Received: 2 July 2021 | Accepted: 11 May 2022

DOI: 10.1111/rssc.12575

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Series C Applied Statistics

Heterogeneous graphical model for non-negative and non-Gaussian $PM_{2.5}$ data

Jiaqi Zhang[1] | **Xinyan Fan**[1] | **Yang Li**[1,2] | **Shuangge Ma**[3]

[1]Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China

[2]RSS and China-Re Life Joint Lab on Public Health and Risk Management, Renmin University of China, Beijing, China

[3]Department of Biostatistics, Yale University, New Haven, USA

Correspondence
Xinyan Fan, Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China.
Email: 20198102@ruc.edu.cn

Yang Li, Center for Applied Statistics, School of Statistics, and RSS and China-Re Life Joint Lab on Public Health and Risk Management, Renmin University of China, Beijing, China.
Email: yang.li@ruc.edu.cn

Abstract

Studies on the conditional relationships between $PM_{2.5}$ concentrations among different regions are of great interest for the joint prevention and control of air pollution. Because of seasonal changes in atmospheric conditions, spatial patterns of $PM_{2.5}$ may differ throughout the year. Additionally, concentration data are both non-negative and non-Gaussian. These data features pose significant challenges to existing methods. This study proposes a heterogeneous graphical model for non-negative and non-Gaussian data via the score matching loss. The proposed method simultaneously clusters multiple datasets and estimates a graph for variables with complex properties in each cluster. Furthermore, our model involves a network that indicate similarity among datasets, and this network can have additional applications. In simulation studies, the proposed method outperforms competing alternatives in both clustering and edge identification. We also analyse the $PM_{2.5}$ concentrations' spatial correlations in Taiwan's regions using data obtained in year 2019 from 67 air-quality monitoring stations. The 12 months are clustered into four groups: January–March, April, May–September and October–December, and the corresponding graphs have 153, 57, 86 and 167 edges respectively. The results show obvious seasonality, which is consistent with the meteorological literature. Geographically, the $PM_{2.5}$ concentrations of north and south Taiwan regions correlate more respectively. These

results can provide valuable information for developing joint air-quality control strategies.

KEYWORDS

generalized *h*-score matching, heterogeneity, penalization, $PM_{2.5}$ data

1 | INTRODUCTION

Air pollution remains a widespread public concern in many countries and regions globally. One particularly harmful air pollutant is ambient fine particulate matter with an aerodynamic diameter of 2.5 μm or less ($PM_{2.5}$), which can seriously endanger human health. According to Cohen et al. (2017), 4.2 million deaths worldwide in the year 2015 can be attributed to $PM_{2.5}$. Some diseases and symptoms are closely linked to $PM_{2.5}$ exposure (Tseng et al., 2019; Wang et al., 2021; Xu et al., 2021; Zhang et al., 2021). Therefore, solving the problem of excessive $PM_{2.5}$ concentration is vital for people's health. Recent studies have focussed on generating historical predictions of $PM_{2.5}$ concentrations with a high spatiotemporal resolution (Wei et al., 2021a, 2021b), supporting insightful analyses of $PM_{2.5}$ over medium- or small-scale areas. Other studies on air pollution meteorology have explored spatial correlations of $PM_{2.5}$ concentrations in different areas. These correlations can provide useful information on spatial patterns of $PM_{2.5}$ and help develop joint prevention and control strategies on atmospheric pollution for more effective air-quality improvement (Wang et al., 2016; Zhang et al., 2018).

This study focusses on spatial correlations of $PM_{2.5}$ concentrations between different regions in Taiwan. The Pearson correlation is a classical tool to measure the relationships between regions with respect to $PM_{2.5}$ concentration (Jin et al., 2017; Zhang et al., 2018). However, this method has limitations. First, the results may be deceptive because of confounders. The left panel of Figure 1 presents the graph of three Taiwanese regions (Jiayi, Shanhua and Tainan) obtained through Pearson correlation. The figure shows an edge between Jiayi and Tainan, but the high Pearson correlation between these two regions may arise from the connection

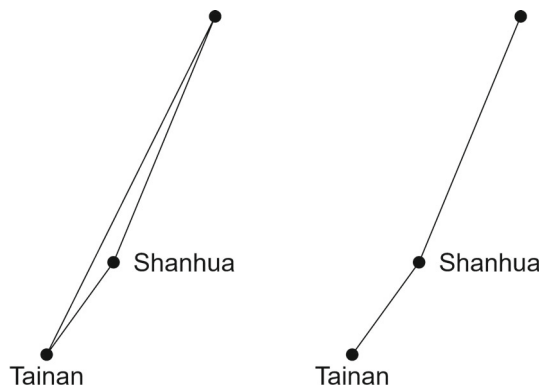

FIGURE 1 Subgraphs in January for Jiayi, Shanhua and Tainan. The left and right panels present the subgraphs of the Pearson correlation and proposed method respectively.

between Jiayi and Shanhua and between Shanhua and Tainan. This is because, since the northeast wind prevails in January (Hsu & Cheng, 2019), Jiayi spreads its $PM_{2.5}$ downstream and indirectly affects Tainan through Shanhua. Second, the Pearson correlation coefficient describes the marginal relationship between two regions' $PM_{2.5}$ concentrations. Due to the pollutants' ability to move through air (Guan et al., 2021; Lv et al., 2015), studies covering a wide geographic area are preferred. For example, Chu et al. (2015) and Wu et al. (2019) conducted an air-quality analysis of Taiwan, and Xie et al. (2018) and Chang et al. (2019) analysed how inter-city transport contributes to $PM_{2.5}$ concentrations in the Yangtze River Delta and Beijing-Tianjin–Hebei region of mainland China. Considering the limitations of the Pearson correlation, we use the graphical model to describe the regions' conditional dependence with respect to $PM_{2.5}$ concentrations. Under this model, the conditional correlation measures the association between two regions, with the other regions' air pollutant levels fixed. This graph analysis strategy can avoid misleading confounders and include all regions of interest from a global perspective.

Graphical models are informative and powerful tools to explore conditional dependence relationships. Specifically, nodes of a graph represent regions, and edges imply their conditional dependencies with respect to $PM_{2.5}$ concentrations. Representative methods of graphical models include neighbourhood selection (Meinshausen & Bühlmann, 2006), graphical lasso (GL) (Friedman et al., 2008), SPACE (Peng et al., 2009) and CONCORD (Khare et al., 2015). Although these approaches are useful, most of them assume that data are homogeneous. They face challenges in the analysis of $PM_{2.5}$ data. Figure 2 presents the histograms of hourly $PM_{2.5}$ concentration records for three regions (Annan, Xindian and Dayuan) in January and June 2019, showing obvious differences in concentration distribution. Moreover, due to changes in temperature, air pressure and atmospheric circulation patterns, the conditional dependence relationships of $PM_{2.5}$

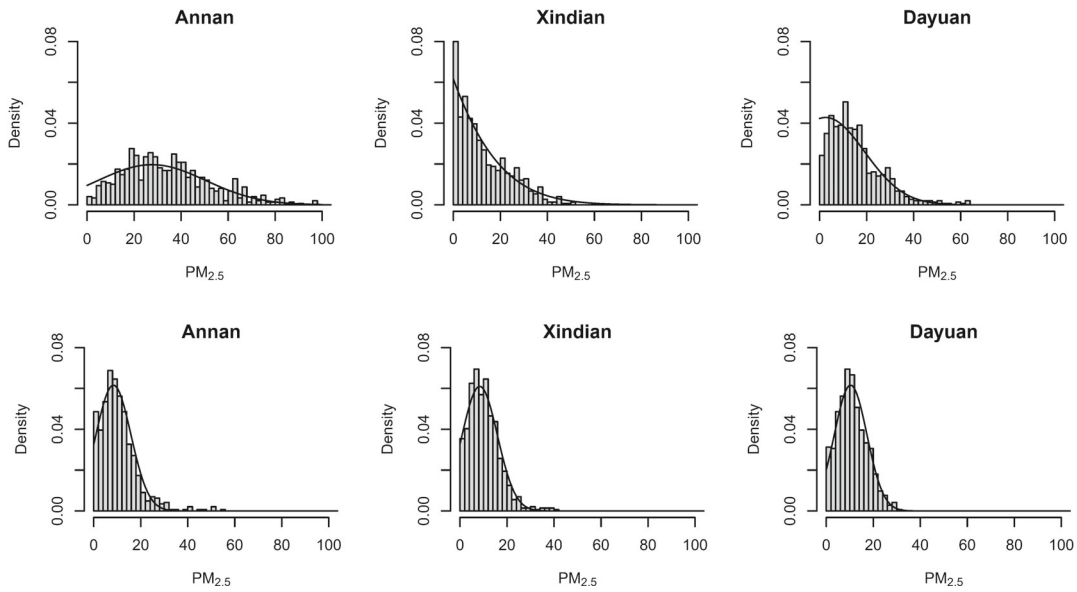

FIGURE 2 Histograms of $PM_{2.5}$ concentration records of three regions in January (upper) and June (lower) in 2019. The solid curves indicate the density functions of a truncated Gaussian distribution with parameters estimated from data.

concentration change from winter to summer. As to be shown in Section 4, the dependence among regions is stronger in winter than in summer. Thus, records may be derived from several distinct subpopulations. There are two primary families of graphical models for heterogeneous data. The first targets the cases in which prior knowledge of class membership is available. Joint estimation methods have been proposed to encourage a common structure (Danaher et al., 2014; Lee & Liu, 2015). However, for many data, population structures are complex and unknown. As such, the second family tackles cases with unknown subpopulation structures. A popular approach is the Gaussian mixture model (Hao et al., 2018), which simultaneously identifies the cluster structure and estimates heterogeneous graphical models. However, determining the number of clusters in mixed models is complicated. Another approach is clustering observations via penalization (Gibberd & Nelson, 2017; Ren et al., 2021), in which the number of clusters is selected with data-driven approaches.

Despite promising results, these heterogeneous graphical models have two limitations. First, most of them assume that the random variables' joint distribution is Gaussian. However, the normality assumption fails when analysing $PM_{2.5}$ data. The concentration records are both non-negative and non-Gaussian, as shown in Figure 2. Second, the observations' structural relationship is not fully utilized. For example, due to seasonal changes in atmospheric conditions (Wu et al., 2019), the conditional dependence relationships of $PM_{2.5}$ concentration tend to be consistent within a certain time period. To overcome these limitations, we propose a heterogeneous graphical model for non-negative data (HGMND) based on score matching. Relevant to this study, Hyvärinen (2005) proposed score matching to eliminate the influence of the multiplicative normalization constant, and this technique was further developed for non-negative data by Hyvärinen (2007) and Yu et al. (2019). This study extends score matching to a heterogeneous graphical model that can simultaneously cluster the months in 2019 and estimate the conditional dependence relationships of $PM_{2.5}$ concentration among 67 regions in Taiwan in each cluster. Our model makes the following contributions. First, our model is extremely flexible, as it accommodates a class of graphical models with distributions supported on the non-negative orthant. Second, the datasets' network structure is fully utilized as it considers more information than the existing studies. Third, the proposed method automatically determines the number of clusters. The analysis results are of great value to understand the spatial and temporal patterns of $PM_{2.5}$ for better air pollution prevention and control.

The remainder of the paper is organized as follows. Section 2 presents the model development and computational algorithm. Simulation studies in Section 3 demonstrate the satisfactory practical performance of the proposed method. The conditional dependence relationships of $PM_{2.5}$ concentration among 67 regions of Taiwan in the 12 months of 2019 are clustered and estimated in Section 4. Section 5 concludes this article with discussions.

2 | MATERIALS AND METHODS

To explore the conditional relationships between $PM_{2.5}$ concentrations of different regions in Taiwan, we collect data on hourly concentration records from 67 air-quality monitoring stations in Taiwan in year 2019. In this section, we first briefly summarize the dataset and split it into multiple datasets according to month. Each dataset is considered homogeneous, while heterogeneity exists across datasets. In each dataset, the records of $PM_{2.5}$ concentration in the 67 regions are

considered as observations of a 67-dimension random vector. A non-negative exponential family is introduced to describe each dataset's population distribution. Then, the conditional relationships among all regions in each dataset are represented by a parameter matrix. We simultaneously cluster the datasets and estimate the parameter matrices by minimizing a generalized h-score matching loss with ℓ_1 and group fused penalties. A common parameter matrix is shared within each cluster. More details are provided in the following subsections.

2.1 | Data summary

Data that support the findings of this study are openly available at https://github.com/zjq-ruc/HGMND. There are 8,760 observations at each station. The $PM_{2.5}$ concentration records at each station are representative of a small region around the station. The data are preprocessed in the following manner. Outliers, whose concentration values are no less than twice the average of the previous and next hours in the same region, are set as missing. The missing values are imputed by the average concentrations of the corresponding region, month and hour. The processed data are used for downstream analysis. Table 1 presents the summary of the data. These $PM_{2.5}$ concentrations show obvious seasonality. The warmer months, April to September, have less $PM_{2.5}$ than the colder months. For example, in January, the monthly average $PM_{2.5}$ concentration is 24.85 $\mu g/m^3$, which is much higher than that in June at 9.76 $\mu g/m^3$. Moreover, 14 days in January have a 24-h average $PM_{2.5}$ concentration of over 25 $\mu g/m^3$, which is a standard of air quality suggested by the World Health Organization.

2.2 | Model

To accommodate heterogeneity across months, we split data into $M = 12$ monthly datasets. In the mth dataset, we assume that the joint density $\Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)})$ of the $PM_{2.5}$ concentration vector $Y^{(m)} = (Y_1^{(m)}, \ldots, Y_p^{(m)})^\top \in \mathbb{R}^p$ is proportional to

$$\exp\left\{-\frac{1}{2a}(y^{(m)})^{a\top}\Theta^{\dagger(m)}(y^{(m)})^a + \eta^{\dagger(m)\top}\frac{(y^{(m)})^b - \mathbf{1}_p}{b}\right\} \mathrm{I}(y^{(m)} \in \mathbb{R}_+^p), \tag{1}$$

where $y^{(m)} = (y_1^{(m)}, \ldots, y_p^{(m)})^\top$ corresponds to a value of $Y^{(m)}$; matrix $\Theta^{\dagger(m)} = (\theta_{lj}^{\dagger(m)}) \in \mathbb{R}^{p\times p}$ and vector $\eta^{\dagger(m)} = (\eta_j^{\dagger(m)}) \in \mathbb{R}^p$ are the true parameters; a, $b > 0$ are known constants; $(y^{(m)})^a = (y_1^{(m)a}, \ldots, y_p^{(m)a})^\top$; $\mathbf{1}_p$ is a p-dimension vector with all elements equal to 1; $\mathrm{I}(\cdot)$ is the indicator function; and $\mathbb{R}_+^p = [0, \infty)^p$ is the non-negative orthant.

Density (1) corresponds to a pairwise graphical model. The conditional dependence relationship between $Y_l^{(m)}$ and $Y_j^{(m)}$ is determined by $\theta_{lj}^{\dagger(m)}$ (Yu et al., 2019). In other words, the $PM_{2.5}$ concentrations of the lth and jth regions, conditional on those of the other regions, are independent if and only if $\theta_{lj}^{\dagger(m)} = 0$. As such, determining the conditional dependence relationships concentrations among the p regions can be formulated as a problem of a sparse estimation of $\Theta^{\dagger(m)}$. When a, $b = 1$, density (1) corresponds to a truncated Gaussian distribution with $Y^{(m)} \sim \mathrm{TN}(\mu^{(m)}, \Sigma^{(m)})$. It is proportional to $\exp\left\{-1/2(y^{(m)} - \mu^{(m)})^\top(\Sigma^{(m)})^{-1}(y^{(m)} - \mu^{(m)})\right\} \mathrm{I}(y^{(m)} \in \mathbb{R}_+^p)$, where $\Sigma^{(m)} = (\Theta^{\dagger(m)})^{-1}$ is positive definite, and $\mu^{(m)} = \Sigma^{(m)}\eta^{\dagger(m)}$. Additionally, when

TABLE 1 Summary of the $PM_{2.5}$ data

	Annual	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct.	Nov.	Dec.
Mean ($\mu g/m^3$)	17.77	24.85	22.02	23.54	20.93	16.42	9.76	10.33	11.64	15.28	21.72	19.36	17.60
SD ($\mu g/m^3$)	12.69	17.50	15.17	15.10	11.63	10.25	6.77	6.82	8.22	9.37	11.83	11.47	10.88
Max ($\mu g/m^3$)	219.00	139.00	219.00	138.00	79.00	101.00	71.00	65.00	85.00	91.00	120.00	129.00	147.00
Days over 25 $\mu g/m^3$ (%)	11.14	45.16	35.71	35.48	20.00	0.00	0.00	0.00	0.00	0.00	25.81	20.00	6.45

$a = 1/2$ and $b = 0$, if we define $(y^b - 1)/b = \log y$ and $\mathbb{R}_+^p = (0, \infty)^p$, the density corresponds to a class of multivariate gamma distributions (Yu et al., 2019). In this study, we set a, $b = 1$ and assume that the $PM_{2.5}$ concentrations follow truncated Gaussian distributions. As shown in Figure 2, the truncated Gaussian distributions' marginally fitted density functions fit the histograms well. The other values of a and b can be considered in follow-up studies.

2.3 | Estimation

We consider the following objective function:

$$\hat{L}(\{\Theta\}, \{\eta\}) = \sum_{m=1}^{M} \hat{J}(\Theta^{(m)}, \eta^{(m)}) + P(\{\Theta\}; E), \tag{2}$$

where $\{\Theta\} = \{\Theta^{(m)}, m = 1, \ldots, M\}$ and $\{\eta\} = \{\eta^{(m)}, m = 1, \ldots, M\}$ are the parameters of the joint density in the form of (1) to approximate the true density. The first part of (2) is the sum of M loss functions, with $\hat{J}(\Theta^{(m)}, \eta^{(m)})$ for the mth dataset. The second term $P(\{\Theta\}; E)$ is a penalty function on $\{\Theta\}$ based on network $\mathcal{G}_M(V, E)$ among the M datasets. Network $\mathcal{G}_M(V, E)$ is set to describe the similarity between the M datasets. In particular, V is the node set, and E is the edge set. Each dataset corresponds to one node in V, and the edge (m_1, m_2) in E indicates that the conditional dependence relationships among variables in the m_1th dataset are close to those in the m_2th dataset. Network $\mathcal{G}_M(V, E)$ is derived based on prior knowledge. As the conditional dependence relationships among the $PM_{2.5}$ concentrations of different regions are similar in adjacent months, $\mathcal{G}_M(V, E)$ is set as a chain network while analysing the $PM_{2.5}$ concentration data.

To define the loss functions, we first introduce notations. We define $h(y^{(m)}) = (h_1(y_1^{(m)}), \ldots, h_p(y_p^{(m)}))^\top$. $h^{1/2}(Y^{(m)})$ is the element-wise square root on vector $h(y^{(m)})$, where $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ are known positive functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$. Let $h_j'(x) = dh_j(x)/dx$ for any $x \in \mathbb{R}_+$. We denote $\mathbf{y}_i^{(m)} = (\mathrm{y}_{i1}^{(m)}, \ldots, \mathrm{y}_{ip}^{(m)})^\top$ as the ith $PM_{2.5}$ concentration records of $Y^{(m)}$, n_m as the number of observations in the mth dataset, and $\mathbf{Y}^{(m)} = (\mathbf{y}_1^{(m)}, \ldots, \mathbf{y}_{n_m}^{(m)})^\top$ as the data matrix. Furthermore, we define $\Phi^{(m)} = (\Theta^{(m)\top}, \eta^{(m)})^\top$. Motivated by the generalized h-score matching loss (Yu et al., 2019), we define the loss function $\hat{J}(\Theta^{(m)}, \eta^{(m)})$ as

$$\hat{J}(\Theta^{(m)}, \eta^{(m)}) = \frac{1}{2}\mathrm{vec}(\Phi^{(m)})^\top \Gamma(\mathbf{Y}^{(m)})\mathrm{vec}(\Phi^{(m)}) - g(\mathbf{Y}^{(m)})^\top \mathrm{vec}(\Phi^{(m)}), \tag{3}$$

where $\Gamma(\mathbf{Y}^{(m)}) \in \mathbb{R}^{(p+1)p \times (p+1)p} = \mathrm{diag}(\Gamma_1^{(m)}, \ldots, \Gamma_p^{(m)})$ is block-diagonal with the jth $(p+1) \times (p+1)$ block

$$\Gamma_j^{(m)} = \frac{1}{n_m} \sum_{i=1}^{n_m} \begin{bmatrix} h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{2a-2}(\mathbf{y}_i^{(m)})^a(\mathbf{y}_i^{(m)})^{a\top} & -h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{a+b-2}(\mathbf{y}_i^{(m)})^a \\ -h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{a+b-2}(\mathbf{y}_i^{(m)})^{a\top} & h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{2b-2} \end{bmatrix},$$

for $j = 1, \ldots, p$. Furthermore, $g(\mathbf{Y}^{(m)}) = \text{vec}((g_1^{(m)\top}, g_2^{(m)\top})^\top) \in \mathbb{R}^{(p+1)p}$, where $g_1^{(m)} \in \mathbb{R}^{p\times p}$ and $g_2^{(m)} \in \mathbb{R}^p$. The jth column of $g_1^{(m)}$, $g_{1,j}^{(m)}$, and jth entry of $g_2^{(m)}$, $g_{2,j}^{(m)}$ are given by

$$
\begin{aligned}
g_{1,j}^{(m)} &= \frac{1}{n_m}\sum_{i=1}^{n_m}\left(h_j'(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{a-1} + (a-1)h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{a-2}\right)(\mathbf{y}_i^{(m)})^a \\
&\quad + a h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{2a-2} e_j, \\
g_{2,j}^{(m)} &= \frac{1}{n_m}\sum_{i=1}^{n_m} - h_j'(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{b-1} - (b-1)h_j(\mathrm{y}_{ij}^{(m)})(\mathrm{y}_{ij}^{(m)})^{b-2},
\end{aligned}
$$

where $e_j \in \mathbb{R}^p$ is a vector with 1 at the jth position and 0 elsewhere. Both $\Gamma(\mathbf{Y}^{(m)})$ and $g(\mathbf{Y}^{(m)})$ are independent of $\Theta^{(m)}$ and $\eta^{(m)}$. Thus, loss $\hat{J}(\Theta^{(m)}, \eta^{(m)})$ is quadratic and convex in $\text{vec}(\Phi^{(m)})$, as $\Gamma(\mathbf{Y}^{(m)})$ is positive semidefinite. The details for deriving $\Gamma(\mathbf{Y}^{(m)})$ and $g(\mathbf{Y}^{(m)})$ are as follows.

Loss function $\hat{J}(\Theta^{(m)}, \eta^{(m)})$ is the empirical version of the generalized h-score matching loss $J(\Theta^{(m)}, \eta^{(m)})$, which takes the form

$$
\begin{aligned}
J(\Theta^{(m)}, \eta^{(m)}) = \int_{\mathbb{R}_+^p} \frac{1}{2} \Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)}) \,\|\, &\nabla \log \Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)}) \circ h^{1/2}(Y^{(m)}) \\
&- \nabla \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)}) \circ h^{1/2}(Y^{(m)})\|_2^2 dy^{(m)},
\end{aligned} \tag{4}
$$

where $\circ$ is the Hadamard product operator, and ∇ is the gradient operator; that is, for any function $f(y^{(m)})$, $\nabla f(y^{(m)}) = (\partial f(y^{(m)})/\partial y_j^{(m)}) \in \mathbb{R}^p$. Loss $J(\Theta^{(m)}, \eta^{(m)})$ compares the gradients of the model log-density $\log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})$ and true log-density $\log \Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)})$ with weight $h^{1/2}(y^{(m)})$. The generalized h-score matching loss is an extension of the score matching loss, which is formed by the expected squared Euclidean distance between the gradients of the model's log-densities and true distribution. Due to discontinuities at the boundary of $\mathbb{R}_+^p$, partial integration underlying the score matching estimator may fail (Hyvärinen, 2007; Yu et al., 2019). To tackle this issue, h functions are introduced. Following Yu et al. (2019), we set $h_j(x) = (\tau x - x^2/2c)\mathrm{I}(0 \le x \le c\tau) + 1/2c\tau^2 \mathrm{I}(x > c\tau)$ for $j = 1, \ldots, p$, where τ and c are two parameters. In numerical study, we set $\tau = 1$ and $c = 5$. Loss $J(\Theta^{(m)}, \eta^{(m)})$ is uniquely minimized when $\Theta^{(m)} = \Theta^{\dagger(m)}$ and $\eta^{(m)} = \eta^{\dagger(m)}$. As the loss depends on only the gradient of $\log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})$, it gets rid of the intractable normalization constant. Detailed procedures for deriving (3) from (4) are presented in Appendix A.

We propose the following penalty:

$$
P(\{\Theta\}; E) = \lambda_1 \sum_{m=1}^{M} |\Theta^{(m)-}| + \lambda_2 \sum_{(m_1,m_2)\in E} \|\Theta^{(m_1)-} - \Theta^{(m_2)-}\|_F,
$$

where $\Theta^{(m)-} = \Theta^{(m)} - \text{diag}(\theta_{11}^{(m)}, \ldots, \theta_{pp}^{(m)})$, $|\Theta^{(m)-}| = \sum_{l\neq j} |\theta_{lj}^{(m)}|$, $\|\cdot\|_F$ represents the Frobenius norm, and λ_1, λ_2 are non-negative tuning parameters. The penalty function comprises two terms. The first term results in sparse estimates for $\Theta^{(m)}$, $m = 1, \ldots, M$. Larger values of λ_1 lead to sparser estimates. In other words, there are more regions whose $PM_{2.5}$ concentrations are conditional independent. The fused group lasso term penalizes the difference between two connected datasets' conditional relationships. The fused group lasso can help borrow strength across the M

datasets when estimating $\Theta^{(m)}$. Furthermore, it can help identify the cluster structure of $\Theta^{(m)}$'s. Specifically, if $\Theta^{(m_1)-} = \Theta^{(m_2)-}$, we conclude that the conditional relationships of $PM_{2.5}$ concentrations in the m_1th month and those in the m_2th month are the same, and that the m_1th and m_2th months belong to the same cluster. Tuning λ_2 controls the degree of heterogeneity among the estimates of $\Theta^{(m)}$'s. The values of $\Theta^{(m)}$ are all different if $\lambda_2 = 0$, while they are the same if $\lambda_2 = \infty$. A similar penalty can be found in Gibberd and Nelson (2017), which is suitable for only the chain network. Other related studies have used the fused lasso, $\lambda_2 \sum_{(m_1,m_2)\in E} |\Theta^{(m_1)-} - \Theta^{(m_2)-}|$, instead of the group fused lasso to cluster datasets (Monti et al., 2014); however, the clustering performance is poor.

Remark 1. We can adapt our objective function to a class of pairwise graphical models with distributions supported on $\mathbb{R}^p$ with minor changes. We only need to replace the generalized h-score matching loss with the original score matching loss. A exact form of the loss can be found in Lin et al. (2016) and Yu et al. (2019).

2.4 | Computation

We minimize objective function (2) with an alternating direction method of multipliers (ADMM) algorithm. We rewrite objective function (2) as

$$\begin{aligned}\hat{L}(\{\Theta\},\{\eta\}) = \frac{1}{2}\sum_{m=1}^{M}\mathrm{vec}(\Phi^{(m)})^\top\Gamma(\mathbf{Y}^{(m)})\mathrm{vec}(\Phi^{(m)}) - g(\mathbf{Y}^{(m)})^\top\mathrm{vec}(\Phi^{(m)}) \\ + \lambda_1\sum_{m=1}^{M}|Z^{(m)-}| + \lambda_2\sum_{(m_1,m_2)\in E}\|Z^{(m_1)-} - Z^{(m_2)-}\|_F, \\ \text{subject to } Z^{(m)} = \Theta^{(m)}, \quad \text{for } m = 1,\ldots,M.\end{aligned}$$

The scaled augmented Lagrangian for this objective function is

$$\begin{aligned}Q(\{\Theta\},\{\eta\},\{Z\},\{U\}) = \frac{1}{2}\sum_{m=1}^{M}\mathrm{vec}(\Phi^{(m)})^\top\Gamma(\mathbf{Y}^{(m)})\mathrm{vec}(\Phi^{(m)}) - g(\mathbf{Y}^{(m)})^\top\mathrm{vec}(\Phi^{(m)}) \\ + \lambda_1\sum_{m=1}^{M}|Z^{(m)-}| + \lambda_2\sum_{(m_1,m_2)\in E}\|Z^{(m_1)-} - Z^{(m_2)-}\|_F \\ + \frac{\rho}{2}\sum_{m=1}^{M}\left(\left\|\Theta^{(m)} - Z^{(m)} + U^{(m)}\right\|_F^2 - \left\|U^{(m)}\right\|_F^2\right),\end{aligned}$$

where $\{Z\} = \{Z^{(m)}, m = 1,\ldots,M\}$ and $\{U\} = \{U^{(m)}, m = 1,\ldots,M\}$. $\{U\}$ indicates dual variables, and ρ serves as a penalty parameter. In this study, we set ρ to 1, which leads to reasonable performance and fast convergence. Denote $\{A_{(t)}\} = \{A_{(t)}^{(m)}, m = 1,\ldots,M\}$ for matrices (or vectors) $A_{(t)}^{(m)}$'s at the tth iteration. Following the approach of Boyd et al. (2011), we consider two convergence criteria: $r_1 = \sum_{m=1}^{M}\|\Theta_{(t)}^{(m)} - Z_{(t)}^{(m)}\|_F^2 < \epsilon_1 = 10^{-3}$ and $r_2 = \sum_{m=1}^{M}\|Z_{(t)}^{(m)} - Z_{(t-1)}^{(m)}\|_F^2 < \epsilon_2 = 10^{-3}$. The proposed algorithm is summarized in Algorithm 1. Details on updating $\{\Theta_{(t)}\}$, $\{\eta_{(t)}\}$, and $\{Z_{(t)}\}$ are given in Appendix B.

Algorithm 1: Outline of the ADMM algorithm for HGMND

Input: $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(M)}$, λ_1, λ_2.

Output: HGMND estimates $\hat{\Theta}^{(m)}$, $m = 1, \ldots, M$.

Initialize: $\Theta_{(0)}^{(m)} = Z_{(0)}^{(m)} = U_{(0)}^{(m)} = \mathbf{0}$, a $p \times p$ matrix with all zeroes, $m = 1, \ldots, M$, and $t = 0$.

while not convergent **do**

  Let $t = t + 1$,

  Update $\{\Theta_{(t)}\}$ and $\{\eta_{(t)}\}$:

$$(\{\Theta_{(t)}\}, \{\eta_{(t)}\}) = \underset{\{\Theta\},\{\eta\}}{\arg\min}\, Q\Big(\{\Theta\}, \{\eta\}, \{Z_{(t-1)}\}, \{U_{(t-1)}\}\Big); \tag{5}$$

  Update $\{Z_{(t)}\}$:

$$\{Z_{(t)}\} = \underset{\{Z\}}{\arg\min}\, Q\Big(\{\Theta_{(t)}\}, \{\eta_{(t)}\}, \{Z\}, \{U_{t-1}\}\Big); \tag{6}$$

  Update $\{U_{(t)}\}$: **for** $m = 1, \ldots, M$ **do**

    $U_{(t)}^{(m)} = U_{(t-1)}^{(m)} + \Theta_{(l)}^{(m)} - Z_{(t)}^{(m)}$.

  end

end

Following Monti et al. (2014), we select the tuning parameters with an AIC-type criterion:

$$\begin{aligned}\mathrm{AIC}(\lambda_1, \lambda_2) = &\sum_{m=1}^{M} \mathrm{vec}(\hat{\Phi}^{(m)})^{\top} \Gamma(\mathbf{Y}^{(m)}) \mathrm{vec}(\hat{\Phi}^{(m)}) - g(\mathbf{Y}^{(m)})^{\top} \mathrm{vec}(\hat{\Phi}^{(m)}) \\ &+ \sum_{l \neq j} \sum_{\substack{(m_1, m_2) \in E \\ m_1 < m_2}} \mathrm{I}(\hat{\theta}_{lj}^{(m_1)} \neq \hat{\theta}_{lj}^{(m_2)}) \cdot \mathrm{I}(\hat{\theta}_{lj}^{(m_2)} \neq 0),\end{aligned}$$

where $\hat{\Phi}^{(m)}$, $\hat{\theta}_{lj}^{(m_1)}$, and $\hat{\theta}_{lj}^{(m_2)}$ denote the estimated parameters. We conduct a grid search to select λ_1 and λ_2. It takes <2 min to run each optimization on a 5×5 searching grid under the setting of truncated Gaussian distribution, with $n_m = 100$, $p = 50$, $M = 20$ and $K = 2$, using a six-core 2.2GHz CPU using R.

3 | EVALUATION WITH SYNTHETIC DATA

Simulation studies are conducted to assess the proposed method. We set M datasets, which belong to K clusters, and randomly generate n_m observations of $Y^{(m)}$ from Equation (1). To mimic

seasonal changes in the conditional dependence relationships of $PM_{2.5}$ concentrations, we consider a chain network. We suppose that there exist $K-1$ changepoints (cp) $\{\tau_1, \ldots, \tau_{K-1}\} \subset \{1, \ldots, M\}$ in the chain network. For each $k \in \{1, \ldots, K-1\}$, $\Theta^{\dagger(m_1)} = \Theta^{\dagger(m_2)}$ if $\tau_{k-1} < m_1, m_2 \leq \tau_k$, where $\tau_0 = 0$. Other networks can also be considered, and more examples are provided in Appendix C. $\Theta^{\dagger(m)}$'s in each cluster are the same and generated as follows. We first generate K graphs with the p features belonging to $p/10$ equally sized unconnected subgraphs. An edge between two nodes in each subgraph is generated with a probability of 0.1. For a given graph structure, the elements in $\Theta^{\dagger(m)}$ not corresponding to the edges are set to zero, and those that correspond to the edges are generated randomly from a uniform distribution of $[-1,-0.5] \cup [0.5,1]$. To ensure positive definiteness, we set the diagonal entries as $\theta_{jj}^{\dagger(m)} = 0.1 + \sum_{l \neq j} |\theta_{lj}^{\dagger(m)}|$ for $j = 1, \ldots, p$. Next, we scale $\Theta^{\dagger(m)}$ as follows. Let $\Sigma^{(m)} = \Theta^{\dagger(m)-1}$, and $\Sigma^{*(m)}$ is the correlation coefficient matrix corresponding to $\Sigma^{(m)}$. Then, we set $\Theta^{\dagger(m)} = \Sigma^{*(m)-1}$. For each dataset, $\eta^{\dagger(m)}$ is generated from a multivariate Gaussian distribution. As shown in Figure 2, the truncated Gaussian densities fit the histograms of $PM_{2.5}$ concentrations well. Thus, we first set $a = 1$ and $b = 1$. In addition to the truncated Gaussian densities, the multivariate gamma distribution may also be appropriate to describe the $PM_{2.5}$ data; thus, we also consider $a = 1/2$ and $b = 0$. For data matrix $\boldsymbol{Y}^{(m)} = (y_{ij}^{(m)}) \in \mathbb{R}^{n_m \times p}$, we scale each element of the jth column by $\sqrt{\sum_{i=1}^{n_m} (y_{ij}^{(m)})^2/(n_m - 1)}$. Data are generated with package `genscore` in R.

Four measures are calculated to assess performance. The Rand index (RI) is used to measure the accuracy of clustering, while the F_1 score, true positive rate (TPR) and false positive rate (FPR) are used to measure the sparsity recovery performance: $F_1 = M^{-1}\sum_{m=1}^{M} 2TP^{(m)}/(2TP^{(m)} + FN^{(m)} + FP^{(m)})$, $TPR = M^{-1}\sum_{m=1}^{M} TP^{(m)}/(TP^{(m)} + FN^{(m)})$, and $FPR = M^{-1}\sum_{m=1}^{M} FP^{(m)}/(TN^{(m)} + FP^{(m)})$, where $TP^{(m)}$, $TN^{(m)}$, $FP^{(m)}$ and $FN^{(m)}$ are the numbers of true positives, true negatives, false positives and false negatives, respectively, for dataset m. We compare the proposed method with five existing methods: graphical lasso (GL), independent fused graphical lasso (IFGL), group fused graphical lasso (GFGL), rank-based group fused graphical lasso (rGFGL), and logarithm-based group fused graphical lasso (lGFGL). GL (Friedman et al., 2008) separately analyses the M datasets. IFGL (Monti et al., 2014) and GFGL (Gibberd & Nelson, 2017) jointly estimate the interaction matrices with different penalties. These three methods make a Gaussian assumption and adopt the negative log-likelihood as the loss. The last two methods, rGFGL and lGFGL, are almost the same as GFGL except for the construction of sample covariance matrices. rGFGL uses a rank-based method (Xue & Zou, 2012), while lGFGL takes a logarithm of data before estimation. All numerical results are based on 100 replicates.

Then, we set $M = 10, 15, 20, 30$, $K = 2, 3$, $p = 20, 50$, $n_m = 30, 50, 100$, and cp = {5}, {10} {10, 20}, {15, 20}, {5, 10}. Detailed combinations of M, K, p, n_m, and cp are shown in Tables 2 and 3. We set $\eta^{\dagger(m)} \sim \mathcal{N}(\mathbf{0}_p, 0.1^2 I_p)$ for the truncated multivariate normal distributions and $\eta^{\dagger(m)} \sim \mathcal{N}(\mathbf{2}_p, 0.1^2 I_p)$ for the multivariate gamma distributions, where $\mathbf{0}_p$ and $\mathbf{2}_p$ are two p-dimension vectors with all 0s and 2s respectively. I_p is the $p \times p$ identity matrix. Results are provided in Tables 2 and 3. The observed patterns for the two distribution types are very similar. HGMND, GFGL, rGFGL and lGFGL, which use a group fused term to jointly analyse the M datasets, have better performance on RI than IFGL and GL. In terms of the sparsity recovery performance, HGMND performs better than the other five methods with higher F_1 score in most settings. For example, when $M = 10, K = 2$, cp = {5}, $n_m = 50, p = 20$, HGMND's F_1 score value is 0.588, which is larger than 0.546, 0.557, 0.493, 0.440 and 0.270 for GFGL, rGFGL, lGFGL, IFGL and GL respectively. As expected, the sparsity recovery performance improves as n_m increases and p decreases.

TABLE 2 Comparison under non-centred truncated Gaussian distributions with a chain network. For each setting, the best F_1, TPR, FPR and RI are in boldface. Numbers in parentheses are standard errors

Method	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI
	$n_m = 50, p = 20$				$n_m = 50, p = 20$				$n_m = 50, p = 20$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{5\}$				$M = 10, K = 2$, cp $= \{5\}$			
HGMND	**0.767**	**0.833**	**0.018**	0.988	**0.769**	**0.846**	0.022	**0.964**	**0.588**	0.643	**0.028**	**0.997**
	(0.057)	(0.077)	(0.006)	(0.022)	(0.051)	(0.074)	(0.008)	(0.067)	(0.061)	(0.104)	(0.008)	(0.015)
GFGL	0.700	0.780	0.023	**0.999**	0.767	0.829	**0.019**	0.963	0.546	0.644	0.037	0.990
	(0.064)	(0.105)	(0.006)	(0.007)	(0.053)	(0.069)	(0.007)	(0.075)	(0.065)	(0.116)	(0.011)	(0.062)
rGFGL	0.651	0.806	0.035	**0.999**	0.696	0.818	0.029	0.937	0.557	0.629	0.033	0.990
	(0.057)	(0.094)	(0.009)	(0.007)	(0.061)	(0.073)	(0.011)	(0.103)	(0.069)	(0.106)	(0.011)	(0.036)
lGFGL	0.575	0.770	0.049	0.992	0.555	0.805	0.059	0.856	0.493	0.503	**0.028**	0.940
	(0.069)	(0.097)	(0.013)	(0.019)	(0.070)	(0.062)	(0.017)	(0.128)	(0.070)	(0.102)	(0.010)	(0.113)
IFGL	0.563	0.802	0.055	0.962	0.658	0.803	0.033	0.937	0.440	**0.668**	0.073	0.965
	(0.043)	(0.076)	(0.013)	(0.040)	(0.054)	(0.058)	(0.012)	(0.084)	(0.054)	(0.103)	(0.020)	(0.089)
GL	0.184	0.750	0.349	0.526	0.182	0.762	0.353	0.395	0.161	0.645	0.352	0.556
	(0.010)	(0.034)	(0.023)	(0.000)	(0.009)	(0.031)	(0.023)	(0.000)	(0.013)	(0.056)	(0.022)	(0.000)
	$n_m = 50, p = 20$				$n_m = 100, p = 20$				$n_m = 50, p = 20$			
	$M = 30, K = 3$, cp $= \{10, 20\}$				$M = 30, K = 3$, cp $= \{15, 20\}$				$M = 15, K = 3$, cp $= \{5, 10\}$			
HGMND	**0.684**	**0.854**	**0.027**	**0.997**	**0.840**	0.873	**0.012**	0.995	**0.533**	**0.817**	**0.053**	**0.971**
	(0.045)	(0.057)	(0.005)	(0.008)	(0.036)	(0.051)	(0.002)	(0.014)	(0.045)	(0.075)	(0.010)	(0.039)
GFGL	0.585	0.822	0.040	0.983	0.829	**0.875**	0.013	**0.998**	0.486	0.787	0.060	0.958
	(0.048)	(0.097)	(0.008)	(0.039)	(0.029)	(0.036)	(0.002)	(0.011)	(0.054)	(0.093)	(0.021)	(0.051)
rGFGL	0.584	0.771	0.035	0.985	0.649	0.792	0.025	0.988	0.458	0.688	0.054	0.951
	(0.053)	(0.090)	(0.005)	(0.029)	(0.057)	(0.075)	(0.005)	(0.033)	(0.046)	(0.089)	(0.015)	(0.067)
lGFGL	0.468	0.765	0.062	0.959	0.474	0.765	0.053	0.986	0.350	0.679	0.093	0.922
	(0.042)	(0.102)	(0.015)	(0.063)	(0.048)	(0.094)	(0.016)	(0.029)	(0.049)	(0.097)	(0.022)	(0.080)
IFGL	0.446	0.699	0.059	0.964	0.611	0.705	0.020	0.963	0.362	0.708	0.097	0.938
	(0.032)	(0.085)	(0.010)	(0.025)	(0.060)	(0.054)	(0.010)	(0.029)	(0.061)	(0.110)	(0.042)	(0.034)
GL	0.167	0.827	0.358	0.690	0.210	0.851	0.264	0.632	0.255	0.689	0.171	0.714
	(0.008)	(0.026)	(0.021)	(0.000)	(0.014)	(0.025)	(0.021)	(0.000)	(0.022)	(0.043)	(0.021)	(0.000)

TABLE 2 (Continued)

Method	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI
	$n_m = 30, p = 20$				$n_m = 50, p = 50$				$n_m = 100, p = 50$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{10\}$			
HGMND	0.593	0.749	0.042	**0.988**	**0.603**	0.638	**0.009**	0.988	**0.757**	**0.894**	**0.009**	**1.000**
	(0.056)	(0.105)	(0.015)	(0.023)	(0.039)	(0.055)	(0.001)	(0.022)	(0.024)	(0.034)	(0.001)	(0.005)
GFGL	0.591	**0.790**	0.048	0.985	0.569	0.607	0.010	**1.000**	0.696	0.835	0.011	**1.000**
	(0.058)	(0.096)	(0.018)	(0.027)	(0.037)	(0.061)	(0.002)	(0.000)	(0.029)	(0.056)	(0.001)	(0.000)
rGFGL	**0.599**	0.694	**0.033**	0.982	0.559	0.589	0.010	0.999	0.682	0.881	0.014	1.000
	(0.056)	(0.100)	(0.012)	(0.029)	(0.036)	(0.054)	(0.001)	(0.007)	(0.023)	(0.043)	(0.002)	(0.000)
lGFGL	0.434	0.686	0.078	0.983	0.465	**0.672**	0.024	0.967	0.572	0.816	0.021	1.000
	(0.052)	(0.094)	(0.018)	(0.034)	(0.031)	(0.049)	(0.004)	(0.042)	(0.030)	(0.047)	(0.003)	(0.005)
IFGL	0.430	0.757	0.095	0.967	0.509	0.642	0.017	0.930	0.572	0.810	0.021	0.937
	(0.049)	(0.091)	(0.024)	(0.068)	(0.031)	(0.062)	(0.004)	(0.046)	(0.040)	(0.072)	(0.007)	(0.038)
GL	0.138	0.734	0.489	0.526	0.149	0.499	0.105	0.526	0.153	0.753	0.164	0.526
	(0.007)	(0.035)	(0.034)	(0.000)	(0.007)	(0.020)	(0.004)	(0.000)	(0.004)	(0.017)	(0.005)	(0.000)

TABLE 3 Comparison under non-centred gamma distributions with a chain network. For each setting, the best F_1, TPR, FPR and RI are in boldface. Numbers in parentheses are standard errors

Method	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI
	$n_m = 50, p = 20$				$n_m = 50, p = 20$				$n_m = 50, p = 20$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{5\}$				$M = 10, K = 2$, cp $= \{5\}$			
HGMND	**0.793**	0.860	**0.016**	0.998	**0.789**	0.818	**0.014**	**0.966**	0.636	0.842	0.044	0.973
	(0.072)	(0.133)	(0.005)	(0.009)	(0.052)	(0.079)	(0.004)	(0.075)	(0.071)	(0.093)	(0.018)	(0.041)
GFGL	0.732	0.837	0.024	0.998	0.781	0.813	0.015	0.953	**0.637**	0.830	**0.042**	0.970
	(0.067)	(0.119)	(0.007)	(0.010)	(0.043)	(0.068)	(0.004)	(0.097)	(0.059)	(0.101)	(0.015)	(0.047)
rGFGL	0.789	**0.869**	0.018	0.998	0.779	0.799	**0.014**	0.943	0.612	**0.891**	0.056	**0.984**
	(0.057)	(0.088)	(0.005)	(0.009)	(0.054)	(0.082)	(0.004)	(0.124)	(0.058)	(0.078)	(0.018)	(0.037)
lGFGL	0.666	0.776	0.029	**0.999**	0.788	**0.821**	0.015	0.930	0.586	0.841	0.056	0.946
	(0.068)	(0.133)	(0.008)	(0.011)	(0.047)	(0.078)	(0.005)	(0.108)	(0.059)	(0.089)	(0.021)	(0.054)
IFGL	0.604	0.831	0.048	0.959	0.691	0.820	0.029	0.876	0.532	0.784	0.063	0.831
	(0.038)	(0.074)	(0.009)	(0.116)	(0.047)	(0.042)	(0.009)	(0.101)	(0.061)	(0.096)	(0.024)	(0.083)
GL	0.232	0.728	0.269	0.526	0.324	0.602	0.109	0.395	0.252	0.699	0.215	0.556
	(0.017)	(0.043)	(0.032)	(0.000)	(0.019)	(0.040)	(0.008)	(0.000)	(0.022)	(0.054)	(0.028)	(0.000)
	$n_m = 50, p = 20$				$n_m = 100, p = 20$				$n_m = 50, p = 20$			
	$M = 30, K = 3$, cp $= \{10, 20\}$				$M = 30, K = 3$, cp $= \{5, 20\}$				$M = 15, K = 3$, cp $= \{5, 10\}$			
HGMND	**0.804**	**0.938**	**0.018**	**1.000**	**0.841**	**0.952**	**0.020**	0.982	**0.548**	0.750	**0.040**	**0.980**
	(0.041)	(0.028)	(0.005)	(0.003)	(0.039)	(0.057)	(0.004)	(0.030)	(0.051)	(0.098)	(0.013)	(0.026)
GFGL	0.669	0.843	0.028	0.987	0.692	0.844	0.025	0.989	0.519	**0.790**	0.050	0.945
	(0.051)	(0.087)	(0.008)	(0.020)	(0.053)	(0.070)	(0.006)	(0.019)	(0.041)	(0.089)	(0.014)	(0.073)
rGFGL	0.707	0.933	0.031	0.993	0.726	0.931	0.028	**0.992**	0.514	0.700	0.041	0.954
	(0.034)	(0.017)	(0.005)	(0.013)	(0.034)	(0.016)	(0.005)	(0.015)	(0.043)	(0.091)	(0.008)	(0.087)
lGFGL	0.647	0.922	0.041	0.977	0.696	0.882	0.028	0.973	0.511	0.709	0.041	0.953
	(0.058)	(0.052)	(0.014)	(0.025)	(0.050)	(0.079)	(0.010)	(0.024)	(0.047)	(0.088)	(0.014)	(0.064)
IFGL	0.406	0.819	0.098	0.967	0.471	0.815	0.066	0.946	0.374	0.775	0.096	0.927
	(0.065)	(0.123)	(0.047)	(0.033)	(0.117)	(0.070)	(0.042)	(0.035)	(0.038)	(0.079)	(0.017)	(0.036)
GL	0.318	0.671	0.106	0.690	0.298	0.671	0.107	0.632	0.319	0.675	0.107	0.714
	(0.016)	(0.036)	(0.007)	(0.000)	(0.017)	(0.036)	(0.007)	(0.000)	(0.025)	(0.051)	(0.011)	(0.000)

TABLE 3 (Continued)

Method	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI
	$n_m = 30, p = 20$				$n_m = 50, p = 50$				$n_m = 100, p = 50$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{10\}$				$M = 20, K = 2$, cp $= \{10\}$			
HGMND	0.632	**0.853**	0.045	0.986	**0.653**	0.774	0.012	**1.000**	**0.793**	0.907	**0.008**	0.998
	(0.067)	(0.105)	(0.016)	(0.027)	(0.025)	(0.054)	(0.002)	(0.000)	(0.022)	(0.026)	(0.001)	(0.012)
GFGL	**0.637**	0.850	0.044	0.983	0.607	**0.804**	0.017	0.996	0.685	**0.913**	0.015	**1.000**
	(0.047)	(0.098)	(0.014)	(0.033)	(0.030)	(0.057)	(0.002)	(0.014)	(0.021)	(0.043)	(0.002)	(0.000)
rGFGL	0.628	0.806	**0.040**	**0.999**	0.648	0.750	**0.011**	**1.000**	0.772	0.910	0.009	0.996
	(0.054)	(0.106)	(0.011)	(0.008)	(0.032)	(0.057)	(0.001)	(0.000)	(0.020)	(0.022)	(0.001)	(0.014)
lGFGL	0.523	0.841	0.074	0.993	0.569	0.768	0.019	0.998	0.668	0.865	0.015	**1.000**
	(0.055)	(0.098)	(0.021)	(0.023)	(0.034)	(0.057)	(0.003)	(0.014)	(0.022)	(0.048)	(0.001)	(0.000)
IFGL	0.541	0.759	0.055	0.909	0.444	0.807	0.041	0.953	0.567	0.906	0.026	0.894
	(0.054)	(0.118)	(0.019)	(0.066)	(0.088)	(0.099)	(0.020)	(0.038)	(0.022)	(0.036)	(0.002)	(0.211)
GL	0.178	0.653	0.323	0.526	0.174	0.619	0.121	0.526	0.203	0.805	0.140	0.526
	(0.013)	(0.042)	(0.037)	(0.000)	(0.009)	(0.023)	(0.008)	(0.000)	(0.012)	(0.019)	(0.011)	(0.000)

4 | $PM_{2.5}$ DATA ANALYSIS

In this section, we first cluster the 12 months and estimate a graph for each cluster with the proposed method. Then, we analyse the results. Finally, we compare the proposed and alternative methods.

4.1 | Graph clustering and estimation

Motivated by the exploratory analysis, we apply the proposed method based on a truncated multivariate Gaussian assumption and a chain network. To better illustrate our results, we add a map of Taiwan from http://www.tianditu.gov.cn in the background of our estimated graphs. Figure 3 presents the clustering of months and graphs illustrating the conditional dependence relationships across the 67 regions. Looking back on the right panel of Figure 1, $PM_{2.5}$ concentrations in Tainan and Jiayi are independent given the pollution levels of the other regions, whereas those of Tainan and Shanhua as well as Shanhua and Jiayi are conditionally dependent. This result illustrates the proposed method's ability to avoid misleading confounders, as is mentioned in Introduction. Besides, the graphs show obvious seasonality. For example, the graphs for June to September are highly different from those for October to December. Figure 4 shows the number of edges selected in each graph and node degree distributions in January and June. We find that the estimated graphs for the colder days have more edges. This seasonality has been studied in the literature. As Wu et al. (2019) indicated, air quality in Taiwan is poorer during the cold dry season, which corroborates our results. The relationship between the number of edges and average daily precipitation per week is also examined in Figure 4. The figure shows that clusters of the months correspond to the changes in average daily precipitation per week. Thus,

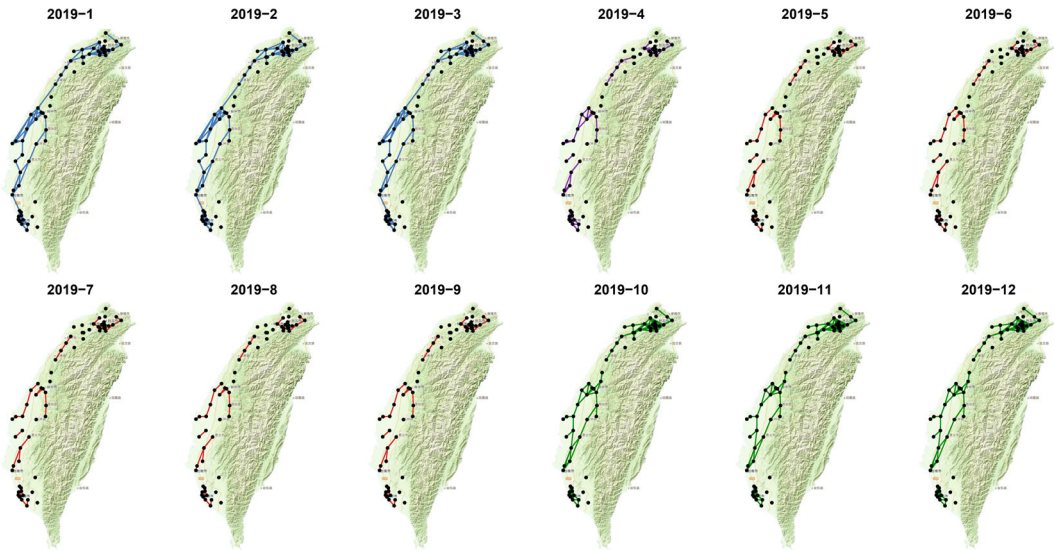

FIGURE 3 Monthly estimated graphs for the 67 regions in 2019, where months in the same cluster share the same colour of edges, while graphs with different edge colours mean that the corresponding months are in different clusters. [Colour figure can be viewed at wileyonlinelibrary.com]

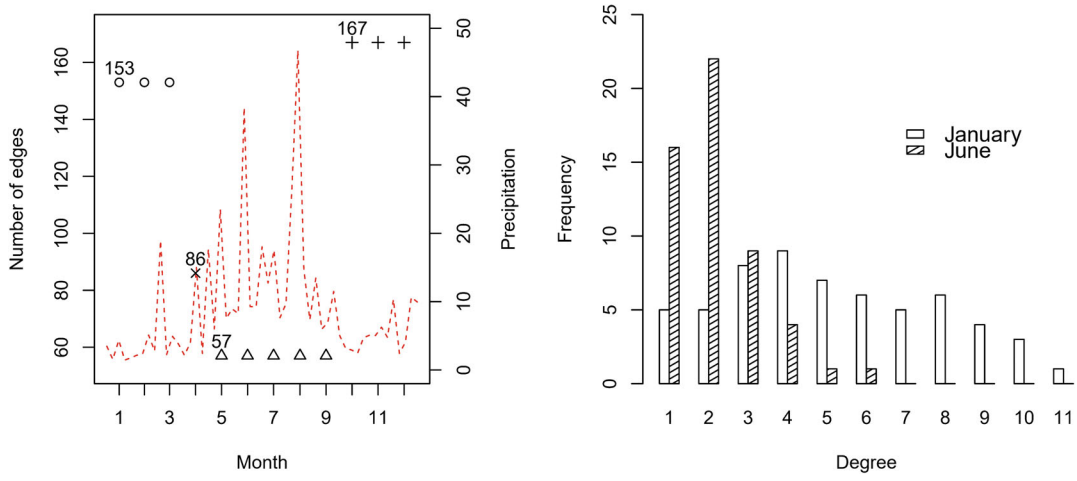

FIGURE 4 Number of estimated edges for each month and node degree distribution. Left panel: The points denote the number of selected edges per month, with one type of shape denoting one cluster. The red dotted line represents the average daily precipitation per week. Right panel: Frequency of degrees for the nodes of January and June. [Colour figure can be viewed at wileyonlinelibrary.com]

the heterogeneous structure of $PM_{2.5}$ concentrations can be highly attributed to the changes in climatic conditions.

From a geographical perspective, on the colder days, the northern and southern areas are more connected, likely because air movement can influence pollution levels in neighbouring cities (Lv et al., 2015). Figure 5 presents the estimated graph for January 2019, with the 67 regions divided into five districts by the environmental protection administration that monitors air quality. The right panel shows that the numbers of estimated edges within the same district are mostly larger than those between districts, which is the same in other months. This implies that the proposed method can identify urban agglomerations in $PM_{2.5}$ analysis. Moreover, the northern area—districts 1 and 2—are rarely connected to the southern area—districts 3–5. This result is consistent with Wu et al. (2019), as the atmospheric condition in southern Taiwan differs from that in the north.

Therefore, based on these results, we suggest that Taiwan should put more effort into formulating joint prevention and control strategies for $PM_{2.5}$ on cold and dry days, in the period from January to March and October to December. Meanwhile, more efforts should be made to comprehensively consider groups of closely connected regions, which may have higher consistency in pollution emissions owing to similarities in pollution source types, emission levels and other factors. For example, in January, in the coastal industrial complex of southern Kaohsiung City, regions near the stations of Fengshan, Fuxing, Qianjin, Qianzhen, Renwu, Xiaogang and Zuoying are highly connected with node degrees of 8 to 10, and each of them also has an annual $PM_{2.5}$ concentration over 20 $\mu g/m^3$. Air pollution prevention and control strategies for industrial production activities can be implemented jointly in these regions for more effective air-quality improvement.

4.2 | Comparison with alternative methods

GFGL, rGFGL, lGFGL and two naive methods are adopted for comparison. The first naive method connects the closest 10% pairs of regions with respect to the geographical distance. The second

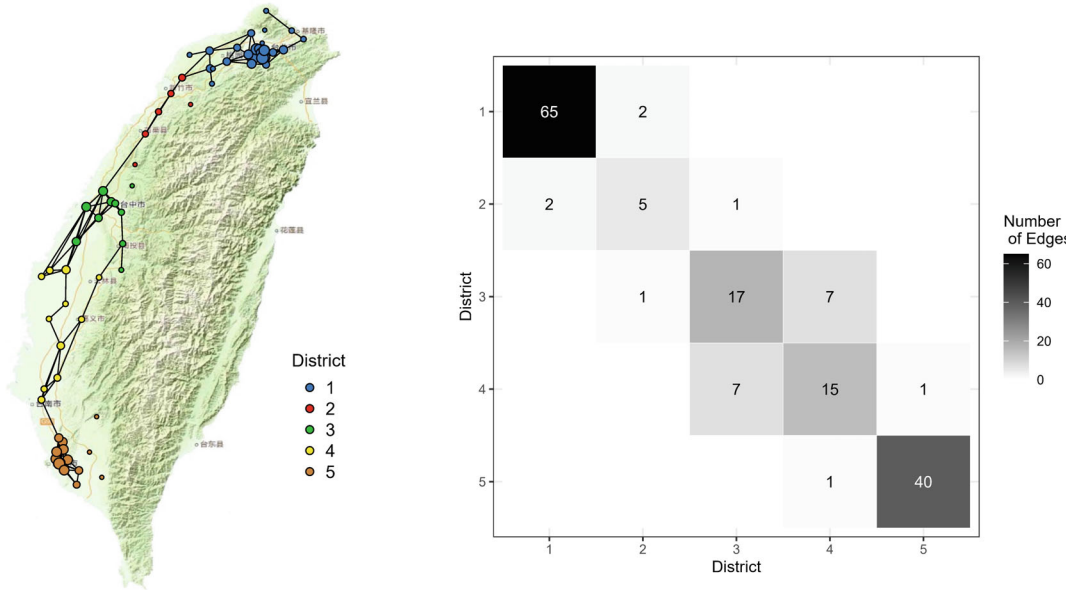

FIGURE 5 Left panel: Estimated graph for January 2019. Nodes in different districts have different colours, and nodes with larger degrees have larger sizes. Right panel: Number of edges between districts. The diagonal elements indicate the number of estimated edges within each district. [Colour figure can be viewed at wileyonlinelibrary.com]

naive method obtains the graph for each month by conducting hard thresholding on the sample Pearson correlation matrices (Bickel & Levina, 2008).

The two naive methods yield poor results (Figure 6). The first method obtains the same graph for all 12 months and does not use the $PM_{2.5}$ data at all. For the second method, as mentioned in Section 1, Pearson correlation reveals the marginal relationship between the two variables. The thresholding method leads to very dense graphs with almost all pairs of regions connected. To further control the sparsity level of the graphs induced by Pearson correlations, we attempt to remove the edges corresponding to small absolute values in the sample correlation matrices and ensure that the graphs have the same numbers of edges as those obtained by the proposed method. However, there still exist some edges connecting regions far apart that are difficult to explain, such as the graph for July. This is possibly because, in July, under the influence of easterly winds and the Central Mountain Range, which extends from the northeast to the southwest of the island, local circulation is produced in the southwestern area and its coastal waters. Thus, Pearson correlations of the southwestern region may be confounded. However, only adjacent regions are connected by the proposed method, showing the method's ability to restrict confounders when studying conditional relationships. Additionally, the proposed method's results reveal the seasonality of the $PM_{2.5}$ data with much higher interpretability, while the two naive methods do not model the clustering structure of the $PM_{2.5}$ data.

It is difficult to assess clustering and edge identification accuracy in real data analysis. We use a random sampling approach to compare the proposed method with the remaining three methods, GFGL, rGFGL and lGFGL. To measure stability, we sample 60% of the observations from each of the 12 datasets and compare the corresponding results with those based on the complete datasets. Specifically, we compute the RI and probability that each identified edge is identified

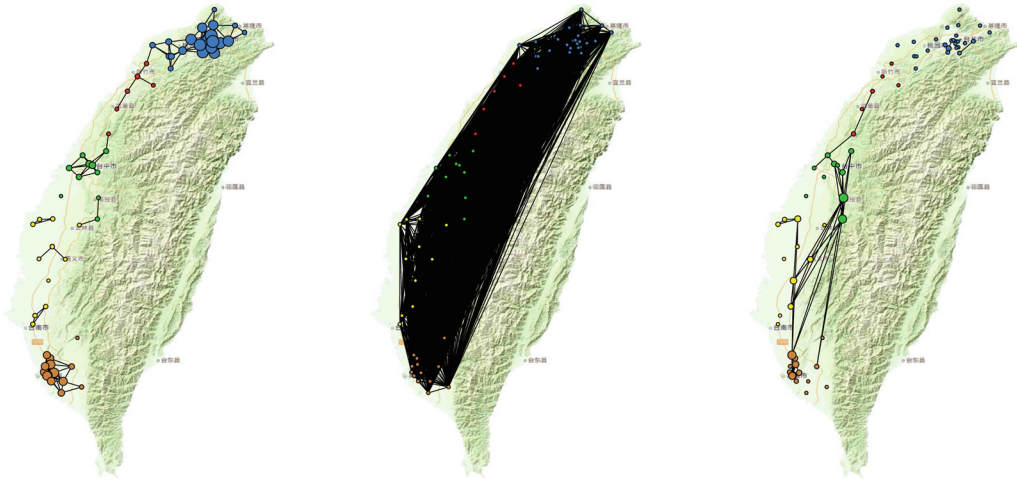

FIGURE 6 Results of alternative methods. Left panel: Graph based on geographical distance. Middle panel: Graph for January with the hard-thresholding method. Right panel: Graph for July after removing edges that correspond to small absolute values in the sample correlation matrix. [Colour figure can be viewed at wileyonlinelibrary.com]

again. According to the literature (Huang & Ma, 2010), this can be a measure of stability. This procedure is repeated 100 times. The average values of RI for HGMND, GFGL, rGFGL and lGFGL are 0.96, 0.89, 0.94 and 0.93 respectively. HGMND yields more stable clustering results. Furthermore, HGMND is much more stable for selecting edges with the average values of identified edge probabilities equal to 0.94, compared with 0.86, 0.86 and 0.78 for GFGL, rGFGL and lGFGL respectively.

5 | CONCLUSIONS

Research on the spatial pattern of $PM_{2.5}$ is critical for socioeconomic and epidemiological studies (Huang et al., 2019; Su et al., 2020). In Taiwan, $PM_{2.5}$ is one of the most serious air pollutants that causes significant health risks for residents. Therefore, Taiwan's air quality needs to be improved urgently. Accordingly, determining the conditional relationships among $PM_{2.5}$ concentrations across regions is important for joint air pollution prevention and control. However, since $PM_{2.5}$ concentrations are non-negative, non-normal and heterogeneous, classical methods are not suitable. To tackle this issue, we extend the generalized h-score matching loss based on group fused lasso to simultaneously cluster datasets and estimate graphs of variables with complex distributions. Furthermore, we consider the network among datasets, which adds additional information to the model. Simulations show that the proposed approach has satisfactory performance. Finally, the proposed method is used to cluster the months of 2019 and estimate the conditional dependence relationships of $PM_{2.5}$ concentration records among 67 regions in Taiwan in each cluster. We find that the graphs have significant seasonality, and various regions have closer links in winter than in summer. Geographically, north and southwest Taiwan have dense internal connections, while not many external connections exist between them. We suggest putting more effort in developing joint air pollution prevention and control strategies among the more connected regions. Furthermore, although motivated by $PM_{2.5}$ concentrations, the

proposed method is suitable for broader applications. It only requires the data to be non-negative and heterogeneous.

This study can be extended in multiple directions. First, the hourly $PM_{2.5}$ concentration records for the 67 regions are intensive and can be considered as data for 67 functions at different hours. A functional graphical model for non-negative data can be developed in future research. Second, since air quality is sensitive to climate change and modern transportation and production greatly contribute to air pollution, covariates such as climate conditions and socioeconomic factors can be included for better estimation of the conditional relationships. Third, adding geographic location adjacency information to the model is another aspect worth exploring. Location information is critical for air pollutant diffusion. A model with this information can be used to study the transboundary contributions of $PM_{2.5}$ in neighbouring regions as well as their changes over time. Fourth, analysis of the $PM_{2.5}$ data involves the non-centred truncated Gaussian assumption. Therefore, it may be worthwhile to apply more complex distributions, and testing methods can be studied to determine distribution. Finally, many types of air pollutants, such as $PM_{2.5}$, PM_{10}, and sulphur dioxide can be regarded as a system. We may study the conditional relationships among different regions for this system in follow-up research.

ACKNOWLEDGEMENTS

We thank the editor and reviewers for the valuable comments that greatly improved the paper. Dr. Y Li is supported by Platform of Public Health & Disease Control and Prevention, Major Innovation & Planning Interdisciplinary Platform for the 'Double-First Class' Initiative, Renmin University of China. Dr. S Ma is supported by the National Science Foundation (1916251). Dr. X Fan is supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (2020030259).

REFERENCES

Bickel, P.J. & Levina, E. (2008) Covariance regularization by thresholding. *Annals of Statistics*, 36, 2577–2604.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 1–122.

Chang, X., Wang, S., Zhao, B., Xing, J., Liu, X., Wei, L. et al. (2019) Contributions of inter-city and regional transport to $PM_{2.5}$ concentrations in the Beijing-Tianjin-Hebei region and its implications on regional joint air pollution control. *Science of the Total Environment*, 660, 1191–1200.

Chu, H.-J., Huang, B. & Lin, C.-Y. (2015) Modeling the spatio-temporal heterogeneity in the pm10-pm2.5 relationship. *Atmospheric Environment*, 102, 176–182.

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K. et al. (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389, 1907–1918.

Danaher, P., Wang, P. & 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.

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

Gibberd, A.J. & Nelson, J.D.B. (2017) Regularized estimation of piecewise constant Gaussian graphical models: the group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, 26, 623–634.

Guan, P., Wang, X., Cheng, S. & Zhang, H. (2021) Temporal and spatial characteristics of $PM_{2.5}$ transport fluxes of typical inland and coastal cities in China. *Journal of Environmental Sciences*, 103, 229–245.

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

Hsu, C.-H. & Cheng, F.-Y. (2019) Synoptic weather patterns and associated air pollution in Taiwan. *Aerosol and Air Quality Research*, 19, 1139–1151.

Huang, J. & Ma, S. (2010) Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, 16, 176–195.

Huang, G., Zhou, W., Qian, Y. & Fisher, B. (2019) Breathing the same air? Socioeconomic disparities in $PM_{2.5}$ exposure and the potential benefits from air filtration. *Science of the Total Environment*, 657, 619–626.

Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 695–709.

Hyvärinen, A. (2007) Some extensions of score matching. *Computational Statistics & Data Analysis*, 51, 2499–2512.

Jin, Q., Fang, X., Wen, B. & Shan, A. (2017) Spatio-temporal variations of $PM_{2.5}$ emission in China from 2005 to 2014. *Chemosphere*, 183, 429–436.

Khare, K., Oh, S.Y. & Rajaratnam, B. (2015) A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 803–825.

Lee, W. & Liu, Y. (2015) Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research*, 16, 1035–1062.

Lin, L., Drton, M. & Shojaie, A. (2016) Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10, 806–854.

Lv, B., Liu, Y., Yu, P., Zhang, B. & Bai, Y. (2015) Characterizations of $PM_{2.5}$ pollution pathways and sources analysis in four large cities in China. *Aerosol and Air Quality Research*, 15, 1836–1843.

Meinshausen, N. & Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34, 1436–1462.

Monti, R.P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C. & Montana, G. (2014) Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 103, 427–443.

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

Ren, M., Zhang, S., Zhang, Q. & Ma, S. (2021) Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics*. Available from: https://doi.org/10.1111/biom.13426

Su, P.-F., Sie, F.-C., Yang, C.-T., Mau, Y.-L., Kuo, S. & Ou, H.-T. (2020) Association of ambient air pollution with cardiovascular disease risks in people with type 2 diabetes: a Bayesian spatial survival analysis. *Environmental Health*, 19, 110.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Tseng, C.-H., Tsuang, B.-J., Chiang, C.-J., Ku, K.-C., Tseng, J.-S., Yang, T.-Y. et al. (2019) The relationship between air pollution and lung cancer in nonsmokers in Taiwan. *Journal of Thoracic Oncology*, 14, 784–792.

Wang, H., Zhao, L., Xie, Y. & Hu, Q. (2016) APEC blue-The effects and implications of joint pollution prevention and control program. *Science of the Total Environment*, 553, 429–438.

Wang, F., Chen, T., Chang, Q., Kao, Y.-W., Li, J., Chen, M. et al. (2021) Respiratory diseases are positively associated with $PM_{2.5}$ concentrations in different areas of Taiwan. *PLoS ONE*, 16, 1–11.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W. et al. (2021a) Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications. *Remote Sensing of Environment*, 252, 112136.

Wei, J., Li, Z., Pinker, R.T., Wang, J., Sun, L., Xue, W. et al. (2021b) Himawari-8-derived diurnal variations in ground-level $PM_{2.5}$ pollution across China using the fast space-time Light Gradient Boosting Machine (LightGBM). *Atmospheric Chemistry and Physics*, 21, 7863–7880.

Wu, C.-H., Tsai, I.-C., Tsai, P.-C. & Tung, Y.-S. (2019) Large-scale seasonal control of air quality in Taiwan. *Atmospheric Environment*, 214, 116868.

Xie, Y., Zhao, L., Xue, J., Gao, H.O., Li, H., Jiang, R. et al. (2018) Methods for defining the scopes and priorities for joint prevention and control of air pollution regions based on data-mining technologies. *Journal of Cleaner Production*, 185, 912–921.

Xu, H., Guo, B., Qian, W., Ciren, Z.C., Guo, W., Zeng, Q. et al. (2021) Dietary pattern and long-term effects of particulate matter on blood pressure: a large cross-sectional study in Chinese adults. *Hypertension*, 78, 184–194.

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

Yu, S., Drton, M. & Shojaie, A. (2019) Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20, 1–70.

Yuan, M. & Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zhang, N.-N., Ma, F., Qin, C.-B. & Li, Y.-F. (2018) Spatiotemporal trends in $PM_{2.5}$ levels from 2013 to 2017 and regional demarcations for joint prevention and control of atmospheric pollution in China. *Chemosphere*, 210, 1176–1184.

Zhang, Y., Wei, J., Shi, Y., Quan, C., Ho, H.C., Song, Y. et al. (2021) Earlylife exposure to submicron particulate air pollution in relation to asthma development in Chinese preschool children. *Journal of Allergy and Clinical Immunology*, 148, 771–782.e12.

How to cite this article: Zhang, J., Fan, X., Li, Y. & Ma, S. (2022) Heterogeneous graphical model for non-negative and non-Gaussian $PM_{2.5}$ data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(5), 1303–1329. Available from: https://doi.org/10.1111/rssc.12575

APPENDIX. A

Appendix A provides details for deriving (3) from (4). Based on Yu et al. (2019), we assume that the following two conditions hold.

(C1) For any $y_{-j}^{(m)} \in \mathbb{R}^{p-1}$, $\eta^{(m)} \in \mathbb{R}^p$, and an arbitrary $p \times p$ symmetric matrix $\Theta^{(m)}$,

$$\Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)}) h_j(y_j^{(m)}) \frac{\partial \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} \Big|_{y_j^{(m)} \to 0^+}^{y_j^{(m)} \to +\infty} = 0,$$

where $y_{-j}^{(m)} = (y_1^{(m)}, \ldots, y_{j-1}^{(m)}, y_{j+1}^{(m)}, \ldots, y_p^{(m)})^\top$, and $f(y)|_{y_j \to 0^+}^{y_j \to +\infty} = \lim_{y_j \to +\infty} f(y) - \lim_{y_j \to 0^+} f(y)$ for $j = 1, \ldots, p$ and $m = 1, \ldots, M$.

(C2) For any $\eta^{(m)} \in \mathbb{R}^p$ and an arbitrary $p \times p$ symmetric matrix $\Theta^{(m)}$,

$$\mathbb{E} \| \nabla \log \Pr(Y^{(m)}; \Theta^{(m)}, \eta^{(m)}) \circ h^{1/2}(Y^{(m)}) \|_2^2 < +\infty,$$
$$\mathbb{E} \parallel D\{\nabla \log \Pr(Y^{(m)}; \Theta^{(m)}, \eta^{(m)}) \circ h(Y^{(m)})\} \parallel_1 < +\infty,$$

where $D\{(f_1(y^{(m)}), \ldots, f_p(y^{(m)}))^\top\} = (\partial f_1(y^{(m)})/\partial y_1^{(m)}, \ldots, \partial f_p(y^{(m)})/\partial y_p^{(m)})^\top$, and the expectation is taken under $\Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)})$. For a vector $v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, $\|v\|_1 = \sum_{i=1}^p |v_i|$ and $\|v\|_2 = (\sum_{i=1}^p v_i^2)^{1/2}$.

Under conditions (C1) and (C2), $J(\Theta^{(m)}, \eta^{(m)})$ can be written as

$$J(\Theta^{(m)}, \eta^{(m)}) = \int_{\mathbb{R}_+^p} \Pr(y^{(m)}; \Theta^{\dagger(m)}, \eta^{\dagger(m)}) \sum_{j=1}^{p} \Bigg\{ h_j'(y_j^{(m)}) \frac{\partial \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} + h_j(y_j^{(m)}) \frac{\partial^2 \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)2}} + \frac{1}{2} h_j(y_j^{(m)}) \left(\frac{\partial \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} \right)^2 \Bigg\} dy^{(m)} + \text{const},$$

where const is a constant independent of $\Theta^{(m)}$ and $\eta^{(m)}$, and

$$\frac{\partial \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} = -\frac{1}{2} \sum_{l=1}^{p} (\theta_{lj}^{(m)} + \theta_{jl}^{(m)}) (y_l^{(m)})^a (y_j^{(m)})^{a-1} + \eta_j^{(m)} (y_j^{(m)})^{b-1},$$

$$\frac{\partial^2 \log \Pr(y^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)2}} = -\frac{a-1}{2} \sum_{l=1}^{p} \left(\theta_{lj}^{(m)} + \theta_{jl}^{(m)} \right) (y_l^{(m)})^a (y_j^{(m)})^{a-2} - a\, \theta_{jj}^{(m)} (y_j^{(m)})^{2a-2} + (b-1) \eta_j^{(m)} (y_j^{(m)})^{b-2}.$$

The corresponding empirical version can be written as

$$\hat{J}(\Theta^{(m)}, \eta^{(m)}) = \frac{1}{n_m} \sum_{i=1}^{n_m} \sum_{j=1}^{p} \Bigg\{ h_j'(\mathrm{y}_{ij}^{(m)}) \frac{\partial \log \Pr(\mathbf{y}_i^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} + h_j(\mathrm{y}_{ij}^{(m)}) \frac{\partial^2 \log \Pr(\mathbf{y}_i^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)2}} + \frac{1}{2} h_j(\mathrm{y}_{ij}^{(m)}) \left(\frac{\partial \log \Pr(\mathbf{y}_i^{(m)}; \Theta^{(m)}, \eta^{(m)})}{\partial y_j^{(m)}} \right)^2 \Bigg\} + \text{const},$$

where const is omitted, and we still use the notation $\hat{J}(\Theta^{(m)}, \eta^{(m)})$.

Then, we have

$$\hat{J}(\Theta^{(m)}, \eta^{(m)}) = \frac{1}{2} \mathrm{vec}(\Phi^{(m)})^\top \Gamma(\mathbf{Y}^{(m)}) \mathrm{vec}(\Phi^{(m)}) - g(\mathbf{Y}^{(m)})^\top \mathrm{vec}(\Phi^{(m)}).$$

Here, we omit the const term.

APPENDIX. B

Appendix B presents details on updating $\{\Theta_{(t)}\}$, $\{\eta_{(t)}\}$, and $\{Z_{(t)}\}$. Since (5) is separable for $m = 1, \ldots, M$, we provide only the details on updating $\Theta^{(m)}$ and $\eta^{(m)}$ with the superscript $\cdot^{(m)}$ omitted for notational simplicity. For any matrix A and sets S_1 and S_2, we denote $(A)_{S_1,S_2}$ as the submatrix of A, where the columns and rows are restricted to sets S_1 and S_2 respectively. We define $C_1 = \{1, \ldots, p\}$ and $C_2 = \{p+1\}$. We denote $\Gamma_{11,j} = (\Gamma_j^{(m)})_{C_1,C_1}$,

$\Gamma_{12,j} = (\Gamma_j^{(m)})_{c_1,c_2}$, $\Gamma_{22,j} = (\Gamma_j^{(m)})_{c_2,c_2}$, $\Gamma_{11} = \mathrm{diag}(\Gamma_{11,1}, \ldots, \Gamma_{11,p})$, $\Gamma_{12} = \mathrm{diag}(\Gamma_{12,1}, \ldots, \Gamma_{12,p})$, and $\Gamma_{22} = \mathrm{diag}(\Gamma_{22,1}, \ldots, \Gamma_{22,p})$. In other words, Γ_{11}, Γ_{12}, and Γ_{22} are block-diagonal matrices.

At the tth iteration, after simple calculations, we have

$$\eta_{(t)} = \Gamma_{22}^{-1}\left(g_2 - \Gamma_{12}^{\top}\mathrm{vec}(\hat{\Theta}_{(t)})\right),$$

and $\hat{\Theta}_{(t)}$ is the minimizer of the profiled loss:

$$\hat{\Theta}_{(t)} = \arg\min_{\Theta} \frac{1}{2}\mathrm{vec}(\Theta)^{\top}\Gamma_{11|2}\mathrm{vec}(\Theta) - \left(\mathrm{vec}(g_1) - \Gamma_{12}\Gamma_{22}^{-1}g_2\right)^{\top}\mathrm{vec}(\Theta) + \frac{\rho}{2}\left\|\mathrm{vec}(\Theta) - \mathrm{vec}(Z_{(t-1)} - U_{(n-1)})\right\|_2^2,$$

where $\Gamma_{11|2} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{12}^{\top}$. In other words, $\mathrm{vec}(\Theta) = [\Gamma_{11|2} + \rho\, I_{p^2}]^{-1}[\mathrm{vec}(g_1) - \Gamma_{12}\Gamma_{22}^{-1}g_2 + \rho\ \mathrm{vec}(Z_{(t-1)} - U_{(t-1)})]$, and I_{p^2} is a $p^2 \times p^2$ identity matrix.

In the tth iteration,

$$\{Z_{(t)}\} = \arg\min_{\{Z\}} \frac{\rho}{2}\sum_{m=1}^{M}\left\|\Theta_{(t)}^{(m)} - Z^{(m)} + U_{(t-1)}^{(m)}\right\|_F^2 + \lambda_1\sum_{m=1}^{M}\left\|Z^{(m)-}\right\|_1 \quad + \lambda_2\sum_{(m_1,m_2)\in E}\left\|Z^{(m_1)-} - Z^{(m_2)-}\right\|_F. \tag{A1}$$

Since we do not penalize the diagonal elements, they are easy to update via

$$Z_{(t);jj}^{(m)} = \Theta_{(t);jj}^{(m)} + U_{(t-1);jj}^{(m)},\ j = 1, \ldots, p,\ m = 1, \ldots, M.$$

Noting that $Z^{(m)}$ is symmetric, we only need to deal with the upper-triangle elements. We vectorize the upper triangle of $Z^{(m)}$ by row, in symbols $\tilde{z}^{(m)} = (Z_{lj}^{(m)} \parallel j > l) \in \mathbb{R}^{p(p-1)/2}$, and then construct $\mathbf{Z} = [\tilde{z}^{(1)}, \ldots, \tilde{z}^{(M)}]^{\top} \in \mathbb{R}^{M\times p(p-1)/2}$. The same operations are performed on $\{\Theta_{(t)}\}$ and $\{U_{(t)}\}$ to form $\mathbf{\Theta}_{(t)}$ and $\mathbf{U}_{(t)}$ respectively.

We can rewrite Equation (A1) as

$$H(\mathbf{Z}; \overline{\lambda}_1, \overline{\lambda}_2) = \frac{1}{2}\parallel \mathbf{A} - \mathbf{Z} \parallel_F^2 + \underbrace{\overline{\lambda}_1 \parallel \mathbf{Z} \parallel_1}_{R_1(\mathbf{Z})} + \underbrace{\overline{\lambda}_2 \parallel \mathbf{DZ} \parallel_{2,1}}_{R_2(\mathbf{Z})}, \tag{A2}$$

where $\mathbf{A} = \mathbf{\Theta}_{(t)} + \mathbf{U}_{(t-1)}$, $\overline{\lambda}_1 = \lambda_1/\rho$, $\overline{\lambda}_2 = \lambda_2/\rho$, and $\parallel \mathbf{DZ} \parallel_{2,1} = \sum_{l=1}^{M-1}\|(\mathbf{DZ})_{l\cdot}\|_2$; $(\mathbf{DZ})_{l\cdot}$ is the lth row of matrix $\mathbf{DZ}$. Matrix $\mathbf{D} = (D_{lj})$ is determined by network $\mathcal{G}_M(V, E)$. It has M columns and the same number of rows as the size of E. On the lth row of $\mathbf{D}$, which corresponds to the edge $(m_1, m_2) \in E$, the m_1th entry of this row is set to -1 and the m_2th is 1, for $m_1 < m_2$. We denote

$$\mathrm{prox}_{R_1}(\mathbf{A}) = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{A} - \mathbf{Z}\|_F^2 + \overline{\lambda}_1\|\mathbf{Z}\|_1, \tag{A3}$$

$$\mathrm{prox}_{R_2}(\mathbf{A}) = \arg\min_{\mathbf{Z}} \frac{1}{2}\parallel \mathbf{A} - \mathbf{Z} \parallel_F^2 + \overline{\lambda}_2 \parallel \mathbf{DZ} \parallel_{2,1}. \tag{A4}$$

Problem (A3) can be solved with a soft-thresholding operator (Tibshirani, 1996), while problem (A4) can be solved with a group least-angle regression algorithm (Yuan & Lin, 2006) after variable

transformation (Gibberd & Nelson, 2017). We obtain the solution $\text{prox}_{R_1+R_2}(\mathbf{A})$ to Equation (A2) using the Dykstras iterative projection algorithm. The details are given in Algorithm 2.

Algorithm 2: Dykstras iterative projection algorithm

Input: $\mathbf{A}$.

Output: $\text{prox}_{R_1+R_2}(\mathbf{A})$.

Initialize: $\mathbf{Z}_{(0)} = \mathbf{A}$, $\mathbf{P}_{(0)} = \mathbf{0}$, $\mathbf{Q}_{(0)} = \mathbf{0}$, and $t_2 = 0$.

while not convergent **do**

$t_2 = t_2 + 1$;

$\mathbf{V}_{(t_2-1)} = \text{prox}_{R_2}\left(\mathbf{Z}_{(t_2-1)} + \mathbf{P}_{(t_2-1)}\right)$;

$\mathbf{P}_{(t_2)} = \mathbf{P}_{(t_2-1)} + \mathbf{Z}_{(t_2-1)} - \mathbf{V}_{(t_2-1)}$;

$\mathbf{Z}_{(t_2)} = \text{prox}_{R_1}\left(\mathbf{V}_{(t_2-1)} + \mathbf{Q}_{(t_2-1)}\right)$;

$\mathbf{Q}_{(t_2)} = \mathbf{Q}_{(t_2-1)} + \mathbf{V}_{(t_2-1)} - \mathbf{Z}_{(t_2)}$.

end

APPENDIX. C

Appendix C provides an example that considers two spatial networks among M datasets. Figure C1 presents the network structure and true clustering structure. We set $M = 20, 30$; $K = 2, 3$; $p = 20, 50$; and $n_m = 50,100$. Details about the combinations of M, K, p, n_m, and cp are

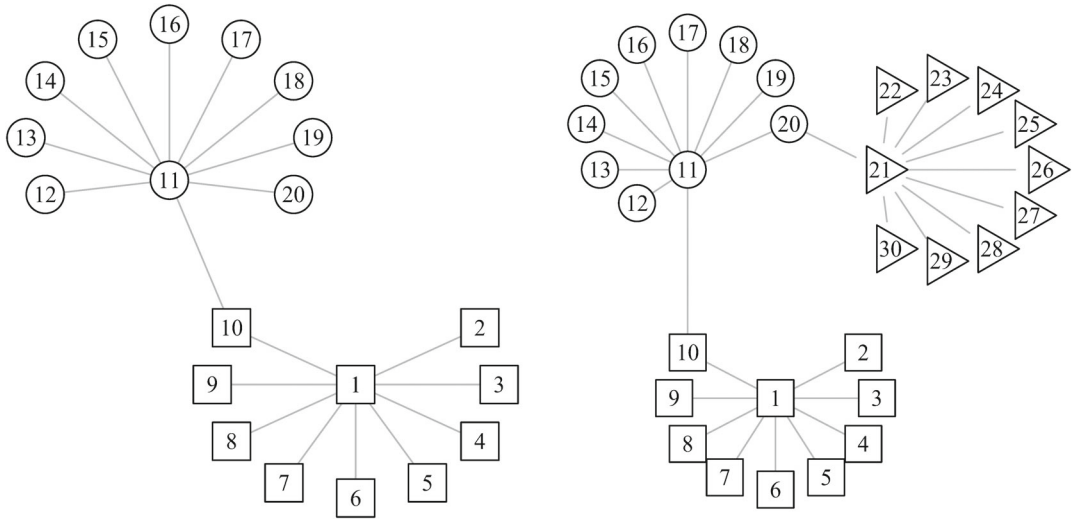

FIGURE C1 Two spatial networks among datasets for $M = 20$ (left panel) or $M = 30$ (right panel). Different shapes denote different clusters.

TABLE C1 Comparison under non-centred truncated Gaussian and non-centred gamma settings with two types of spatial networks. For each setting, the best F_1, TPR, FPR and RI are in boldface. Numbers in parentheses are standard errors

Method	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI	F_1	TPR	FPR	RI
	Truncated Gaussian											
	$n_m = 50, p = 20$				$n_m = 50, p = 20$				$n_m = 100, p = 50$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 30, K = 3$, cp $= \{10, 20\}$				$M = 20, K = 2$, cp $= \{10\}$			
HGMND	**0.795**	**0.917**	**0.021**	0.998	**0.710**	0.929	0.033	**1.000**	**0.751**	**0.815**	0.007	0.996
	(0.039)	(0.041)	(0.005)	(0.010)	(0.024)	(0.016)	(0.004)	(0.000)	(0.024)	(0.034)	(0.001)	(0.013)
GFGL	0.703	0.837	0.028	0.994	0.638	0.744	**0.025**	0.994	0.738	0.762	**0.005**	**1.000**
	(0.043)	(0.039)	(0.006)	(0.015)	(0.053)	(0.075)	(0.008)	(0.018)	(0.025)	(0.012)	(0.001)	(0.000)
rGFGL	0.783	0.903	**0.021**	**0.999**	0.692	**0.933**	0.037	**1.000**	0.746	0.768	0.006	**1.000**
	(0.038)	(0.049)	(0.005)	(0.007)	(0.030)	(0.012)	(0.005)	(0.003)	(0.027)	(0.038)	(0.001)	(0.000)
lGFGL	0.436	0.684	0.077	0.990	0.362	0.626	0.070	0.982	0.407	0.777	0.040	0.997
	(0.047)	(0.101)	(0.020)	(0.024)	(0.030)	(0.033)	(0.010)	(0.016)	(0.021)	(0.022)	(0.004)	(0.011)
IFGL	0.543	0.703	0.046	0.985	0.360	0.758	0.099	0.986	0.545	0.786	0.022	0.988
	(0.043)	(0.096)	(0.012)	(0.022)	(0.058)	(0.074)	(0.032)	(0.014)	(0.020)	(0.029)	(0.002)	(0.021)
GL	0.343	0.439	0.058	0.526	0.233	0.694	0.167	0.690	0.384	0.604	0.029	0.526
	(0.016)	(0.020)	(0.004)	(0.000)	(0.010)	(0.024)	(0.008)	(0.000)	(0.009)	(0.012)	(0.001)	(0.000)
	Gamma											
	$n_m = 50, p = 20$				$n_m = 50, p = 20$				$n_m = 100, p = 50$			
	$M = 20, K = 2$, cp $= \{10\}$				$M = 30, K = 3$, cp $= \{10, 20\}$				$M = 20, K = 2$, cp $= \{10\}$			
HGMND	**0.835**	**0.880**	**0.012**	0.994	**0.818**	**0.888**	**0.012**	0.998	0.717	0.907	**0.012**	**1.000**
	(0.063)	(0.108)	(0.005)	(0.015)	(0.045)	(0.065)	(0.004)	(0.012)	(0.025)	(0.044)	(0.002)	(0.000)
GFGL	0.745	0.851	0.023	0.997	0.633	0.760	0.024	0.994	0.703	0.893	0.013	0.998
	(0.078)	(0.126)	(0.007)	(0.012)	(0.055)	(0.092)	(0.005)	(0.017)	(0.038)	(0.065)	(0.001)	(0.010)
rGFGL	0.795	0.879	0.018	**1.000**	0.797	0.870	0.017	**1.000**	**0.720**	**0.928**	0.013	**1.000**
	(0.037)	(0.045)	(0.004)	(0.005)	(0.036)	(0.045)	(0.004)	(0.005)	(0.014)	(0.015)	(0.001)	(0.000)
lGFGL	0.435	0.683	0.077	0.990	0.435	0.683	0.077	0.990	0.365	0.794	0.051	0.978
	(0.046)	(0.100)	(0.021)	(0.024)	(0.046)	(0.100)	(0.021)	(0.024)	(0.017)	(0.016)	(0.004)	(0.024)
IFGL	0.603	0.739	0.037	0.988	0.512	0.630	0.031	0.983	0.573	0.903	0.025	0.983
	(0.040)	(0.135)	(0.013)	(0.023)	(0.055)	(0.122)	(0.018)	(0.011)	(0.024)	(0.043)	(0.003)	(0.090)
GL	0.317	0.569	0.106	0.526	0.318	0.671	0.106	0.690	0.200	0.809	0.142	0.526
	(0.018)	(0.039)	(0.008)	(0.000)	(0.016)	(0.036)	(0.007)	(0.000)	(0.010)	(0.018)	(0.011)	(0.000)

shown in Table C1. Furthermore, we set $\eta_0^{(m)} \sim \mathcal{N}(\mathbf{2}_p, 0.1^2\, I_p)$ for both the truncated multivariate normal distributions and multivariate gamma distributions. The results are provided in Table C1. Results similar to those in Section 3 are obtained. HGMND is competitive in clustering and sparse structure estimation.