.1847, 0.0096)$
	ILSBio	C8	
Group 2	Christiana Healthcare	A7	$\pi_2 = (0.4244, 0.1990, 0.0597, 0.2324, 0.0845)$
	International Genomics Consortium	AC	
	Asterand	E9	
	University of Chicago	OL	
	Albert Einstein Medical Center	S3	
Group 3	UCSF	A1	$\pi_3 = (0.1250, 0.0625, 0.1250, 0.2500, 0.4375)$
	Walter Reed	A2	
	Cureline	AN	
	MSKCC	AO	
	Mayo	AR	
	Duke	B6	
	University of Pittsburgh	BH	
	Greater Poland Cancer Center	D8	
	Roswell Park	E2	
	University of Miami	EW	
	MD Anderson	GM	
Group 4	Candler	LL	$\pi_4 = (0.3070, 0.3340, 0.2923, 0.0667, 0.0000)$

ing structure. Our methodology incorporates several advancements to enhance the analysis of heterogeneity. Firstly, we have utilized the association between variables to enable efficient detection and characterization of heterogeneity. Additionally, we have developed a data-driven approach to determine the number of components in a fully automated way. We have estimated both the means and precision matrices to capture the underlying structure of the data accurately, and we have introduced the grouping strategy to model clustered data flexibly, further enhancing the versatility and applicability of our approach. The theoretical development of our methodology is challenging, as demonstrated in the accompanying Supplementary Material, while also providing implications for other high-dimensional heterogeneity and network analyses. We have developed a generalized Expectation-Conditional-Maximization (ECM) algorithm, facilitated by the Alternating Direction Method of Multipliers (ADMM) technique, for effectively estimating sparse parameters and unknown grouping parameters. Our numerical studies have demonstrated the promising performance of the Grouped-HGGM in accurately identifying the number of components in clustered data. Our analysis of real-world breast invasive carcinoma data has provided valuable insights into disease heterogeneity, highlighting the potential application of our methodology in other cancer

types and diverse fields such as neuroscience (Galbraith, Daniel and Vissel, 2010), personalized medicine (Fokkema et al., 2018), and labor economics (Pereda-Fernandez, 2021).

Our study opens up several potential avenues for future research. The penalized fusion technique employed in our heterogeneity analysis can be extended to more complex scenarios, such as investigating conditional network structures, thereby enabling data-driven determination of the number and structure of components. Furthermore, the grouping strategy represents an improvement over random effects analysis by providing an enhanced approach for addressing heterogeneity. This strategy can also be explored in the analysis of heterogeneity in numerous non-Gaussian data types, or alternatively, considered in supervised and semi-supervised studies, which can be a prospective avenue for future research. In addition, while we have not focused on allowing different clusters to have different K_0 values, this could be achieved by imposing sparsity of the group-specific mixture probabilities $\boldsymbol{\pi}_{g_i}^*$, thereby serving other analytical objectives. By addressing these directions, our research has the potential to significantly advance the understanding and analysis of heterogeneity in clustered data, providing valuable insights in a wide range.

Supplementary Materials

Contain the additional computational, theoretical and numerical results in the online supplementary materials.

Acknowledgements

We thank the Editor, Associate Editor, and two reviewers for their careful review and insightful comments. This study is supported by the Humanities and Social Science Foundation of Ministry of Education of China 24YJA910007, NIH CA204120, and NSF 2209685.

References

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

Chen, X., Feng, Z. and Peng, H. (2023). Estimation and order selection for multivariate exponential power mixture models. *Journal of Multivariate Analysis* **195**, 105140.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. Roy. Statist. Soc. B* **76**, 373–397.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T. and Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav. Res. Methods* **50**, 2016–2034.

Galbraith, S., Daniel, J. A. and Vissel, B. (2010). A study of clustered data and approaches to its analysis. *J. Neurosci.* **30**, 10601–10608.

Gao, C., Zhu, Y., Shen, X. and Pan, W. (2016). Estimation of multiple networks in gaussian mixture models. *Electron. J. Stat.* **10**, 1133–1154.

Göbler, K., Drton, M., Mukherjee, S., & Miloschewski, A. (2024). High-dimensional undirected graphical models for arbitrary mixed data. *Electron. J. Stat.* **18**, 2339–2404.

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

Hao, B., Sun, W. W., Liu, Y. and Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *J. Mach. Learn. Res.* **18**, 7981–8038.

KEGG. (Kyoto Encyclopedia of Genes and Genomes). https://www.genome.jp/pathway/hsa05224. Accessed on 7/16/2023.

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

McLachlan, G. J. and Peel, D. (2000). Finite mixture models, New York: Wiley.

Pei, Y., Peng, H. and Xu, J. (2022). A latent class Cox model for heterogeneous time-to-event

data. *Journal of Econometrics* **239**, 105351.

Pereda-Fernandez, S. (2021). Copula-based random effects models for clustered data. *J. Bus. Econom. Statist.* **39**, 575–588.

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

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008). The nested dirichlet process. *J. Amer. Statist. Assoc.* **103**, 1131–1154.

Sugasawa, S. (2021). Grouped heterogeneous mixture modeling for clustered data. *J. Amer. Statist. Assoc.* **116**, 999–1010.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249.

TCGA. (The Cancer Genome Atlas). https://portal.gdc.cancer.gov/projects/TCGA-BRCA. Accessed on 7/16/2023.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.

Wang, B., Zhang, Y., Sun, W. W. and Fang, Y. (2018). Sparse convex clustering. *J. Comput. Graph. Statist.* **27**, 393–403.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann.*

Statist. **38**, 894–942.

Xin Zeng, Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen, China

E-mail: zengxin@stu.xmu.edu.cn

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

E-mail: shuangge.ma@yale.edu

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

E-mail: qzzhang@xmu.edu.cn