Grouped Network Vector Autoregression

Xuening Zhu[1] and Rui Pan[2]

[1] *School of Data Science, Fudan University, Shanghai, China;*

[2] *Corresponding Author, School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China*

Abstract

In the study of time series analysis, it is of great interest to model a continuous response for all the individuals at equally spaced time points. With the rapid advance of social network sites, network data are becoming increasingly available. In order to incorporate the network information among individuals, Zhu et al. (2017) developed a network vector autoregression (NAR) model. The response of each individual can be explained by its lagged value, the average of its neighbors, and a set of node-specific covariates. However, all the individuals are assumed to be homogeneous since they share the same autoregression coefficients. To express individual heterogeneity, we develop in this work a grouped NAR (GNAR) model. Individuals in a network can be classified into different groups, characterized by different sets of parameters. The strict stationarity of the GNAR model is established. Two estimation procedures are further developed as well as the asymptotic properties. Numerical studies are conducted to evaluate the finite sample performance of our proposed methodology. At last, two real data examples are presented for illustration purpose. They are the studies of user posting behavior on Sina Weibo platform and air pollution pattern (especially $\mathrm{PM}_{2.5}$) in mainland China.

KEY WORDS: EM Algorithm; Network Data; Ordinary Least Square Estimator; Vector Autoregression.

[*] Xuening Zhu[1] is supported by NIDA, NIH grants P50 DA039838, a NSF grant DMS 1512422 and National Nature Science Foundation of China (NSFC, 11690015); Rui Pan[2] is supported by National Nature Science Foundation of China (NSFC, 11601539, 11631003, 71771224), the Fundamental Research Funds for the Central Universities, and China's National Key Research Special Program Grant 2016YFC0207704.

1. INTRODUCTION

An important sign of the rapid development of Internet and mobile Internet is the rise of social networks. Typical representatives include Facebook, Twitter, Sina Weibo, and many others. Accordingly, network data are becoming increasingly available. On one side, users (i.e., nodes) in a social network are no longer independent with each other, but related through various relationships (e.g., friendship). On the other side, plentiful covariates can be collected for each user, such as personal information, consuming behavior, and textual records. As a result, network data play an important role in various disciplines. They can be used to provide site user portraits (Lewis et al., 2008), characterize social capital flow patterns (Bohn et al., 2014), and analyze consumer behavior (Hofstra et al., 2015).

Mathematically, we use an adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ to represent the network structure, where N is the total number of nodes. If the ith node follows the jth one, we set $a_{ij} = 1$; otherwise $a_{ij} = 0$. For convenience, we always let $a_{ii} = 0$. Other than that, we assume that a continuous response $Y_{it} \in \mathbb{R}^1$ can be observed for each node over time t. On social network platform, Y_{it} could be the number of characters posted by node i at time t, reflecting nodal activeness. Furthermore, we denote $\mathbb{Y}_t = (Y_{1t}, \cdots, Y_{Nt})^\top \in \mathbb{R}^N$, and we are particularly interested in studying the dynamic pattern of $\mathbb{Y}_t$. To this end, vector autoregression (VAR) models and the corresponding dimension reduction methods are extensively used in the past literatures, especially the factor models (Pan and Yao, 2008; Lam and Yao, 2012). Recently, Zhu et al. (2017) proposed a network vector autoregression (NAR) model, which takes network structure into account when modeling the dynamics of $\mathbb{Y}_t$.

By NAR, it is assumed that the response Y_{it} is influenced by four factors, (a) its lagged value $Y_{i(t-1)}$, (b) its socially connected neighbors $n_i^{-1}\sum_j a_{ij}Y_{j(t-1)}$ with $n_i =$

$\sum_j a_{ij}$, (c) a set of node-specific covariates $V_i \in \mathbb{R}^p$, and (d) an independent noise ε_{it}. As a result, the model is spelled out as

$$Y_{it} = \beta_0 + \beta_1 n_i^{-1} \sum_j a_{ij} Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + V_i^\top \gamma + \varepsilon_{it}, \tag{1.1}$$

where β_0, β_1, β_2, and γ are referred to as baseline effect, network effect, momentum effect, and nodal effect respectively.

Although model (1.1) can be used to study the dynamic pattern of $\mathbb{Y}_t$ when network information is available, it treats all the nodes to be homogenous. For instance, by the NAR model, the node-irrelevant network effect β_1 implies that all the nodes are influenced by their neighbors to the same extent. This is obviously unrealistic in practice. Take Sina Weibo as an example, which is one of the most popular social network platforms in China. Some nodes on the platform are super stars or political leaders, and they have millions of fans. These nodes are referred to as opinion leaders and less influenced by others (Wasserman and Faust, 1994). As a result, the network effect (i.e., β_1) for the opinion leaders should be small. On contrary, their followers are more likely to be affected, which leads to a relatively large network effect for those ordinary nodes.

From the above discussion, one can conclude that the baseline effect, network effect, momentum effect, and nodal effect might be distinct for different group of nodes. By the real data analysis, we indeed find that nodes in a network can be classified into K groups, characterizing by different sets of parameters (e.g., β_{1k} with $k = 1, \cdots, K$). Figure 1 shows that for the Sina Weibo dataset, nodes are classified into 3 groups, with totally different coefficient estimates. To be more specific, compared to group 2, the estimated network effect is much smaller of group 3 (i.e., $\hat{\beta}_{12} = 0.026$ vs. $\hat{\beta}_{13} = 0.002$).

On the other hand, group 3 has a larger estimated momentum effect than that of group 2 (i.e., $\hat{\beta}_{22} = 0.396$ vs. $\hat{\beta}_{23} = 0.958$). This indicates that nodes in group 2 tend to be affected by their connected neighbors, while those in group 3 are more likely to be self-influenced.

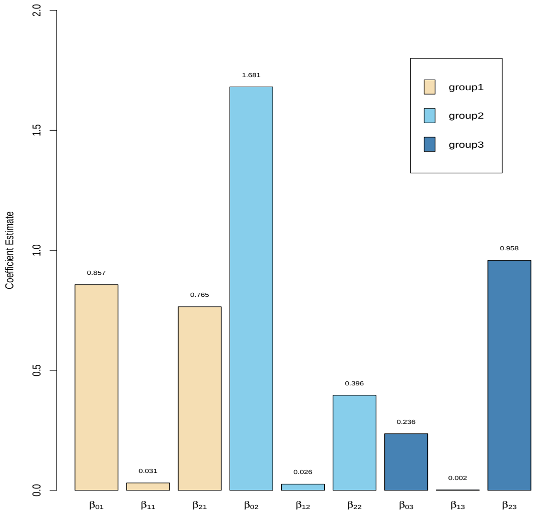

Figure 1: Coefficient estimates for 3 different groups. Distinct characteristics can be obviously detected for different groups of nodes.

In order to capture this interesting phenomenon, we propose in this work a grouped network vector autoregression (GNAR) model. The GNAR model basically assumes that nodes in a network can be classified into different groups, characterized by different sets of parameters. The proposed model is related to the literature of clustering time series data, where the most popularly used technique is model-based clustering established with finite mixture models (Fröhwirth-Schnatter and Kaufmann, 2008; Juárez and Steel, 2010; Wang et al., 2013). In this approach, each time series is assumed to belong to one specific group, and each group is characterized by a different data generating mechanism. The method is widely applied to gene expression classification (Luan and Li, 2003; Heard et al., 2006), financial data modelling (Frühwirth-Schnatter and Kaufmann, 2006; Bauwens and Rombouts, 2007) and economic growth analysis (Fröhwirth-Schnatter and Kaufmann, 2008; Juárez and Steel, 2010; Wang et al., 2013). To our best knowledge, most of the above methods deal with independent univariate time series and can be difficult to directly apply to network data.

In this article, we consider to group users according to their dynamic network behaviors. The network information is employed and embedded into modelling. Specifically, Section 2 explicitly introduces the GNAR model, including the establishment of the strict stationarity of $\mathbb{Y}_t$. In section 3, two estimation methods are developed, an EM algorithm and a two step estimation procedure. The corresponding asymptotic properties are further built. A number of simulation studies are conducted in Section 4 in order to demonstrate the finite sample performance of our methodology. Two real examples are studied in Section 5. The first dataset is about user posting collected from Sina Weibo platform (the largest Twitter type social media in China). The second one is a PM$_{2.5}$ dataset, which are recorded across mainland China. At last, some concluding remarks are given in Section 6. All the technical proofs are left in the separate supplementary material.

2. GROUPED NETWORK VECTOR AUTOREGRESSION

2.1. Model and Notations

Recall the NAR model defined in (1.1). We are interested in modeling the dynamics of $\mathbb{Y}_t$. It can be noted that all the effects are invariant with node, which implies all the nodes are homogenous. However, as discussed above, this assumption might be too stringent in real practice. To fix this problem, we assume nodes in the network can be classified into K groups, where each group is characterized by a specific set of parameters $\theta_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \gamma_k^\top)^\top \in \mathbb{R}^{p+3}$ for $1 \le k \le K$. Let $\mathcal{F}_t$ be the σ-field generated by $\{Y_{is} : 1 \leq i \leq N, 1 \leq s \leq t\}$. Given $\mathcal{F}_{t-1}$, $Y_{1t}, \cdots, Y_{Nt}$ are assumed to be independent and follow a mixture Gaussian distribution

$$\sum_{k=1}^{K} \alpha_k f\Big(\beta_{0k} + \beta_{1k} n_i^{-1} \sum_j a_{ij} Y_{j(t-1)} + \beta_{2k} Y_{i(t-1)} + V_i^\top \gamma_k, \sigma_k^2\Big), \tag{2.1}$$

where $\alpha_k \geq 0$ satisfying $\sum_{k=1}^{K} \alpha_k = 1$ is the group ratio, and $f(\mu, \sigma^2)$ is the probability density function for normal distribution with mean μ and variance σ^2. Model (2.1) is referred to as grouped network vector autoregression model. Essentially, the GNAR model specifies different dynamical patterns for each group through different set of parameters. Following the NAR model, we refer to β_{0k}, β_{1k}, β_{2k}, and γ_k as *grouped* baseline effect, network effect, momentum effect, and nodal effect respectively.

In (2.1), it is not specified which group each node belongs to. We then assume the ith node carries a latent variable $z_{ik} \in \{0, 1\}$. Specifically, $z_{ik} = 1$ if i is from the kth group, otherwise $z_{ik} = 0$. As a result, the GNAR model (2.1) can be written as

$$Y_{it} = \sum_{k=1}^{K} z_{ik}\Big(\beta_{0k} + \beta_{1k} n_i^{-1} \sum_j a_{ij} Y_{j(t-1)} + \beta_{2k} Y_{i(t-1)} + V_i^\top \gamma_k + \sigma_k \varepsilon_{it}\Big), \tag{2.2}$$

where ε_{it} is the independent noise term, and follows standard normal distribution. One could further represent the GNAR model in a random coefficient form as

$$Y_{it} = b_{0i} + b_{1i} n_i^{-1} \sum_j a_{ij} Y_{j(t-1)} + b_{2i} Y_{i(t-1)} + V_i^\top r_i + \delta_i \varepsilon_{it}, \tag{2.3}$$

where $b_{ji} = \sum_k z_{ik} \beta_{jk}$ for $0 \leq j \leq 2$, $r_i = \sum_k z_{ik} \gamma_k$, and $\delta_i = \sum_{ik} z_{ik} \sigma_k$. Note that (2.3) can be seen as a generalized extension of the NAR model. The main differences lie in two aspects, (a) the effects (i.e., coefficients) are all node-specific, reflecting the heterogenous characteristics of each node, and (b) all the parameters are random (i.e., linear combination of the latent variables z_{ik}). This makes the GNAR model (2.3) more flexible and realistic in practice.

Remark 1. The GNAR model (2.3) takes only one lag information into consideration. As a flexible extension, one could consider the GNAR(p) model by taking more historical information as,

$$Y_{it} = b_{0i} + \sum_{m=1}^{q} b_{1i}^{(m)} n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-m)} + \sum_{m=1}^{p} b_{2i}^{(m)} Y_{i(t-m)} + V_i^\top r_i + \delta_i \varepsilon_{it}, \tag{2.4}$$

where $b_{1i}^{(m)} = \sum_k z_{ik}\beta_{1k}^{(m)}$ and $b_{2i}^{(m)} = \sum_k z_{ik}\beta_{2k}^{(m)}$. Similarly, the theoretical properties and estimation methods can be extended with the GNAR(p) model (2.4). In this work, we only focus on the GNAR model with one lag for simplicity.

Recall $\mathbb{Y}_t = (Y_{1t}, \cdots, Y_{Nt})^\top \in \mathbb{R}^N$ is the vector of responses at time t. Let $D_k = \text{diag}\{z_{ik} : 1 \le i \le N\} \in \mathbb{R}^{N\times N}$ with $1 \le k \le K$. Further define $\mathbb{V} = (V_1, \cdots, V_N)^\top \in \mathbb{R}^{N\times p}$ and $\mathcal{B}_0 = \sum_{k=1}^{K} D_k(B_{0k} + \mathbb{V}\gamma_k) \in \mathbb{R}^N$, where $B_{0k} = \beta_{0k}\mathbf{1} \in \mathbb{R}^N$ and $\mathbf{1} = (1, \cdots, 1)^\top$ with compatible dimension. Similarly, write $\mathcal{B}_1 = \sum_{k=1}^{K} D_k B_{1k} \in \mathbb{R}^{N\times N}$ and $\mathcal{B}_2 = \sum_{k=1}^{K} D_k B_{2k} \in \mathbb{R}^{N\times N}$, where $B_{jk} = \beta_{jk} I \in \mathbb{R}^{N\times N}$ for $j = 1, 2$ and I is the identity matrix with compatible dimension. Then the GNAR model can be written in a vector form as

$$\mathbb{Y}_t = \mathcal{B}_0 + \mathcal{G}\mathbb{Y}_{t-1} + \mathcal{E}_t, \tag{2.5}$$

where $\mathcal{G} = \mathcal{B}_1 W + \mathcal{B}_2$, $W = \text{diag}\{n_1^{-1}, \cdots, n_N^{-1}\}A$ is the row-normalized adjacency matrix, and $\mathcal{E}_t = (\delta_1\varepsilon_{1t}, \cdots, \delta_N\varepsilon_{Nt})^\top \in \mathbb{R}^N$ is the noise vector.

2.2. Strict Stationarity of GNAR

As long as we derive (2.5), it is important to study the strict stationarity of the GNAR model. When N is fixed, we have the following theorem.

Theorem 1. *Assume $E\|V_i\| < \infty$ and N is fixed. If $\max_{1\le k\le K}(|\beta_{1k}| + |\beta_{2k}|) < 1$, then there exists a unique stationary solution $\{\mathbb{Y}_t\}$ with $E\|\mathbb{Y}_t\| < \infty$ to the GNAR model (2.5). The solution takes the form:*

$$\mathbb{Y}_t = (I - \mathcal{G})^{-1}\mathcal{B}_0 + \sum_{j=0}^{\infty} \mathcal{G}^j \mathcal{E}_{t-j}. \tag{2.6}$$

The proof of Theorem 1 is given in Section 2 in the supplementary material. Regarding Theorem 1, we have the following remarks.

Remark 2. Given group label $\boldsymbol{Z} = \{z_{ik} : 1 \le i \le N, 1 \le k \le K\}$, define conditional expectation of $\mathbb{Y}_t$ as $\mu_Y = E(\mathbb{Y}_t|\boldsymbol{Z}) = (I - \mathcal{G})^{-1}b_0$, where $b_0 = (b_{01}, \cdots, b_{0N})^\top \in \mathbb{R}^N$. More specifically, denote $\mu_Y = (\mu_1, \cdots, \mu_N)^\top \in \mathbb{R}^N$. As discussed before, Y_{it} could be the number of posts a node made on social network platform. As a result, μ_Y can be seen as nodal activeness level and is of great interest to be investigated. Further let $\mathcal{M}_k = \{i_1, \cdots, i_{N_k}\}$ be the collection of node indexes of the kth group, and $|\mathcal{M}_k| = N_k$ is the group size. It can be verified that the conditional expectation for nodes belonging to the same group is identical, i.e., $\mu_{i_1} = \cdots = \mu_{i_{N_k}} = \nu_k$.

Remark 3. In addition to the conditional mean, we also study the conditional covariance of $\mathbb{Y}_t$. For any integer h, define the auto covariance function of $\mathbb{Y}_t$ given $\boldsymbol{Z}$ as $\Gamma(h) = \text{cov}(\mathbb{Y}_t, \mathbb{Y}_{t-h}|\boldsymbol{Z})$. It can be verified that $\Gamma(0) = (I - \mathcal{G})^{-1}\Sigma_{\mathbb{V}}(I - \mathcal{G}^\top)^{-1} + \Sigma_{\mathcal{E}}$, where $\Sigma_{\mathbb{V}} = \text{diag}\{\sum_{k=1}^K z_{ik}(\gamma_k^\top \Sigma_V \gamma_k) : 1 \le i \le N\}$ with $\Sigma_V = \text{cov}(V_1)$, and $\text{vec}(\Sigma_{\mathcal{E}}) = (I - \mathcal{G} \otimes \mathcal{G})^{-1}\text{vec}(\Sigma_e)$ with $\Sigma_e = \text{diag}\{\sum_{k=1}^K z_{ik}\sigma_k^2 : 1 \le i \le N\}$. It can be further verified that $\Gamma(h) = \mathcal{G}^h\Gamma(0)$ for $h > 0$ and $\Gamma(h) = \Gamma(0)(\mathcal{G}^\top)^{-h}$ for $h < 0$.

To better understand (2.6), we consider a special network structure, the "core-periphery" network. Specifically, there are two groups of nodes in this kind of network, the core (i.e., group 1) and the periphery (i.e., group 2). Nodes in the core group are often celebrities who has a number of followers. While nodes in the periphery group have very few followers and they are influenced by the core. Figure 2 is a diagram of the core-periphery network.

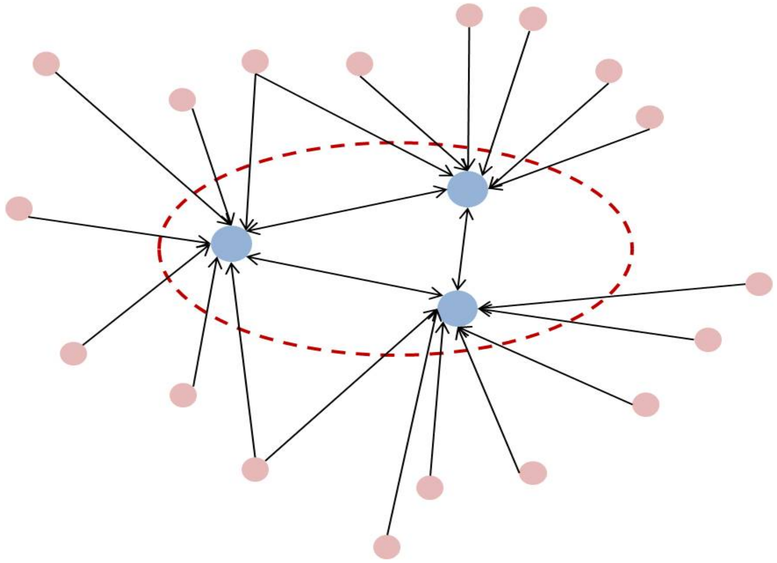

Figure 2: The core-periphery network structure. Blue circle represents the core, while pink the periphery. The core and the periphery can be seen as two different groups, identifying by their own regression coefficients. The arrow represents the direction of the relationship.

Without loss of generality, let the first s nodes be the first group and the left $N - s$ another. Accordingly, let $W = (W_{11}, W_{12}; W_{21}, W_{22})$ be the partition of the two groups. Edges are seldom observed from the core to the periphery or among the periphery. Accordingly , we set $W_{12} = \mathbf{0}$ and $W_{22} = \mathbf{0}$. It can be analytically computed that the conditional expectation of the two groups are $\nu_1 = \beta_{01}/(1 - \beta_{21} - \beta_{11})$ and $\nu_2 = (1 - \beta_{22})^{-1}(\beta_{02} + \beta_{12}\nu_1)$. In such a case, the conditional mean for the core is only determined by its own coefficients (i.e., β_{01}, β_{11}, and β_{21}). However, the activeness level of the periphery is also influenced by the core through the term $\beta_{12}\nu_1$.

3. PARAMETER ESTIMATION

In this section, we discuss the estimation of the GNAR model. Note that the group label z_{ik} is latent. Therefore, parameter estimation and group detection need to be conducted at the same time. Since the procedure might not be straightforward, as a starting point, we assume the group label is known. In fact, this can be useful when the groups are pre-determined by some preliminary knowledge.

3.1. Estimation When Group Label is Known

Define $\mathbb{Y}_t^{(k)} = (Y_{it} : i \in \mathcal{M}_k)^\top \in \mathbb{R}^{N_k}$, $W^{(k)} = (w_{ij} : i \in \mathcal{M}_k, 1 \leq j \leq N) \in \mathbb{R}^{N_k \times N}$, $\mathbb{V}^{(k)} = (V_i : i \in \mathcal{M}_k)^\top \in \mathbb{R}^{N_k \times p}$, and $\mathcal{E}_t^{(k)} = (\varepsilon_{it} : i \in \mathcal{M}_k)^\top \in \mathbb{R}^{N_k}$. Then the GNAR model (2.3) can be rewritten as

$$\mathbb{Y}_t^{(k)} = \beta_{0k} + \beta_{1k} W^{(k)} \mathbb{Y}_{t-1} + \beta_{2k} \mathbb{Y}_{t-1}^{(k)} + \mathbb{V}^{(k)} \gamma_k + \sigma_k \mathcal{E}_t^{(k)}, \tag{3.1}$$

for $k = 1, \cdots, K$. Let $X_{it} = (1, w_i^\top \mathbb{Y}_t, Y_{it}, V_i^\top)^\top \in \mathbb{R}^{p+3}$, where w_i is the ith row of W. Further let $\mathbb{X}_t^{(k)} = (X_{it}^\top : i \in \mathcal{M}_k) \in \mathbb{R}^{N_k \times (p+3)}$. Recall that $\theta_k = (\beta_{0k}, \beta_{1k}, \beta_{2k}, \gamma_k^\top)^\top \in \mathbb{R}^{p+3}$. Then (3.1) could be written as $\mathbb{Y}_t^{(k)} = \mathbb{X}_t^{(k)} \theta_k + \sigma_k \mathcal{E}_t^{(k)}$. Subsequently, the ordinary least squares (OLS) estimator can be obtained for the kth group as

$$\hat{\theta}_k = \Big(\sum_{t=1}^T \mathbb{X}_{t-1}^{(k)\top} \mathbb{X}_{t-1}^{(k)} \Big)^{-1} \Big(\sum_{t=1}^T \mathbb{X}_{t-1}^{(k)\top} \mathbb{Y}_t^{(k)} \Big). \tag{3.2}$$

It is then of great interest to investigate the asymptotic properties of $\hat{\theta}_k$.

Define $\mu_Y^{(k)} = (\mu_i : i \in \mathcal{M}_k)^\top \in \mathbb{R}^{N_k}$. In addition, let $\Sigma_Y = \Gamma(0) = (\sigma_{y,ij}) \in \mathbb{R}^{N \times N}$, $\Sigma_Y^{(k)} = (\sigma_{y,ij} : i \in \mathcal{M}_k, 1 \leq j \leq N) \in \mathbb{R}^{N_k \times N}$, and $\Sigma_Y^{(k,k)} = (\sigma_{y,ij} : i \in \mathcal{M}_k, j \in \mathcal{M}_k) \in \mathbb{R}^{N_k \times N_k}$. The following technical conditions are required.

(C1) (Group Size) Assume that $\min_k N_k = O(N^\delta)$, where $0 < \delta \leq 1$.

(C2) (Independence Assumption) Assume that V_is are independent and identically distributed random vectors with $E(V_1) = \mathbf{0}$, $\mathrm{cov}(V_1) = \Sigma_V \in \mathbb{R}^{p\times p}$, and finite fourth order moment. Assume ε_{it}s are independent and identically distributed. In addition, assume $\{V_i\}$ and $\{\varepsilon_{it}\}$ are mutually independent.

(C3) (Network Structure) Assume W is a sequence of matrices indexed by N. They are assumed to be non-stochastic.

(C3.1) (Connectivity) Treat W as a transition probability matrix of a Markov chain with state space to be the set of all the nodes in the network (i.e., $\{1, \cdots, N\}$). Suppose the Markov chain is irreducible and aperiodic. Further define $\pi = (\pi_1, \cdots, \pi_N)^\top \in \mathbb{R}^N$ as the stationary distribution of the Markov chain, such that (a) $\pi_i \geq 0$ with $\sum_{i=1}^N \pi_i = 1$, and (b) $\pi = W^\top \pi$. Furthermore, $\sum_{i=1}^N \pi_i^2$ is assumed to converge to 0 as $N \to \infty$.

(C3.2) (Uniformity) Define $W^* = W + W^\top$ as a symmetric matrix. Assume $\lambda_{\max}(W^*) = O(\log N)$ and $\lambda_{\max}(WW^\top) = O(N^{\delta'})$ for $\delta' < \delta$, where $\lambda_{\max}(M)$ stands for the largest absolute eigenvalue of an arbitrary symmetric matrix M, and δ is defined in (C1).

(C4) (Law of Large Numbers) Assume the following limits exist: $c_{1\beta}^{(k)} = \lim_{N_k\to\infty} N_k^{-1}(\mathbf{1}^\top W^{(k)}\mu_Y)$, $c_{2\beta}^{(k)} = \lim_{N_k\to\infty} N_k^{-1}(\mathbf{1}^\top \mu_Y^{(k)})$, $\Sigma_1^{(k)} = \lim_{N_k\to\infty} N_k^{-1}\{\mu_Y^{(k)\top}\mu_Y^{(k)} + \mathrm{tr}(W^{(k)\top}W^{(k)}\Sigma_Y)\}$, $\Sigma_2^{(k)} = \lim_{N_k\to\infty} N_k^{-1}\{(\mu_Y^{(k)\top}W^{(k)}\mu_Y) + \mathrm{tr}(W^{(k)}\Sigma_Y^{(k)\top})\}$, and $\Sigma_3^{(k)} = \lim_{N_k\to\infty} N_k^{-1}\{(\mu_Y^{(k)\top}\mu_Y^{(k)}) + \mathrm{tr}(\Sigma_Y^{(k,k)})\}$ for $k = 1, \cdots, K$.

Condition (C1) is an assumption on group size, which assumes that the diverging speed of all groups should be at least faster than $O(N^\delta)$ for $\delta > 0$. It is remarkable that the unbalanced group size is allowed, which could widely exist in real practice. Next, condition (C2) is a regular assumption imposed on the nodal covariates Z_i and noise term ε_{it}. Condition (C3) sets constraints on the network structure W. Specifically, condition (C3.1) requires certain extent of connectivity should exist for the network. Here a sufficient condition for the irreducibility of the Markov chain is that, there should exist a path with finite length between two arbitrary nodes. Condition (C3.2) restricts the heterogeneity of the nodes in the network, which requires the divergence rate of $\lambda_{\max}(W^*)$ and $\lambda_{\max}(WW^\top)$ should not be too fast. Lastly, condition (C4) is a law of large numbers condition for each group. It assumes the limits of certain network features exist as $N_k \to \infty$ for $k = 1, \cdots, K$.

Theorem 2. *Assume* $\max_k(|\beta_{1k}| + |\beta_{2k}|) < 1$ *and Conditions (C1)–(C4). We have* $\sqrt{N_k T}(\hat{\theta}_k - \theta_k) = O_p(1)$ *as* $\min\{N_k, T\} \to \infty$.

The proof of Theorem 2 is given in Section 3 in a supplementary material. By Theorem 2, one could see that the $\sqrt{N_k T}$-consistency can be obtained for the estimator $\hat{\theta}_k$.

3.2. An EM Algorithm

Although the OLS estimation in (3.2) is simple and straightforward, it can be limited since the group label is unknown. Recall that the latent variable $z_{ik} \in \{0, 1\}$ indicates whether the ith user belongs to the kth group. Denote Θ as the parameter space. The full likelihood function is given as

$$L(\Theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\prod_{t=1}^{T} \alpha_k \phi\big\{\sigma_k^{-1}(Y_{it} - X_{it}^\top \theta_k)\big\} \right]^{z_{ik}}, \tag{3.3}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution. We then adopt an EM algorithm for parameter estimation. In particular, after setting an initial value $\hat{\theta}^{(0)}$, we iterate the following steps. Specifically, in the mth ($m \geq 1$) iteration, we have that

E-Step. Estimate z_{ik} by its posterior mean $z_{ik}^{(m)}$. Here,

$$z_{ik}^{(m)} = E\big(z_{ik}|\hat{\theta}^{(m-1)}\big) = \frac{\hat{\alpha}_k^{(m-1)} \prod_{t=1}^{T} \phi(\hat{\Delta}_{it,k}^{(m-1)})}{\sum_{k=1}^{K} \hat{\alpha}_k^{(m-1)} \prod_{t=1}^{T} \phi(\hat{\Delta}_{it,k}^{(m-1)})}, \tag{3.4}$$

where $\hat{\Delta}_{it,k}^{(m-1)} = (Y_{it} - X_{i(t-1)}^{\top}\hat{\theta}_k^{(m-1)})/\hat{\sigma}_k^{(m-1)}$, and $\hat{\theta}_k^{(m-1)}$, $\hat{\sigma}_k^{(m-1)}$ are the estimates from the $(m-1)$th iteration.

M-Step. Given $z_{ik}^{(m)}$, we then maximize (3.3) with regarding to α_k, θ_k, and σ_k. Particularly, we have

$$\hat{\theta}_k^{(m)} = \Big(\sum_i z_{ik}^{(m)} \sum_t X_{it}X_{it}^{\top}\Big)^{-1}\Big(\sum_i z_{ik}^{(m)} \sum_t X_{it}Y_{it}\Big), \tag{3.5}$$

$$\big(\hat{\sigma}_k^2\big)^{(m)} = \Big(T\sum_i z_{ik}\Big)^{-1}\Big\{\sum_i z_{ik}^{(m)} \sum_t (Y_{it} - X_{it}^{\top}\hat{\theta}_k^{(m)})^2\Big\}, \;\; \hat{\alpha}_k^{(m)} = N^{-1}\big(\sum_i^N z_{ik}^{(m)}\big). \tag{3.6}$$

Repeat the above steps until the EM algorithm converges and the final results are the desired estimators.

It can be noted that the estimation given by (3.5) is in spirit similar to (3.2). Particularly, the EM estimation of θ_k can be treated as a weighted OLS estimator, where the weights are the latent group variables z_{ik}. In addition, the estimation of σ_k^2 and α_k in (3.6) can be comprehended in similar way.

3.3. A Two Step Estimation Method

In real practice, the computation of the E-Step (3.4) might be not stable when the time dimension T is large. That makes the estimation result in M-Step might not be reliable. Note that (2.3) can be treated as a random coefficient model with node-specific coefficients. Motivated by this fact, we consider a two step estimation procedure as an alternative. In the first step, we estimate the coefficient at the nodal level. Secondly, these estimates are pooled together to obtain the parameter estimation $\hat{\theta}_k$ for $k = 1, \cdots, K$. For convenience, we assume $(\beta_{1k}, \beta_{2k})^\top$ are not the same between different groups.

Specifically, let $b_i = (b_{0i} + V_i^\top \gamma_i, b_{1i}, b_{2i})^\top \in \mathbb{R}^3$. Write $\boldsymbol{X}_{it} = (1, w_i^\top \mathbb{Y}_t, Y_{it})^\top \in \mathbb{R}^3$. Then the estimates for b_i can be obtained as

$$\hat{b}_i = \Big(\sum_{t=1}^{T} \boldsymbol{X}_{i(t-1)} \boldsymbol{X}_{i(t-1)}^\top\Big)^{-1} \Big(\sum_{t=1}^{T} \boldsymbol{X}_{i(t-1)} Y_{it}\Big). \tag{3.7}$$

Note that (3.7) is the ordinary least squares estimation for each node. Intuitively, this estimate will approximate the true value b_i well when T is sufficiently large.

Theorem 3. *Assume $N = o(\exp(T))$, the stationary condition $\max_k(|\beta_{1k}| + |\beta_{2k}|) < 1$, and conditions (C1)–(C4). In addition, assume there exists $\tau > 0$, such that $\min_i\{(e_i^\top \Sigma_Y e_i)(w_i^\top \Sigma_Y w_i) - (e_i^\top \Sigma_Y w_i)^2\} \geq \tau$ with probability tending to 1. Then we have $\sup_{1 \leq i \leq N} \|\hat{b}_i - b_i\| = o_p(1)$.*

The proof of Theorem 3 is given in Section 4 in a supplementary material. Regarding the term appeared above $(e_i^\top \Sigma_Y e_i)(w_i^\top \Sigma_Y w_i) - (e_i^\top \Sigma_Y w_i)^2$, we rewrite it as $\sum_i \sum_{j_1, j_2} \Delta_{ij_1j_2} w_{ij_1} w_{ij_2} (\widetilde{\sigma}_{y,j_1j_2} - \widetilde{\sigma}_{y,ij_1}\widetilde{\sigma}_{y,ij_2})$, where $\Delta_{ij_1j_2} = \sigma_{y,ii}\sigma_{y,j_1j_1}\sigma_{y,j_2j_2}$ and $\widetilde{\sigma}_{y,ij} = \mathrm{cor}(Y_{it}, Y_{jt})$. Then the condition can be satisfied if $\sigma_{y,ii}$ and $\widetilde{\Delta}_{ij_1j_2} = \widetilde{\sigma}_{y,j_1j_2} - \widetilde{\sigma}_{y,ij_1}\widetilde{\sigma}_{y,ij_2}$ are lower bounded away from 0, with probability tending to 1 for triplets set $\{(i, j_1, j_2) : a_{ij_1}a_{ij_2} = 1, i \neq j_1, i \neq j_2\}$. Given the results in Theorem 3, it is noteworthy that the overall estimation bias (i.e., $\sup_i \|\hat{b}_i - b_i\|$) can be controlled as the diverging speed of time T is slightly faster than $\log(N)$ (i.e., log transformed network size).

Based on the theoretical result of Theorem 3, we consider the second step for estimation. Ideally, the estimated values $\hat{b}_i$ will form K clusters (i.e., groups) by the cluster algorithm. The corresponding group members are collected in $\widehat{\mathcal{M}}_k$, where $\widehat{N}_k = |\widehat{\mathcal{M}}_k|$. Then the group ratio α_k can be directly estimated by $\hat{\alpha}_k = \widehat{N}_k/N$. Subsequently, given this estimated group information, one can be able to conduct the estimation by using the same procedure (3.2) in section 3.1. Specifically, θ_k can be estimated by

$$\hat{\theta}_k^{TS} = \Big(\sum_{t=1}^{T}\sum_{i\in\widehat{\mathcal{M}}_k} X_{i(t-1)}X_{i(t-1)}^\top\Big)^{-1}\Big(\sum_{t=1}^{T}\sum_{i\in\widehat{\mathcal{M}}_k} X_{i(t-1)}Y_{it}\Big),$$

which is referred to as the two step (TS) estimator. Theoretically, one could expect the consistency result of $\hat{\theta}_k^{TS}$ if all nodes are clustered into their true groups with probability tending to 1 (Hartigan, 1981; Pollard, 1981; Von Luxburg et al., 2008). This can be guaranteed by the result of Theorem 3 when abundant time information can be obtained.

4. NUMERICAL STUDIES

4.1. Simulation Models

To demonstrate the finite sample performance of our proposed methodology, we conduct a number of numerical studies in this section. Specifically, the first two examples are presented with different types of network structures. The third example is displayed to study the parameter estimation and prediction accuracy when the number of groups pis misspecified. In each example, different estimation methods (EM and TS) are employed and compared.

For each example, we fix the number of groups $K = 3$ and generate the random innovations ε_{it} from a standard normal distribution. For convenience, we set $\delta_k = 1$

for $k = 1, \cdots, K$. In addition, nodal covariates $V_i = (V_{i1}, \cdots, V_{i5})^\top \in \mathbb{R}^5$ are independently sampled from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma_v = (\sigma_{j_1 j_2})$ with $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$. The true value of the parameters for each group is listed in Table 1. Furthermore, let $\sigma^2 = 1$ for Example 1 and 2, while $\sigma^2 = 4$ for Example 3. Given the initial value $\mathbb{Y}_0 = \mathbf{0}$, the time series $\mathbb{Y}_t$ is generated according to the GNAR model (2.3). Particularly, the first 50 replications are dropped to ensure the time series to achieve stationarity.

It should be particularly noted that different network and momentum effects are employed for each group to distinguish nodal behaviors. As shown in Table 1, Group 1 has relatively lower activeness level with small positive network and momentum effects (i.e., β_1 and β_2). Group 2 is characterized by its negative network effect (i.e., β_1), which implies nodal behaviors in this group exhibit a negative correlated pattern with their connected friends. Lastly, compared with the other two groups, Group 3 occupies a larger portion (i.e., α) and has higher momentum effect (i.e., β_2). Subsequently, we introduce two typical network structures employed in the simulation studies.

Example 1. (Stochastic Block Model) First of all, we consider the block structure network, which is also known as the stochastic block model (Wang and Wong, 1987; Nowicki and Snijders, 2001; Zhao et al., 2012). This model assumes that nodes in the same block are more likely to be connected. To generate such model, we follow Zhu et al. (2017) to set $J \in \{5, 10, 20\}$ blocks and randomly assign each node a block label with equal probability. Next, let $P(a_{ij} = 1) = 0.3N^{-0.3}$ if i and j are from the same block, otherwise set $P(a_{ij} = 1) = 0.3N^{-1}$. Consequently, nodes within the same block will have higher probability to connect than nodes from different blocks.

Example 2. (Power-law Model) In real network, it can be observed that a small portion of nodes (e.g., super stars and opinion leaders) have a large amount of network links, but the majority have limited number of connections. This phenomenon can be described by the power-law model (Barabási and Albert, 1999). Specifically, we generate the nodal in-degrees $d_i = \sum_j a_{ji}$ from a power-law distribution, i.e., $P(d_i = d) = cd^{-\alpha}$, where c is a normalizing constant and α is the exponent parameter. We set $\alpha = 2.5$ as suggested by Clauset et al. (2009), which is based on empirical studies with real social network data.

Example 3. (Number of Groups) In this example, we evaluate the impact on parameter estimation and prediction accuracy when the numberp of groups K is incorrectly specified. Specifically, data are generated from the power-law model described in Example 2 with total time periods $(T + 20)$. The first T time periods are used for parameter estimation, and the rest 20 periods for prediction. Lastly, we set $K = 1, 2, 3, 5, 7$, where $K = 3$ is the true number of groups.

4.2. Performance Measurements and Simulation Results

For each simulation example, we consider different network sizes $N = 100, 200, 500$. Accordingly, to evaluate the performances of the two proposed estimation methods, we employ two settings of T as $T = N/2$ and $T = 2N$ respectively. For a reliable result, we randomly repeat the simulation experiments for $R = 1000$ times. Let $(\hat{\beta}_{0k}^{(r)}, \hat{\beta}_{1k}^{(r)}, \hat{\beta}_{2k}^{(r)}, \hat{\gamma}_k^{(r)\top})^\top \in \mathbb{R}^{p+3}$ be the estimator of the kth group obtained from the rth replication. In addition, for each node, we are able to obtain its group label as $\hat{z}_i^{(r)}$ for $i = 1, \cdots, N$. Specifically, for the EM algorithm, the group label is defined as $\hat{z}_i^{(r)} = \arg\max_k \{\hat{z}_{ik}\}$. For the two step estimation, the group label is the same with the cluster label after the first step estimation. Subsequently, we consider the following measurements for evaluation of numerical results.

First, for a given parameter, the root mean square error (RMSE) is employed to evaluate the estimation accuracy. Take the network effect $\beta_1 = (\beta_{1k} : 1 \leq k \leq K)^\top \in \mathbb{R}^K$ for example. The RMSE is calculated over all groups as $\text{RMSE}_{\beta_j} = \{(RK)^{-1}\sum_{k=1}^K \sum_{r=1}^R (\hat{\beta}_{jk}^{(r)} - \beta_{jk})^2\}^{1/2}$. Similarly, the RMSE can be computed for baseline effect (i.e., RMSE_{β_0}) and momentum effect (i.e., RMSE_{β_2}) respectively. In addition, the RMSE for the nodal effect is defined as $\text{RMSE}_\gamma = \{(RK)^{-1}\sum_{k=1}^K \sum_{r=1}^R \|\hat{\gamma}_k^{(r)} - \gamma_k\|^2\}^{1/2}$. Next, given the estimated groups $\hat{z}_i^{(r)}$, the misclassification rate (MCR) can be calculated as $\text{MCR} = (NR)^{-1}\sum_{r=1}^R \sum_{i=1}^N I(\hat{z}_i^{(r)} \neq z_i)$, where z_i is the true group label of the node i. Lastly, the average network density (i.e., $\{N(N-1)\}^{-1}\sum_{i_1,i_2} a_{i_1 i_2}$) is also reported.

Lastly, when the number of groups K is misspecified (i.e., Example 3), we evaluate the impact on parameter estimation and prediction accuracy. Denote $\widehat{\mathbb{Y}}_t^{(K)}$ as the fitted response for $t = 1, \cdots, T$ and predicted value for $t = T+1, \cdots, T+20$, where the superscript K indicates the number of groups. In order to evaluate the parameter estimation accuracy, we compare the fitted value $\widehat{\mathbb{Y}}_t^{(K)}$ against the conditional expectation $E(\mathbb{Y}_t | \mathcal{F}_{t-1}, \boldsymbol{Z})$. This is because the comparison cannot be directly conducted for parameter estimation error when group number K is misspecified. We then define the estimation error as

$$\text{Err}_{est}^{(K)} = \big\{(NT)^{-1}\sum_{t=1}^T \big\|\widehat{\mathbb{Y}}_t^{(K)} - E(\mathbb{Y}_t | \mathcal{F}_{t-1}, \boldsymbol{Z})\big\|^2\big\}^{1/2},$$

where $\mathcal{F}_{t-1}$ is the σ-field generated by $\{\mathbb{Y}_s : s \leq t-1\}$ and $E(\mathbb{Y}_t | \mathcal{F}_{t-1}, \boldsymbol{Z})$ is the conditional expectation based on the historical and group information. Next, the prediction error is measured by

$$\text{Err}_{pred}^{(K)} = \big\{(20N)^{-1}\sum_{t=T+1}^{T+20} \big\|\widehat{\mathbb{Y}}_t^{(K)} - \mathbb{Y}_t\big\|^2\big\}^{1/2},$$

which is the RMSE for predicted values. The median values of both $\mathrm{Err}_{est}^{(K)}$ and $\mathrm{Err}_{pred}^{(K)}$ over all replications are reported.

The detailed results are given in Tables 2–4. For the first two examples, it is found that as the network size N and time period T increase, the RMSEs of all estimated parameters decrease towards 0 for both EM algorithm and two step (TS) estimation. In addition, similar pattern can be observed for the MCR, which drops as the network size and time period (i.e., N and T) increase. For a finite sample comparison, it can be observed that the EM algorithm outperforms TS estimation in Scenario 1 when less time information can be obtained (i.e., small T). Specifically, lower RMSE and MCR values are observed. However, the TS estimation has greater advantage over the EM algorithm in Scenario 2 in both parameter estimation and group classification. Lastly, for Example 3, it is found that both the estimation and prediction errors drop sharply from $K \leq 2$ to $K = 3$, where the model is correctly specified with $K = 3$. In the meanwhile, for $K \geq 3$, it is observed that the estimation and prediction errors perform relatively steady.

5. CASE STUDY

In this section, we conduct two case studies to evaluate our proposed methods. The first is about user posting behavior on social network platform. The second is the study of dynamic and spatial pattern of $\mathrm{PM}_{2.5}$. The adjacency matrix is constructed between cities by taking advantage of their spatial locations.

5.1. User Behavior Analysis: A Sina Weibo Dataset

We first apply the proposed GNAR model to a social network dataset. The data are collected from Sina Weibo, which is the largest Twitter type social media in China.

Users are allowed to follow other users, create user profiles, and post Weibo to express their opinions. In addition to ordinary users, the celebrities, public media, as well as companies and organizations are also allowed to register on Sina Weibo. Therefore, the background of users can be diversified, leading to different user behaviour patterns.

Data Description

To investigate user behaviour on Weibo, we collect data of $N = 2,021$ followers of an official account, starting from 2014-01-01 for a total of $T = 11$ consecutive weeks. The response Y_{it} is defined to be $\log(1 + x)$-transformed average Weibo post length (i.e., the average number of characters posted by the user in a week), which can be seen as a representative of nodal activeness level. The histogram of the response is displayed in Figure 3, where an approximately symmetric shape can be observed. In addition, two node-specific variables are recorded, the gender of the user (i.e., male = 1 and female = 0), and the number of personal labels (i.e., keywords created by the Weibo users to describe their life status and interests).

The network adjacency matrix A can be constructed as follows, $a_{ij} = 1$ if the ith user follows the jth one on Weibo, otherwise $a_{ij} = 0$. Particularly, the adjacency matrix is asymmetric since users are not required to be mutually connected on Weibo. We visualize the distribution of nodal in-degree (i.e., $a_{+i} = \sum_j a_{ji}$) and out-degree (i.e., $a_{i+} = \sum_j a_{ij}$) in Figure 4. It can be detected that the distribution of in-degree is more skewed than that of out-degree. This implies there might exist users who attract a large amount of followers. In addition, the network density is 2.7% (i.e., $\sum_{i,j} a_{ij}/\{N(N-1)\}$), which indicates a relatively sparse network.

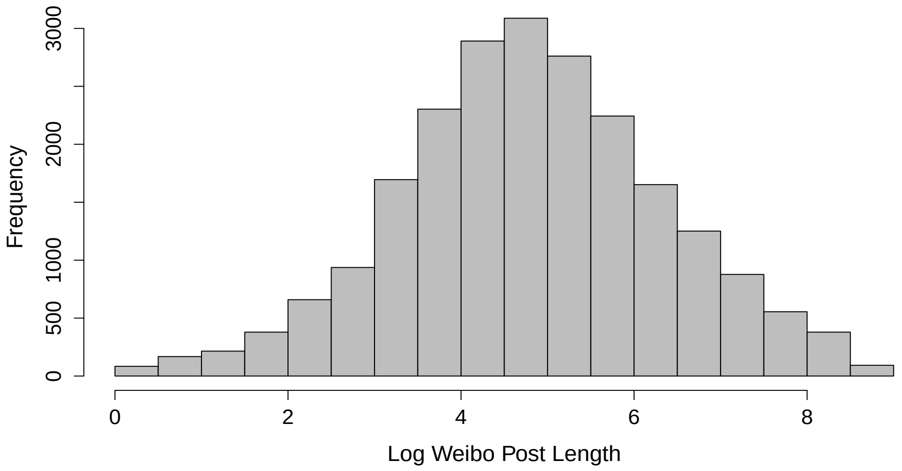

Figure 3: The histogram of responses (i.e., log-transformed weibo post length).

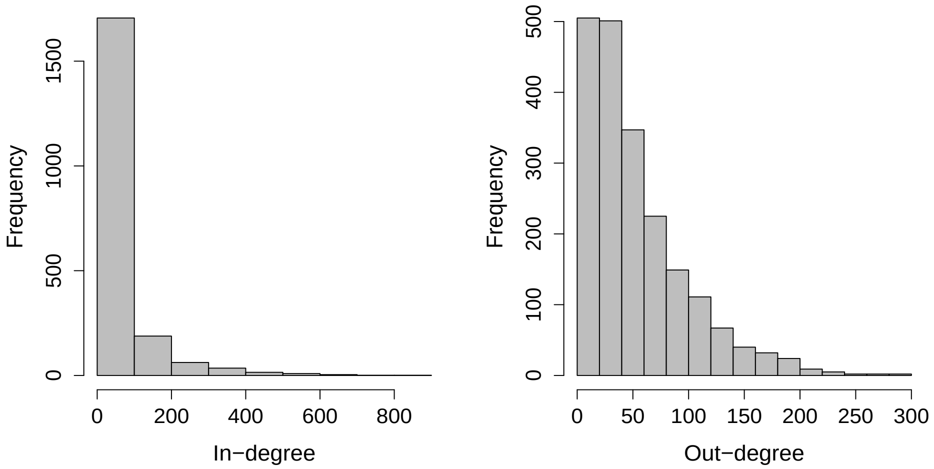

Figure 4: The histogram of nodal in- and out-degree of $N = 2,021$ nodes. A heavily skewed shape can be detected for nodal in-degree, which indicates the existence of "super stars" in the network.

Model Estimation and Explanation

Subsequently, we fit the GNAR model on this dataset. Only EM algorithm is applied, since the network size N is much larger than the number of time periods T. The number of groups is fixed to be $K = 3$ and the estimation results are given in Table 5. One could see that the estimated network effect and momentum effect are all positive for three groups. This suggests user activeness level is positively related to itself as well as that of its following neighbors. Moreover, a stronger momentum effect can be detected compared with the network effect. At last, the estimated nodal effects indicate that the male users with more self-created labels exhibit higher activeness levels.

For further illustration, we conduct comparison among groups. It can be noted that Group 1 and Group 2 occupy a large portion of all network users (with larger estimated α values). Specifically, they both have larger network effects (i.e., the estimated β_1 values) than that of Group 3, implying that users in these two groups tend to be influenced by the ones they follow. When looking at momentum effect (i.e., the estimated β_2 values), it can be observed that users in Group 1 and 3 are more self-motivated than Group 2. Particularly, Group 3 has the largest momentum effect while the smallest network effect. This indicates user behavior of this group can be highly predictable by the history.

Moreover, we draw the boxplot of the responses in a grouped manner in Figure 5. A higher activeness level can be found for Group 3. Actually, users in this group are mostly public media accounts and celebrities with a large amount of followers, such as “Sina Finance”, “Xinhua Views”, “Beijing Youth Daily”, “Phoenix TV”, and many others. These accounts generate contents and release information on the platform frequently such that they can pass information and influence other users. On contrary, most users in Group 1 and Group 2 are ordinary users who play the role of information adopters. Lastly, we conduct a model comparison with the network vector autoregression model (Zhu et al., 2017), and the univariate autoregression (AR) model. The first 9 weeks are used for model training, and the last 2 weeks are employed for prediction evaluation. The predictive root mean square error (RMSE) is used to quantify the prediction accuracy of different models, which are 0.809, 0.850, and 2.312 respectively. It can be observed the predictive RMSE of GNAR is lower than that of NAR and AR, which indicates a better prediction power of the GNAR model.

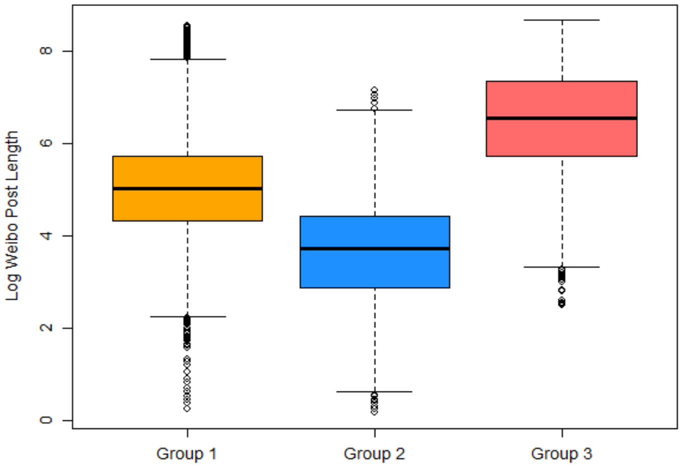

Figure 5: The boxplot of log-transformed weibo post length for each group.

5.4. Air Pollution Analysis: A PM$_{2.5}$ Dataset

In recent years, the issue of air pollution in China has drawn world wide attentions. One particular air pollution is called PM$_{2.5}$, which refers to the airborne particles with aerodynamic diameters less than 2.5 micrometers. There have been evidences that the high concentration of PM$_{2.5}$ may cause severe clinical symptoms, such as lung morbidity, respiratory and cardiovascular diseases. Hence, it is of great importance to understand the PM$_{2.5}$ distribution and diffusion pattern across China.

Data Description

The $PM_{2.5}$ data are collected from air quality monitoring stations over 291 cities in mainland China. Specifically, the daily $PM_{2.5}$ index (unit: $\mu g/m^3$) is recorded from 2015-01-01 to 2015-12-31 with $T = 365$. The left in Figure 6 gives the time series of average daily $PM_{2.5}$ of all cities during 2015. A high $PM_{2.5}$ level can be found in winter (November, December, and January) with highest $PM_{2.5}$ over $100\mu g/m^3$. We then take the yearly average of $PM_{2.5}$ in each city and display it in Figure 7, where darker regions imply higher $PM_{2.5}$ levels. Spatially, the northeastern regions in China (especially in Heibei province) exhibit higher concentration of $PM_{2.5}$.

The response is defined as the log-transformed $PM_{2.5}$ levels, where the histogram is displayed in the right of Figure 6. A symmetric shape can be observed. In order to construct the network structure, we treat each city as a node. The adjacency matrix A is constructed by using spatial distances between any two cities. Let $s_1, \cdots, s_N$ ($s_i \in \mathbb{R}^2$) be the locations of N cities. Then a_{ij} is defined as $a_{ij} = 1/\|s_i - s_j\|$ for $i \neq j$ and $a_{ii} = 0$ for $i = 1, \cdots, N$.

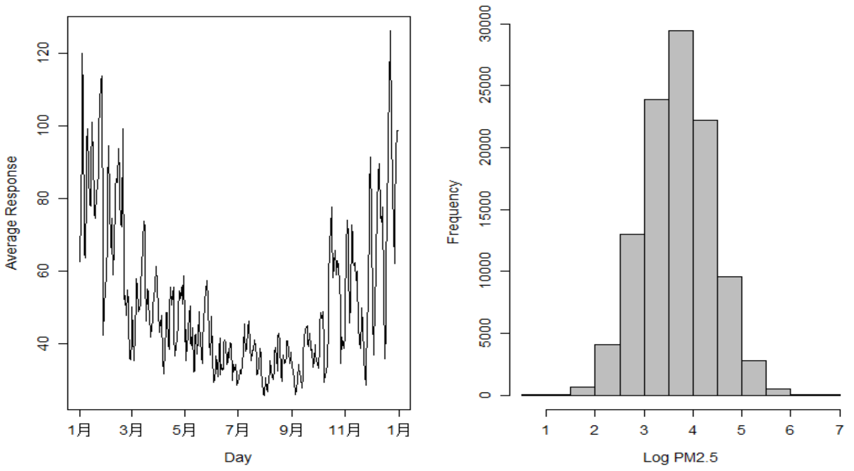

Figure 6: The left panel: daily average $PM_{2.5}$ in the year of 2015; The right panel: the histogram of log-$PM_{2.5}$.

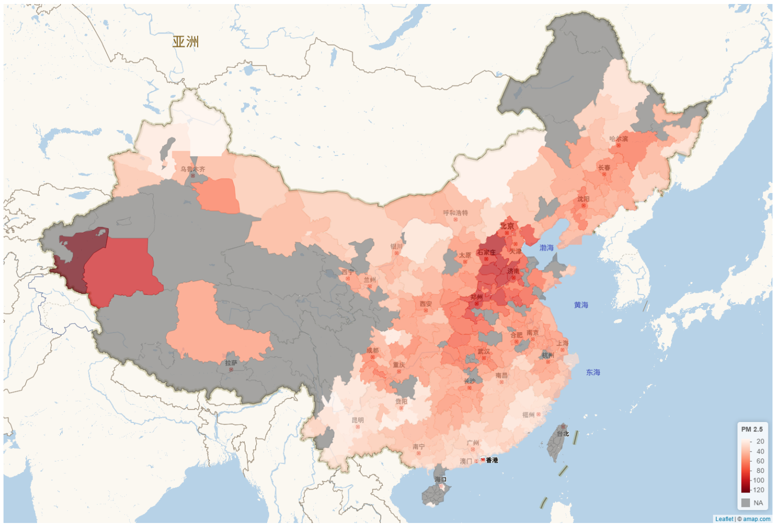

Figure 7: Average $PM_{2.5}$ for each city in the year of 2015. The grey color indicates absence of $PM_{2.5}$ monitoring stations in corresponding cities.

Model Estimation and Explanation

Motivated by the descriptive analysis, we model the dynamic patterns of different seasons separately, which are spring (March to May), summer (June to August), autumn (September to November), and winter (January to February). Intuitively, the number of groups should be large in winter since the pollution level is relatively high. As a result, we set $K = 3$ for winter while $K = 2$ for the other seasons. The GNAR model, NAR model, and AR model are estimated for prediction comparison. The G-NAR model is estimated using the proposed EM algorithm and two step estimation method respectively. For each season, the last 10 days are used to conduct prediction, and the prediction RMSEs are summarized in Table 6. It can be observed that the EM algorithm always outperforms other methods in terms of prediction accuracy. We next illustrate the detailed estimation results by taking advantage of EM algorithm.

The estimated regression coefficients are given in Table 7. We take the results of winter for detailed explanation. First of all, the number of cities in 3 groups is unbalanced, where the proportions are 0.30, 0.12, and 0.58 respectively. It is noteworthy that the first and second group have relatively large baseline effect, which indicates that air pollution in these cities is much more severe. From Figure 8, it can be observed that cities in group 1 and 2 locate in northeastern China. Furthermore, cities in group 1 and 3 have large network effect, which implies that cities in group 1 and 3 are more likely to be influenced by their spatial neighbors. For the other seasons, it can be detected that the patterns in summer and autumn are very similar. This is mainly because that the pollution level is relatively lower in those two seasons.

Regarding this real example, we have two more remarks to make. Firstly, it can be noted that the network structure in this example is symmetric (i.e., $a_{ij} = a_{ji}$). Recall that when the network structure is asymmetric (as in the social network case), the term $n_i^{-1}\sum_j a_{ij}Y_{j(t-1)}$ represents the averaged responses of those nodes that i follows. As a result, network effect β_1 can be viewed as the "influence" that i receives from the nodes it follows (i.e., those js with $a_{ij} = 1$). When the adjacency matrix is symmetric as shown in this example, the term $n_i^{-1}\sum_j a_{ij}Y_{j(t-1)}$ represents the averaged responses of those nodes that i is connected to. The corresponding parameter β_1 can be understood as the "connection" or "correlation" rather than "influence" that node i receives from its connected neighbours. Secondly, in this example, no node-specific covariates are utilized due to our lack of access to more information. It would be an important future research topic to consider nodal effect variables (i.e., V_i) such as temperature, humidity and wind speed into the modelling framework.

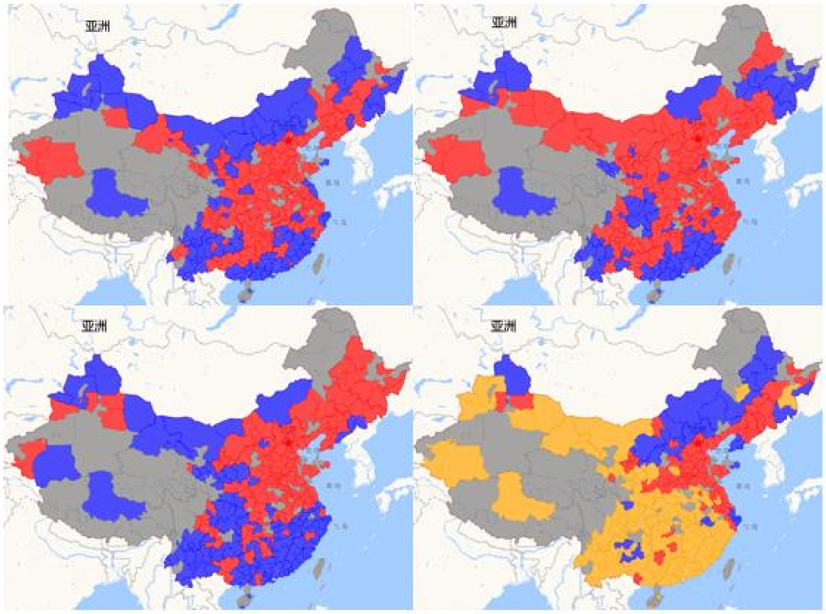

Figure 8: Different groups of cities detected by EM algorithm for spring (left top panel), summer (right top panel), autumn (left bottom panel), and winter (right bottom panel). Cities in Group 1, 2, 3 are marked as red, blue, and yellow.

6. CONCLUDING REMARKS

In this article, we develop a novel GNAR model, which incorporates group specific network autoregression coefficients. To estimate the GNAR model, an EM algorithm and a two step estimation are designed. It is suggested by the numerical result that both methods produce consistent results, while they could have different finite sample performances with different scenarios. Lastly, the Sina Weibo and $PM_{2.5}$ datasets are analyzed for illustration propose, where nodes in the different groups show distinguished behavioral patterns.

To facilitate future research, we discuss here several interesting topics. First, it can be noted that although the estimation and group classification procedure have been developed in this work, however, it is not flexible to conduct inference about the estimated parameters. Therefore, how to make inference can be a problem of interest. Next, for the proposed estimation methods of the GNAR model, the number of groups K needs to pre-specified. Hence how to select K remains to be a challenging task. Lastly, it is assumed that the users can be grouped by their dynamic behavior patterns, which are further quantified by the network autoregression coefficients. As a further extension, one could consider incorporating user network structure information (e.g., the following-followee information of the focal user) together to decide their groups.

APPENDIX

We present here the detailed technical proofs of Lemma 1–Lemma 4 in Appendix A. Next, the proofs of Theorem 1 to Theorem 3 are given in Appendix B, C, and D respectively.

Appendix A. Four Useful Lemmas

Lemma 1. *Let $X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^n$, where X_is are independent and identically distributed random variables with mean zero, variance σ_X^2 and finite fourth order moment. Let $\widetilde{\mathbb{Y}}_t = \sum_{j=0}^\infty G^j U \mathcal{E}_{t-j}$, where $G \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times N}$, and $\{\mathcal{E}_t\}$ satisfy Condition (C1) and are independent of $\{X_i\}$. Then for a matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ and a vector $B = (b_1, \cdots, b_n)^\top \in \mathbb{R}^n$, it holds that*

(a) $n^{-1}B^\top X \to_p 0$ *if* $n^{-2}B^\top B \to 0$ *as* $n \to \infty$.

(b) $n^{-1}X^\top AX \to_p \sigma_X^2 \lim_{n\to\infty} n^{-1}tr(A)$ *if the limit exists, and* $n^{-2}tr(AA^\top) \to 0$ *as* $n \to \infty$.

(c) $(nT)^{-1}\sum_{t=1}^T B^\top \widetilde{\mathbb{Y}}_t \to_p 0$ *if* $n^{-1}\sum_{j=0}^\infty (B^\top G^j UU^\top (G^\top)^j B)^{1/2} \to 0$ *as* $n \to \infty$.

(d) $(nT)^{-1}\sum_{t=1}^{T}\widetilde{\mathbb{Y}}_t^\top A\widetilde{\mathbb{Y}}_t^\top \to_p \lim_{n\to\infty} n^{-1}tr\{A\Gamma(0)\}$ *if the limit exists, and* $n^{-1}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[tr\{U^\top(G^\top)^i AG^jUU^\top(G^\top)^jA^\top G^iU\}]^{1/2}\to 0$ *as* $n\to\infty$.

(e) $(nT)^{-1}\sum_{t=1}^{T}X^\top A\widetilde{\mathbb{Y}}_t^\top \to_p 0$ *if* $n^{-1}\sum_{j=0}^{\infty}[tr\{AG^jUU^\top(G^\top)^jA^\top\}]^{1/2}\to 0$ *as* $n\to\infty$.

Proof: The detailed proof can be found in Lemma 1 of Zhu et al. (2017).

Lemma 2. *Assume* $\min_k N_k = O(N^\delta)$ *and the stationary condition* $c_\beta < 1$, *where* $c_\beta = \max_k(|\beta_{1k}| + |\beta_{2k}|)$. *Further assume Conditions (C1)-(C3) hold. For matrices* $M_1 = (m_{ij}^{(1)}) \in \mathbb{R}^{n\times p}$ *and* $M_2 = (m_{ij}^{(2)}) \in \mathbb{R}^{n\times p}$, *define* $M_1 \preccurlyeq M_2$ *as* $m_{ij}^{(1)} \le m_{ij}^{(2)}$ *for* $1\le i\le n$ *and* $1\le j\le p$. *In addition, define* $|M|_e = (|m_{ij}|)\in\mathbb{R}^{n\times p}$ *for any arbitrary matrix* $M = (m_{ij})\in\mathbb{R}^{n\times p}$. *Then there exists* $J > 0$, *such that*

(a) for any integer $n > 0$, *we have*

$$|\mathcal{G}^n(\mathcal{G}^\top)^n|_e \preccurlyeq n^J c_\beta^{2n} MM^\top, \tag{A.1}$$

$$|\mathcal{G}^n\Sigma_Y|_e \preccurlyeq \alpha n^J c_\beta^{n} MM^\top, \tag{A.2}$$

where $M = C\mathbf{1}\pi^\top + \sum_{j=0}^{J}W^j$, $C > 1$ *is a constant,* π *is defined in (C3.1), and* α *is a finite constant.*

(b) For positive integers $k_1 \le 1$, $k_2\le 1$, *and* $j\ge 0$, *define* $g_{j,k_1,k_2}(\mathcal{G}, W^{(k)}) = |(W^{(k)})^{k_1}\{\mathcal{G}^j(\mathcal{G}^\top)^j\}^{k_2}$
$(W^{(k)\top})^{k_1}|_e \in \mathbb{R}^{N\times N}$. *In addition, define* $(W^{(k)})^0 = \mathcal{I}_k = (I_{N_k}, \mathbf{0})\in\mathbb{R}^{N_k\times N}$. *For integers* $0\le k_1,k_2,m_1,m_2\le 1$, *as* $N\to\infty$ *we have*

$$N^{-1}\sum_{j=0}^{\infty}\Big\{\mu^\top g_{j,k_1,k_2}(\mathcal{G},W^{(k)})\mu\Big\}^{1/2}\to 0, \tag{A.3}$$

$$N^{-1}\sum_{i,j=0}^{\infty}\Big[tr\Big\{g_{i,k_1,k_2}(\mathcal{G},W^{(k)})g_{j,m_1,m_2}(\mathcal{G},W^{(k)})\Big\}\Big]^{1/2}\to 0, \tag{A.4}$$

where $|\mu|_e \preccurlyeq c_\mu \mathbf{1}$ and c_μ is a finite constant.

(c) For integers $0 \le k_1, k_2 \le 1$, define $f_{k_1,k_2}(W^{(k)}, Q) = |(W^{(k)})^{k_1} Q^{k_2} (W^{(k)\top})^{k_1}|_e \in \mathbb{R}^{N\times N}$, where Q is given in (C3). Then for integers $0 \le k_1, k_2, m_1, m_2 \le 1$, as $N \to \infty$ we have

$$N^{-2}\mu^\top f_{k_1,k_2}(W^{(k)}, Q)\mu \to 0, \tag{A.5}$$

$$N^{-2}tr\Big\{f_{k_1,k_2}(W^{(k)}, Q) f_{m_1,m_2}(W^{(k)}, Q)\Big\} \to 0, \tag{A.6}$$

$$N^{-1}\sum_{j=0}^{\infty}\Big[tr\Big\{f_{k_1,k_2}(W^{(k)}, Q) g_{j,m_1,m_2}(\mathcal{G}, W^{(k)})\Big\}\Big]^{1/2} \to 0, \tag{A.7}$$

where $|\mu|_e \preccurlyeq c_\mu \mathbf{1}$ and c_μ is a finite constant.

Proof: The proof is similar in spirit to Zhu et al. (2017). Therefore, we give the guideline of the proof and skip some similar details. Without loss of generality, we let $c_\beta = |\beta_{11}| + |\beta_{21}|$ (i.e., $k = 1$). Consequently, we have $|\mathcal{G}|_e \preccurlyeq |\beta_{11}|W + |\beta_{21}|I$. Let $G = |\beta_{11}|W + |\beta_{21}|I$. Follow similar technique in part (a) in Lemma 2 of Zhu et al. (2017), it can be verified

$$|\mathcal{G}^n|_e \preccurlyeq n^J c_\beta^n M, \tag{A.8}$$

where $M = C\mathbf{1}\pi^\top + \sum_{j=0}^J W^j$ is defined in (a) of Lemma 2. Subsequently, the result (A.1) can be readily obtained. Next, recall that $\Sigma_Y = (I - \mathcal{G})^{-1}\Sigma_{\mathbb{Z}}(I - \mathcal{G}^\top)^{-1} + \sum_{j=0}^\infty \mathcal{G}^j \Sigma_e (\mathcal{G}^\top)^j = (\sum_{j=0}^\infty \mathcal{G}^j)\Sigma_{\mathbb{Z}}(\sum_{j=0}^\infty (\mathcal{G}^\top)^j) + \sum_{j=0}^\infty \mathcal{G}^j \Sigma_e (\mathcal{G}^\top)^j$. Let $\sigma_z^2 = \max_k\{\gamma_k^\top \Sigma_z \gamma_k\}$ and $\sigma_e^2 = \max_k\{\sigma_k^2\}$. Then we have $|\mathcal{G}^n \Sigma_Y|_e \preccurlyeq \sigma_z^2(\sum_{j=0}^\infty |\mathcal{G}^{n+j}|_e)(\sum_{j=0}^\infty |(\mathcal{G}^\top)^j|_e) + \sigma_e^2 \sum_{j=0}^\infty |\mathcal{G}^{n+j}|_e |(\mathcal{G}^\top)^j|_e$. Subsequently, (A.2) can by obtained by applying (A.8). Next, we give the proof of (b) in the following. The conclusion (c) can be proved by similar techniques, which is omitted here to save space.

Let $k_1 = k_2 = 1$. Then we have $g_{j,1,1}(\mathcal{G}, W^{(k)}) = |W^{(k)}\mathcal{G}^j \mathcal{G}^j W^{(k)\top}|$. Recall that

$W^{(k)} = (w_{ij} : i \in \mathcal{M}_k, 1 \leq j \leq N) \in \mathbb{R}^{N_k \times N}$. Since we have $|\mu|_e \preccurlyeq c_\mu \mathbf{1}$, then it suffices to show $\sum_{j=0}^\infty N_k^{-1}\big\{\mathbf{1}^\top g_{j,1,1}(\mathcal{G}, W^{(k)})\mathbf{1}\big\}^{1/2} \to 0$. We first prove (A.3). By (A.8) we have $|W^{(k)}\mathcal{G}^j|_e \preccurlyeq j^K(|\beta_1| + |\beta_2|)^j W^{(k)}M$. As a result, we have

$$|W^{(k)}\mathcal{G}^j(\mathcal{G}^\top)^j W^{(k)\top}|_e \preccurlyeq j^{2K}(|\beta_1| + |\beta_2|)^{2j}\mathcal{M}, \tag{A.9}$$

where $\mathcal{M}$ is defined as $\mathcal{M} = W^{(k)}MM^\top W^{(k)\top}$. As a result, we have $\sum_{j=0}^\infty N_k^{-1}\big\{\mathbf{1}^\top W^{(k)} \mathcal{G}^j(\mathcal{G}^\top)^j W^{(k)\top}\mathbf{1}\big\}^{1/2} \leq N_k^{-1}\alpha_1(\mathbf{1}^\top\mathcal{M}\mathbf{1})^{1/2}$, where $\alpha_1 = \sum_{j=0}^\infty j^K c_\beta^j < \infty$. Then it leads to show $N_k^{-2}\mathbf{1}^\top\mathcal{M}\mathbf{1} \to 0$. It can be shown $\mathbf{1}^\top\mathcal{M}\mathbf{1} = N_k^2 C\sum_j \pi_j^2 + \sum_{j=1}^K \mathbf{1}^\top W^{(k)}W^j(W^\top)^j W^{(k)\top}\mathbf{1} + 2N_k C\sum_j \pi^\top(W^\top)^j W^{(k)\top}\mathbf{1} + \sum_{i\neq j}\mathbf{1}^\top W^{(k)}W^i(W^\top)^j W^{(k)\top}\mathbf{1}$. For the last two terms of $\mathbf{1}^\top\mathcal{M}\mathbf{1}$, by Cauchy inequality, we have

$$N_k\sum_j \pi^\top(W^\top)^j W^{(k)\top}\mathbf{1} \;\; \leq \;\; N_k\Big(\sum_j \pi_j^2\Big)^{1/2}\Big\{\mathbf{1}^\top W^{(k)}W^j(W^\top)^j W^{(k)\top}\mathbf{1}\Big\}^{1/2},$$

and $\sum_{i\neq j}\mathbf{1}^\top W^{(k)}W^i(W^\top)^j W^{(k)\top}\mathbf{1} \leq \sum_{i\neq j}\big\{\mathbf{1}^\top W^{(k)}W^i(W^\top)^i W^{(k)\top}\mathbf{1}\big\}^{1/2}\big\{\mathbf{1}^\top W^{(k)}W^j (W^\top)^j W^{(k)\top}\mathbf{1}\big\}^{1/2}$. As a result, it leads to show

$$\sum_{j=1}^N \pi_j^2 \to 0 \quad \text{and} \quad N_k^{-2}\mathbf{1}^\top W^{(k)}W^j(W^\top)^j W^{(k)\top}\mathbf{1} \to 0 \tag{A.10}$$

for $1 \leq j \leq K+1$. As the first convergence in (A.10) is implied by (C2.1), we next prove $N_k^{-2}\mathbf{1}^\top W^{(k)}W^j(W^\top)^j W^{(k)\top}\mathbf{1} \to 0$ $(1 \leq j \leq K)$. Recall that $W^* = W + W^\top$. Therefore, we have $N_k^{-2}\mathbf{1}^\top W^{(k)}W^j(W^\top)^j W^{(k)\top}\mathbf{1} \leq N^{-2}\mathbf{1}^\top W^{(k)}W^{*2j}W^{(k)\top}\mathbf{1}$. Then it suffices to show $N^{-2}\mathbf{1}^\top W^{(k)}W^{*2j}W^{(k)\top}\mathbf{1} \to 0$. By eigenvalue-eigenvector decomposition of W^* we have $W^* = \sum_k \lambda_k(W^*)u_k u_k^\top$, where $\lambda_k(W^*)$ and $u_k \in \mathbb{R}^N$ are the kth eigenvalue and eigenvector of W^* respectively. As a result, we have $N_k^{-2}\mathbf{1}^\top W^{(k)}W^{*2j}W^{(k)\top}\mathbf{1} \leq N_k^{-2}\lambda_{\max}(W^*)^{2j}(\mathbf{1}^\top W^{(k)}W^{(k)\top}\mathbf{1})$ $(1 \leq j \leq K)$. Further

we have $\mathbf{1}^\top W^{(k)} W^{(k)\top} \mathbf{1} \le N_k \lambda_{\max}(W^{(k)} W^{(k)\top})$. Note that $W^{(k)} W^{(k)\top}$ is a sub-matrix of $WW^\top$ with row and column index in $\mathcal{M}_k$. Therefore, by Cauchy's interlacing Theorem, we have $\lambda_{\max}(W^{(k)}$
$W^{(k)\top}) \le \lambda_{\max}(WW^\top) = O(N^{\delta'})$ for $\delta' < \delta$. Since we have $\min_k N_k = N^\delta$ for $\delta > 0$, then we have $N_k^{-1} \lambda_{\max}(W^{(k)} W^{(k)\top}) \to 0$ as $N \to \infty$. As a consequence, the second term in (A.10) holds. Similarly, it can be proved that (A.10) holds for all $0 \le k_1, k_2 \le 1$. As a result, we have (A.3) holds.

We next prove (A.4) with $k_1 = k_2 = m_1 = m_2 = 1$, and $g_{i,1,1}(\mathcal{G}, W^{(k)}) g_{j,1,1}(\mathcal{G}, W^{(k)}) = |W^{(k)} \mathcal{G}^i (\mathcal{G}^\top)^i W^{(k)\top} W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j W^{(k)\top}|_e$. Then it can be similarly proved for other cases (i.e., $0 \le k_1, k_2, m_1, m_2 \le 1$). Note that by (A.9), we have

$$\Big[\mathrm{tr}\Big\{W^{(k)} \mathcal{G}^i (\mathcal{G}^\top)^i W^{(k)\top} W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j W^{(k)\top}\Big\}\Big]^{1/2} \le i^K j^K (|\beta_1| + |\beta_2|)^{i+j} \mathrm{tr}\big\{\mathcal{M}^2\big\}^{1/2}.$$

It then can be derived that $N_k^{-1} \sum_{i,j=0}^\infty [\mathrm{tr}\{W^{(k)} \mathcal{G}^i (\mathcal{G}^\top)^i W^{(k)\top} W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j W^{(k)\top}\}]^{1/2} \le \alpha^2 N_k^{-1} \mathrm{tr}\big\{\mathcal{M}^2\big\}^{1/2}$. In order to obtain (A.4), it suffices to show that

$$N_k^{-2} \mathrm{tr}\{\mathcal{M}^2\} \to 0. \tag{A.11}$$

Equivalently, by Cauchy inequality, it suffices to prove $(\sum \pi_j^2)^2 \to 0$, and $N_k^{-2} \mathrm{tr}\{W^{(k)} W^j$
${W^j}^\top W^{(k)\top} W^{(k)} W^j {W^j}^\top W^{(k)\top}\} \to 0$ holds for $1 \le j \le K$. It can be easily verified the first term holds by (C2.1). For the second one, we have $N_k^{-2} \mathrm{tr}\{W^{(k)} W^j {W^j}^\top W^{(k)\top} W^{(k)} W^j$
${W^j}^\top W^{(k)\top}\} \le N_k^{-2} \mathrm{tr}\{W^{(k)} (W^*)^{4j} W^{(k)\top}\} \le N_k^{-2} \lambda_{\max}(W^*)^{4j} \mathrm{tr}(W^{(k)} W^{(k)\top}) \le N_k^{-2} N_k$
$\lambda_{\max}(W^*)^{4K} \lambda_{\max}(WW^\top)$. Similarly, due to that $\lambda_{\max}(W^*) = O(\log N)$ and $\lambda_{\max}(WW^\top)$
$= O(N^{\delta'})$ in (C2.2), we have $N_k^{-1} \lambda_{\max}(W^*)^{4K} \lambda_{\max}(WW^\top) \to 0$ as $N \to \infty$. Consequently, we have (A.11) and then (A.4) holds. This completes the proof of (b).

Lemma 3. *Let $\{X_{it} : 1 \le t \le T\}$ and $\{Y_{it} : 1 \le t \le T\}$ be random sub-Gaussian time* *series with mean 0, var*$(X_{it}) = \sigma_{i,xx}$*, var*$(Y_{it}) = \sigma_{i,yy}$*, and cov*$(X_{it}, Y_{it}) = \sigma_{i,xy}$*. Let* $\sigma_{xi,t_1t_2} = cov(X_{it_1}, X_{it_2})$ *and* $\Sigma_{xi} = (\sigma_{xi,t_1t_2} : 1 \leq t_1, t_2 \leq T) \in \mathbb{R}^{T\times T}$*. Similarly, define* σ_{yi,t_1t_2} *and* $\Sigma_{yi} \in \mathbb{R}^{T\times T}$*. Then we have*

$$P\Big(\Big|T^{-1}\sum_{t=1}^{T} X_{it}Y_{it} - \sigma_{i,xy}\Big| > \nu\Big) \leq c_1\Big\{\exp(-c_2\sigma_{xi}^{-2}T^2\nu^2) + \exp(-c_2\sigma_{yi}^{-2}T^2\nu^2)\Big\} \quad (A.12)$$

for $|\nu| \leq \delta$*, where* $\sigma_{xi}^2 = tr(\Sigma_{xi}^2)$*,* $\sigma_{yi}^2 = tr(\Sigma_{yi}^2)$*,* c_1*,* c_2*, and* δ *are finite constants.*

Proof: Let $X_i = (X_{i1}, \cdots, X_{iT})^\top \in \mathbb{R}^T$ and $Y_i = (Y_{i1}, \cdots, Y_{iT})^\top \in \mathbb{R}^T$. In addition, let $Z_i = Z_i + Y_i$. Therefore, we have $Z_i^\top Z_i = 2^{-1}(Z_i^\top Z_i - X_i^\top X_i - Y_i^\top Y_i)$. It can be derived that

$$P\{|T^{-1}(X_i^\top Y_i) - \sigma_{i,xy}| \geq \nu\} \leq P\{|T^{-1}(Z_i^\top Z_i) - (\sigma_{i,xx} + \sigma_{i,yy} + 2\sigma_{i,xy})| \geq \nu_1\}$$
$$+ P\{|T^{-1}(X_i^\top X_i) - \sigma_{i,xx}| \geq \nu_1\} + P\{|T^{-1}(Y_i^\top Y_i) - \sigma_{i,yy}| \geq \nu_1\}, \quad (A.13)$$

where $\nu_1 = 2\nu/3$. Next, we derive the upper bound for the right side of (A.13). Note that $\mathbb{X}_i^\top \mathbb{X}_i$, $Y_i^\top Y_i$, and $Z_i^\top Z_i$ all take quadratic form. Therefore the proofs are similar. For the sake of simplicity, we take $Y_i^\top Y_i$ for an example and derive the upper bound for $P\{|n^{-1}(Y_i^\top Y_i) - \sigma_{i,yy}| \geq \nu_1\}$. Similar results can be obtained for the other two terms.

First we have $Y_i^\top Y_i = Y_i^\top \Sigma_{yi}^{-1/2}\Sigma_{yi}\Sigma_{yi}^{-1/2}Y_i = \widetilde{Y}_i^\top \Sigma_{yi}\widetilde{Y}_i$, where $\widetilde{Y}_i = \Sigma_{yi}^{-1/2}Y_i$ follows sub-Gaussian distribution. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_T$ be the eigenvalues of Σ_{yi}. Since Σ_{yi} is a non-negative definite matrix, The eigenvalue decomposition can be applied to obtain $\Sigma_{yi} = U^\top \Lambda U$, where $U = (U_1, \cdots, U_T)^\top \in \mathbb{R}^{T\times T}$ is an orthogonal matrix and $\Lambda = \text{diag}\{\lambda_1, \cdots, \lambda_T\}$. As a consequence, we have $Y_t^\top Y_t = \sum_t \lambda_t \zeta_t^2$, where $\zeta_t = U_t^\top \widetilde{Y}_t$ and ζ_ts are independent and identically distributed as standard sub-Gaussian. It can be verified $\zeta_t^2 - 1$ satisfies sub-exponential distribution and $T^{-1}(\sum_t \lambda_t) = \sigma_{i,yy}$. In addition, the sub-exponential distribution satisfies condition

(P) on page 45 of Saulis and Statuleviveccius (2012). There exists constants c_1, c_2, and δ such that $P\{|T^{-1}(Y_i^\top Y_i) - \sigma_{i,yy}| \geq \nu_1\} = P\{\sum_t \lambda_t(\zeta_t^2 - 1)| \geq T\nu_1\} \leq c_1 \exp\{-c_2(\sum_t \lambda_t^2)^{-1}T^2\nu^2\} = c_1 \exp\{-c_2\sigma_{yi}^{-1}T^2\nu^2\}$ for $|\nu| < \delta$ by the Theorem 3.3 of Saulis and Statuleviveccius (2012). Consequently, (A.12) can be obtained by appropriately chosen c_1, c_2, and δ.

Lemma 4. *Assume Y_{it} follows the GNAR model (2.3) and $|c_\beta| < 1$. Then there exists finite constants c_1, c_2, and δ, for $\nu < \delta$ we have*

$$P\Big\{\Big|T^{-1}\sum_{t=1}^T Y_{it}^2 - \mu_i^2 - e_i^\top \Sigma_Y e_i\Big| > \nu\Big\} \leq \delta_T, \tag{A.14}$$

$$P\Big\{\Big|T^{-1}\sum_{t=1}^T Y_{it}(w_i^\top \mathbb{Y}_t) - \mu_{Yi}(w_i^\top \mu_Y) - w_i^\top \Sigma_Y e_i\Big| > \nu\Big\} \leq \delta_T \tag{A.15}$$

$$P\Big\{\Big|T^{-1}\sum_{t=1}^T Y_{i(t-1)}\varepsilon_{it}\Big| > \nu\Big\} \leq \delta_T, \quad P\Big\{\Big|T^{-1}\sum_{t=1}^T (w_i^\top \mathbb{Y}_{t-1})\varepsilon_{it}\Big| > \nu\Big\} \leq \delta_T, \tag{A.16}$$

$$P\Big\{\Big|T^{-1}\sum_{t=1}^T Y_{i(t-1)} - \mu_i\Big| > \nu\Big\} \leq \delta_T, \quad P\Big\{\Big|T^{-1}\sum_{t=1}^T w_i^\top \mathbb{Y}_t - w_i^\top \mu_Y\Big| > \nu\Big\} \leq \delta_T, \tag{A.17}$$

where $\delta_T = c_1 \exp(-c_2 T\nu^2)$, $e_i \in \mathbb{R}^N$ is an N-dimensional vector with all elements being 0 but the ith element being 1, and $\mu_i = e_i^\top \mu_Y$.

Proof: For the similarity of proof procedure, we only prove (A.14) in the following. Without loss of generality, let $\mu_Y = \mathbf{0}$. Recall that the group information is denoted as $\boldsymbol{Z} = \{z_{ik} : 1 \leq i \leq N, 1 \leq k \leq K\}$. Define $P^*(\cdot) = P(\cdot|\boldsymbol{Z})$, $E^*(\cdot) = E(\cdot|\boldsymbol{Z})$, and $\mathrm{cov}^*(\cdot) = \mathrm{cov}(\cdot|\boldsymbol{Z})$. Write $\mathcal{Y}_i = (Y_{i1}, \cdots, Y_{iT})^\top \in \mathbb{R}^T$. Given $\boldsymbol{Z}$, $\mathcal{Y}_i$ is a sub-Gaussian random vector with $\mathrm{cov}(\mathcal{Y}_i) = \Sigma_i = (\sigma_{i,t_1t_2}) \in \mathbb{R}^{T\times T}$, where $\sigma_{i,t_1t_2} = e_i^\top \mathcal{G}^{t_1-t_2}\Sigma_Y e_i$ for $t_1 \geq t_2$, $\sigma_{i,t_1t_2} = e_i^\top \Sigma_Y(\mathcal{G}^\top)^{t_2-t_1}e_i$, and $\mathcal{G}$ is pre-defined in (2.5) as $\mathcal{G} = \mathcal{B}_1 W + \mathcal{B}_2$. It can be derived $\mathrm{var}^*(\mathcal{Y}_i^\top \mathcal{Y}_i) \leq c\mathrm{tr}(\Sigma_i^2)$, where c is a positive constant and $\mathrm{tr}(\Sigma_i^2) = T(e_i^\top \Sigma_Y e_i)^2 + 2\sum_{t=1}^{T-1}(T-t)(e_i^\top \mathcal{G}^t \Sigma_Y e_i)^2$. It can be derived $|\Sigma_Y|_e \preccurlyeq \alpha MM^\top$ and $|\mathcal{G}^t\Sigma_Y|_e \preccurlyeq \alpha_1 t^J c_\beta^t MM^\top$ by (A.2) of Lemma 2, where c_β, J and M are defined in Lemma

2, α and α_1 are finite constants. In addition, it can be verified $\sum_{t=1}^{T-1}(T-t)t^{2J}c_\beta^{2t} \le \alpha_2 T$, where α_2 is a finite constant. Therefore we have $\mathrm{tr}(\Sigma_i^2) \le T(\alpha+2\alpha_1\alpha_2)\{(e_i^\top MM^\top e_i)^2\}$. Since we have $e_i^\top MM^\top e_i \le (J+1)e_i^\top M\mathbf{1} \le (J+1)^2 = O(1)$, it can be concluded that $\mathrm{tr}(\Sigma_i^2) \le T\alpha_3$, where $\alpha_3 = (\alpha + 2\alpha_1\alpha_2)(J+1)^2$. By Lemma 3, the (A.14) can be obtained.

Appendix B. Proof of Theorem 1

Let $\lambda_i(M)$ be the ith eigenvalue of $M \in \mathbb{R}^{N\times N}$. We first verify that the solution (2.6) is strictly stationary. By Banerjee et al. (2014), we have $\max_i |\lambda_i(W)| \le 1$. Hence we have

$$\max_{1\le i\le N} |\lambda_i(\mathcal{G})| \le \big(\max_{1\le k\le K} |\beta_{1k}|\big)\big(\max_{1\le i\le N} |\lambda_i(W)|\big) + \max_{1\le k\le K} |\beta_{2k}| < 1. \tag{A.18}$$

Consequently, we have $\lim_{m\to\infty}\sum_{j=0}^{m}\mathcal{G}^j\mathcal{E}_{t-j}$ exists and $\{\mathbb{Y}_t\}$ given by (2.6) is a strictly stationary process. In addition, one could directly verify that $\{\mathbb{Y}_t\}$ satisfies the GNAR model (2.3).

Next, we verify that the strictly stationary solution (2.6) is unique. Assume $\{\widetilde{\mathbb{Y}}_t\}$ is another strictly stationary solution to the GNAR model (2.3) with $E\|\widetilde{\mathbb{Y}}_t\| < \infty$. Then we have $\widetilde{\mathbb{Y}}_t = \sum_{j=1}^{m-1}\mathcal{G}^j(\mathcal{B}_0 + \mathcal{E}_{t-j}) + \mathcal{G}^m\widetilde{\mathbb{Y}}_{t-m}$ for any positive integer m. Let $\rho = \max_k(|\beta_{1k}|+|\beta_{2k}|)$. Then one could verify $E\|\mathbb{Y}_t - \widetilde{\mathbb{Y}}_t\| = E\|\sum_{j=m}^{\infty}\mathcal{G}^j(\mathcal{B}_0+\mathcal{E}_{t-j}) - \mathcal{G}^m\widetilde{\mathbb{Y}}_{t-m}\| \le C\rho^m$, where C is a finite constant unrelated to t and m. Note that m can be chosen arbitrarily. As a result, we have that $E\|\mathbb{Y}_t - \widetilde{\mathbb{Y}}_t\| = 0$, i.e. $\mathbb{Y}_t = \widetilde{\mathbb{Y}}_t$ with probability one. This completes the proof.

Appendix C. Proof of Theorem 2

According to (3.2), $\hat{\theta}_k$ can be explicitly written as $\hat{\theta}_k = \theta_k + \widehat{\Sigma}_k^{-1}\widehat{\zeta}_k$, where $\widehat{\Sigma}_k = (N_kT)^{-1}\sum_{t=1}^{T}\mathbb{X}_{t-1}^{(k)\top}\mathbb{X}_{t-1}^{(k)}$ and $\widehat{\zeta}_k = (N_kT)^{-1}\sum_{t=1}^{T}\mathbb{X}_{t-1}^{(k)\top}\mathcal{E}_t^{(k)}$. Without loss of generality, we assume $\sigma_k^2 = 1$ for $k = 1, \cdots, K$. Let $\Sigma_k = \lim_{N\to\infty} E(\widehat{\Sigma}_k)$. As a result, it suffices to show that

$$\widehat{\Sigma}_k \to_p \Sigma_k, \tag{A.19}$$

$$\sqrt{N_kT}\widehat{\zeta}_k = O_p(1), \tag{A.20}$$

as $\min\{N, T\} \to \infty$. Subsequently, we prove (A.19) in Step 1 and (A.20) in Step 2.

STEP 1. PROOF OF (A.19). Define $Q = (I - \mathcal{G})^{-1}\Sigma_{\mathbb{V}}(I - \mathcal{G}^\top)^{-1}$. In this step, we intend to show that $\widehat{\Sigma}_k =$

$$\frac{1}{N_kT}\sum_{t=1}^{T}\mathbb{X}_{t-1}^{(k)\top}\mathbb{X}_{t-1}^{(k)} = \begin{pmatrix} 1 & \mathbb{S}_{12} & \mathbb{S}_{13} & \mathbb{S}_{14} \\ & \mathbb{S}_{22} & \mathbb{S}_{23} & \mathbb{S}_{24} \\ & & \mathbb{S}_{33} & \mathbb{S}_{34} \\ & & & \mathbb{S}_{44} \end{pmatrix} \to_p \begin{pmatrix} 1 & c_{1\beta} & c_{2\beta} & \mathbf{0}^\top \\ & \Sigma_1 & \Sigma_2 & \kappa_8\gamma^\top\Sigma_z \\ & & \Sigma_3 & \kappa_3\gamma^\top\Sigma_z \\ & & & \Sigma_z \end{pmatrix} = \Sigma_k,$$

where

$$\mathbb{S}_{12} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} w_i^\top\mathbb{Y}_{t-1}, \quad \mathbb{S}_{13} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} Y_{i(t-1)}, \quad \mathbb{S}_{14} = \frac{1}{N_k}\sum_{i\in\mathcal{M}_k} V_i^\top,$$

$$\mathbb{S}_{22} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} (w_i^\top\mathbb{Y}_{t-1})^2, \quad \mathbb{S}_{23} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} w_i^\top\mathbb{Y}_{t-1}Y_{i(t-1)},$$

$$\mathbb{S}_{24} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} w_i^\top\mathbb{Y}_{t-1}V_i^\top, \quad \mathbb{S}_{33} = \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{M}_k} Y_{i(t-1)}^2,$$

$\mathbb{S}_{34} = (N_k T)^{-1} \sum_{t=1}^T \sum_{i \in \mathcal{M}_k} Y_{i(t-1)} V_i^\top$, $\mathbb{S}_{44} = N_k^{-1} \sum_{i \in \mathcal{M}_k} V_i V_i^\top$. By (2.6), we have

$$\mathbb{Y}_t = (I - \mathcal{G})^{-1} b_0 + (I - \mathcal{G})^{-1} b_v + \widetilde{\mathbb{Y}}_t, \tag{A.21}$$

where $b_0 = \sum_k D_k B_{0k}$, $b_v = \sum_k D_k \mathbb{V} \gamma_k$, and $\widetilde{\mathbb{Y}}_t = \sum_{j=0}^\infty \mathcal{G}^j \mathcal{E}_{t-j}$. By the law of large numbers, one could directly obtain that $\mathbb{S}_{44} \to_p \Sigma_v$ and $\mathbb{S}_{14} \to_p \mathbf{0}^\top$. Subsequently, we only show the convergence of $\mathbb{S}_{12}$ and $\mathbb{S}_{23}$ in $\widehat{\Sigma}_k$ as follows.

Convergence of $\mathbb{S}_{12}$. It can be derived that

$$\mathbb{S}_{12} = \frac{1}{N_k T} \sum_{t=1}^T \mathbf{1}^\top W^{(k)} \mathbb{Y}_{t-1} = \frac{\mathbf{1}^\top W^{(k)} \mu_Y}{N_k} + \mathbb{S}_{12a} + \mathbb{S}_{12b},$$

where $\mathbb{S}_{12a} = N_k^{-1} \mathbf{1}^\top W^{(k)} (I - \mathcal{G})^{-1} b_v$ and $\mathbb{S}_{12b} = (N_k T)^{-1} \sum_{t=1}^T \mathbf{1}^\top W^{(k)} \widetilde{\mathbb{Y}}_{t-1}$. Then by (A.5) and (A.3) in Lemma 2, we have $N_k^{-2} \mathbf{1}^\top W^{(k)} Q W^{(k)\top} \mathbf{1} \to 0$ and $N_k^{-1} \sum_{j=0}^\infty \big\{ \mathbf{1}^\top W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j W^{(k)\top} \mathbf{1} \big\}^{1/2} \to 0$, as $N \to \infty$. As a result, it is implied by Lemma 1 (a) and (c) that $\mathbb{S}_{12a} \to_p 0$ and $\mathbb{S}_{12b} \to_p 0$.

Convergence of $\mathbb{S}_{23}$. Note that

$$\begin{aligned} \mathbb{S}_{23} &= \frac{1}{N_k T} \sum_{t=1}^T \sum_{i \in \mathcal{M}_k} w_i^\top \mathbb{Y}_{t-1} Y_{i(t-1)} = \frac{1}{N_k T} \sum_{t=1}^T \mathbb{Y}_{t-1}^{(k)\top} W^{(k)} \mathbb{Y}_{t-1} \\ &= \frac{\mu_Y^{(k)\top} W^{(k)} \mu_Y}{N_k} + \mathbb{S}_{23a} + \mathbb{S}_{23b} + \mathbb{S}_{23c} + \mathbb{S}_{23d} + \mathbb{S}_{23e}, \end{aligned}$$

where $\mathbb{S}_{23a} = N_k^{-1} \widetilde{b}_v^\top \mathcal{I}_k^\top W^{(k)} \widetilde{b}_v$, $\mathbb{S}_{23b} = N_k^{-1} T^{-1} \sum_{t=1}^T \widetilde{\mathbb{Y}}_{t-1}^{(k)\top} W^{(k)} \widetilde{\mathbb{Y}}_{t-1}$ and $\mathbb{S}_{23c} = N_k^{-1} T^{-1} \sum_{t=1}^T (\widetilde{b}_v^\top \mathcal{I}_k^\top W^{(k)} \widetilde{\mathbb{Y}}_{t-1} + \widetilde{\mathbb{Y}}_{t-1}^\top \mathcal{I}_k^\top W^{(k)} \widetilde{b}_v)$, $\mathbb{S}_{23d} = N_k^{-1} (\widetilde{b}_v^\top \mathcal{I}_k^\top \widetilde{\mu}_Y + \mu_Y^\top \mathcal{I}_k^\top \widetilde{b}_v)$, $\mathbb{S}_{23e} = N_k^{-1} T^{-1} \sum_{t=1}^T (\mathbb{Y}_{t-1}^{(k)\top} \widetilde{\mu}_Y + \mu_Y^\top \mathcal{I}_k^\top W^{(k)} \mathbb{Y}_{t-1})$, where $\widetilde{\mu}_Y = W^{(k)} \mu_Y$ and $\widetilde{b}_v = (I - \mathcal{G})^{-1} b_v$.

We next look at the terms one by one. First we have $N_k^{-2} \mathrm{tr}(\mathcal{I}_k Q \mathcal{I}_k^\top W^{(k)} Q W^{(k)\top}) \to 0$ by (A.6) in Lemma 2 (c). Therefore, by (b) in Lemma 1, we have $\mathbb{S}_{23a} \to_p s_{23a}$, where

$s_{23a} = \lim_{N_k \to \infty} E(\mathbb{S}_{23a})$. Next, for $\mathbb{S}_{23b}$ we have $N_k^{-1} \sum_{i,j=0}^{\infty} \mathrm{tr}\{(\mathcal{I}_k \mathcal{G}^i (\mathcal{G}^\top)^i \mathcal{I}_k^\top W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j$ $W^{(k)\top}\} \to 0$ by (A.4) in Lemma 2 (b). Therefore, by (d) in Lemma 1, we have $\mathbb{S}_{23b} \to_p s_{23b}$, where $s_{23b} = \lim_{N_k \to \infty} E(\mathbb{S}_{23b})$. Next, let $\mathbb{S}_{23c} = \mathbb{S}_{23c}^{(1)} + \mathbb{S}_{23c}^{(2)}$, where $\mathbb{S}_{23c}^{(1)} = N_k^{-1} T^{-1} \sum_{t=1}^{T} \widetilde{b}_z^\top \mathcal{I}_k^\top W^{(k)} \widetilde{\mathbb{Y}}_{t-1}$ and $\mathbb{S}_{23c}^{(2)} = N_k^{-1} T^{-1} \sum_{t=1}^{T} \widetilde{\mathbb{Y}}_{t-1}^\top \mathcal{I}_k^\top W^{(k)} \widetilde{b}_v$. Note that we have $N_k^{-1} \sum_{j=0}^{\infty} \mathrm{tr}\{W^{(k)} \mathcal{G}^j (\mathcal{G}^\top)^j W^{(k)\top} \mathcal{I}_k Q \mathcal{I}_k^\top\} \to 0$ and $N_k^{-1} \sum_{j=0}^{\infty} \mathrm{tr}\{\mathcal{I}_k \mathcal{G}^j (\mathcal{G}^\top)^j \mathcal{I}_k^\top W^{(k)}$ $Q W^{(k)\top}\} \to 0$ by (A.7) in Lemma 2 (c). Therefore, $\mathbb{S}_{23c} \to_p s_{23c}$ by (e) in Lemma 1, where $s_{23c} = \lim_{N_k \to \infty} E(\mathbb{S}_{23c})$. Next, by similar proof to the convergence of $\mathbb{S}_{13}$, we have that $\mathbb{S}_{23d} \to_p 0$ and $\mathbb{S}_{23e} \to_p 0$. As a consequence, we have $\mathbb{S}_{23} \to_p \Sigma_2$.

Step 2. Proof of (A.20). It can be verified that $\sqrt{N_k T} E(\widehat{\zeta}_k) = 0$. In addition, we have $\mathrm{var}\{\sqrt{N_k T} \widehat{\zeta}_k\} = E(\widehat{\Sigma}_k) \to \Sigma_k$ as $N_k \to \infty$. Consequently, we have $\sqrt{N_k T} \widehat{\zeta}_k = O_p(1)$.

Appendix D. Proof of Theorem 3

Let $\widehat{\Sigma}_x^{(i)} = T^{-1} \sum_{t=1}^{T} \boldsymbol{X}_{i(t-1)} \boldsymbol{X}_{i(t-1)}^\top = (\hat{\sigma}_{x,ij}) \in \mathbb{R}^{3\times 3}$, and $\widehat{\Sigma}_{xe}^{(i)} = T^{-1}(\sum_{t=1}^{T} \boldsymbol{X}_{i(t-1)} \delta_i$ $\varepsilon_{it})$. We then have

$$\hat{b}_i - b_i = (\widehat{\Sigma}_x^{(i)})^{-1} \Sigma_{xe}^{(i)}.$$

Let $\widehat{\Sigma}_x^{(i)} = (\hat{\sigma}_{x,j_1 j_2} : 1 \le l_1, l_2 \le 3) \in \mathbb{R}^{3\times 3}$, where the index i of $\hat{\sigma}_{x,l_1 l_2}$ is omitted. Specifically, $\hat{\sigma}_{x,11} = 1$, $\hat{\sigma}_{x,12} = T^{-1} \sum_t w_i^\top \mathbb{Y}_{t-1}$, $\hat{\sigma}_{x,13} = T^{-1} \sum_t e_i^\top \mathbb{Y}_{t-1}$, $\hat{\sigma}_{x,22} = T^{-1} \sum_t Y_{i(t-1)}^2$, $\hat{\sigma}_{x,23} = T^{-1} \sum_t Y_{i(t-1)} (w_i^\top \mathbb{Y}_{t-1})$, $\hat{\sigma}_{x,33} = T^{-1} \sum_t (w_i^\top \mathbb{Y}_{t-1})^2$. Mathematically, it can be computed $(\widehat{\Sigma}_x^{(i)})^{-1} = |\widehat{\Sigma}_x^{(i)}|^{-1} \widehat{\Sigma}_x^{*(i)}$, where $|\widehat{\Sigma}_x^{(i)}|$ is the determinant of $\widehat{\Sigma}_x^{(i)}$, and $\widehat{\Sigma}_x^{*(i)}$ is the adjugate matrix of $\widehat{\Sigma}_x^{(i)}$, and $\Sigma_x^{*(i)} = (\hat{\sigma}_{x,l_1 l_2}^*)$, where $\hat{\sigma}_{x,11}^* = \hat{\sigma}_{x,22} \hat{\sigma}_{x,33} - \hat{\sigma}_{x,23}^2$, $\hat{\sigma}_{x,12}^* = \hat{\sigma}_{x,13} \hat{\sigma}_{x,32} - \hat{\sigma}_{x,12} \hat{\sigma}_{x,33}$ $\hat{\sigma}_{x,13}^* = \hat{\sigma}_{x,21} \hat{\sigma}_{x,32} - \hat{\sigma}_{x,22} \hat{\sigma}_{x,31}$, $\hat{\sigma}_{x,22}^* = \hat{\sigma}_{x,11} \hat{\sigma}_{x,33} - \hat{\sigma}_{x,13}^2$, $\hat{\sigma}_{x,23}^* = \hat{\sigma}_{x,13} \hat{\sigma}_{x,32} - \hat{\sigma}_{x,12} \hat{\sigma}_{x,33}$, and $\hat{\sigma}_{x,33}^* = \hat{\sigma}_{x,11} \hat{\sigma}_{x,22} - \hat{\sigma}_{x,12}^2$. It can be derived $|\widehat{\Sigma}_x^{(i)}| = \hat{\sigma}_{x,11}(\hat{\sigma}_{x,22} \hat{\sigma}_{x,33} - \hat{\sigma}_{x,23}^2) - \hat{\sigma}_{x,12}(\hat{\sigma}_{x,12} \hat{\sigma}_{33} - \hat{\sigma}_{13} \hat{\sigma}_{23}) + \hat{\sigma}_{13}(\hat{\sigma}_{12} \hat{\sigma}_{23} - \hat{\sigma}_{22} \hat{\sigma}_{13})$. By the maximum inequality, we have

$$P(\sup_i \|\hat{b}_i - b_i\| > \nu) \leq \sum_{i=1}^{N} P(\|\hat{b}_i - b_i\| > \nu). \tag{A.22}$$

In addition, we have

$$P(\|\hat{b}_i - b_i\| > \nu) \leq P\big(\big||\widehat{\Sigma}_x^{(i)}| - \sigma_x^{(i)}\big| \geq \delta_i\big) + P\big(\big|\widehat{\Sigma}_x^{*(i)}\widehat{\Sigma}_{xe}^{(i)}\big| \geq \delta_i\nu\big), \tag{A.23}$$

where $\sigma_x^{(i)} = \sigma_{x,11}(\sigma_{x,22}\sigma_{x,33} - \sigma_{x,23}^2) - \sigma_{x,12}(\sigma_{x,12}\sigma_{33} - \sigma_{13}\sigma_{23}) + \sigma_{13}(\sigma_{12}\sigma_{23} - \sigma_{22}\sigma_{13}) = (e_i^\top \Sigma_Y e_i)(w_i^\top \Sigma_Y w_i) - (e_i^\top \Sigma_Y w_i)^2$, $\delta_i = \sigma_x^{(i)}/2$. By lemma 4, for each component of $|\widehat{\Sigma}_x^{(i)}|$ we have $P(|\hat{\sigma}_{x,l_1l_2} - \sigma_{x,l_1l_2}| > \nu_0) \leq c_1 \exp(-c_2 T \nu_0^2)$, where $\sigma_{x,l_1l_2} = E(\hat{\sigma}_{x,l_1l_2})$ and ν_0 is a finite positive constant. Moreover, by the conditions of Theorem 3, we have $\sigma_x^{(i)} \geq \tau$ with probability tending to 1. Consequently, it is not difficult to obtain the result $P(||\widehat{\Sigma}_x^{(i)}| - \sigma_x^{(i)}| \geq \delta_i) \leq c_1^* \exp(-c_2^* T\tau^2)$, where c_1^*, c_2^* are finite constants. Subsequently, we have $P(|\widehat{\Sigma}_x^{*(i)}\widehat{\Sigma}_{xe}^{(i)}| \geq \delta_i\nu) \leq P(|\widehat{\Sigma}_x^{*(i)}\widehat{\Sigma}_{xe}^{(i)}| \geq \tau\nu/2)$. By similar technique, one could verify that each element of $\widehat{\Sigma}_x^{*(i)}$ and $\widehat{\Sigma}_{xe}^{(i)}$ converge with probability and the tail probability can be controlled, where the basic results are given in Lemma 4. Consequently, there exists constants c_3^* and c_4^* such that $P(|\widehat{\Sigma}_x^{*(i)}\widehat{\Sigma}_{xe}^{(i)}| \geq \tau\nu/2) \leq c_3^* \exp(-c_4^* T\tau^2\nu^2)$. Consequently, we have $P(\|\hat{b}_i - b_i\| > \nu) \leq c_1^* \exp(-c_2^* T\tau^2) + c_3^* \exp(-c_4^* T\tau^2\nu^2)$ by (A.23). By the condition $N = o(\exp(T))$, the right side of (A.22) goes to 0 as $N \to \infty$. This completes the proof.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, Crc Press.

Barabási, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks," *science*, 286, 509–512.

Bauwens, L. and Rombouts, J. (2007), "Bayesian clustering of many GARCH models," *Econometric Reviews*, 26, 365–386.

Bohn, A., Buchta, C., Hornik, K., and Mair, P. (2014), "Making friends and communicating on Facebook: Implications for the access to social capital," *Social Networks*, 37, 29–41.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009), "Power-law distributions in empirical data," *SIAM review*, 51, 661–703.

Fröhwirth-Schnatter, S. and Kaufmann, S. (2008), "Model-based clustering of multiple time series," *Journal of Business & Economic Statistics*, 26, 78–89.

Frühwirth-Schnatter, S. and Kaufmann, S. (2006), "How do changes in monetary policy affect bank lending? An analysis of Austrian bank data," *Journal of Applied econometrics*, 21, 275–305.

Hartigan, J. A. (1981), "Consistency of Single Linkage for High-Density Clusters," *Journal of the American Statistical Association*, 76, 388–394.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006), "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves," *Journal of the American Statistical Association*, 101, 18–29.

Hofstra, B., Corten, R., and Buskens, V. (2015), "Learning in social networks : S-electing profitable choices among alternatives of uncertain profitability in various networks," *Social Networks*, 43, 100–112.

Juárez, M. A. and Steel, M. F. (2010), "Model-based clustering of non-Gaussian panel data based on skew-t distributions," *Journal of Business & Economic Statistics*, 28, 52–66.

Lam, C. and Yao, Q. (2012), "Factor modeling for high-dimensional time series: inference for the number of factors," *Annals of Statistics*, 40, 694–726.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. A. (2008), "Tastes, ties, and time: A new social network dataset using Facebook.com," *Social Networks*, 30, 330–342.

Luan, Y. and Li, H. (2003), "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics*, 19, 474–482.

Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic block structures," *Journal of the American Statistical Association*, 96, 1077–1087.

Pan, J. and Yao, Q. (2008), "Modelling multiple time series via common factors," *Biometrika*, 95, 365–379.

Pollard, D. (1981), "Strong Consistency of K-Means Clustering," *Annals of Statistics*, 9, 135–140.

Saulis, L. and Statuleviveccius, V. (2012), *Limit theorems for large deviations*, vol. 73, Springer Science & Business Media.

Von Luxburg, U., Belkin, M., and Bousquet, O. (2008), "Consistency of spectral clustering," *Annals of Statistics*, 36, 555–586.

Wang, Y., Tsay, R. S., Ledolter, J., and Shrestha, K. M. (2013), "Forecasting Simultaneously High-Dimensional Time Series: A Robust Model-Based Clustering Approach," *Journal of Forecasting*, 32, 673–684.

Wang, Y. J. and Wong, G. Y. (1987), "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, 82, 8–19.

Wasserman, S. and Faust, K. (1994), *Social network analysis: Methods and applications*, vol. 8, Cambridge university press.

Zhao, Y., Levina, E., and Zhu, J. (2012), "Consistency of community detection in networks under degree-corrected stochastic block models," *The Annals of Statistics*, 40, 2266–2292.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017), "Network vector autoregression," *Annals of statistics*, 45, 1096–1123.

Table 1: Parameter Setup for Examples 1–3 in the Simulation Study.

	α	β_0	β_1	β_2	γ
			Example 1 & 2		
Group 1	0.2	0.0	0.1	0.3	$(0.5, 0.7, 1.0, 1.5, -1.0)^\top$
Group 2	0.3	0.2	-0.3	0.2	$(0.1, 0.9, 0.4, -0.2, -1.5)^\top$
Group 3	0.5	0.5	0.2	0.7	$(0.2, -0.2, 1.4, -0.8, 0.5)^\top$
			Example 3		
Group 1	0.2	5.0	0.2	0.1	$(0.5, 0.7, 1.0, 1.5, -1.0)^\top$
Group 2	0.3	-5.0	-0.4	0.2	$(0.1, 0.9, 0.4, -0.2, -1.5)^\top$
Group 3	0.5	0.0	0.2	0.4	$(0.2, -1.0, 2.0, 3.0, -2.0)^\top$

Table 2: Simulation Results with 1000 Replications for the stochastic block model. The RMSE ($\times 10^2$) are reported for the EM and TS estimation respectively. The network density (ND) and the misclassification rate (MCR) is also reported in percent (%).

N	Est.	α	β_0	β_1	β_2	γ	ND	MCR
Scenario 1. $T = N/2$								
100	EM	3.63	30.80	10.96	14.56	49.64	2.2	11.1
	TS	8.92	110.00	28.13	38.91	175.10	2.2	42.4
200	EM	2.10	14.86	6.42	11.09	26.54	1.1	3.8
	TS	7.56	46.74	22.19	34.66	75.44	1.1	31.3
500	EM	0.82	7.07	3.06	5.71	11.04	0.4	0.9
	TS	6.72	19.00	12.56	22.58	48.59	0.4	14.7
Scenario 2. $T = 2N$								
100	EM	4.08	41.67	12.24	17.60	56.03	2.2	13.3
	TS	6.65	37.43	13.86	21.51	60.08	2.2	15.0
200	EM	2.49	17.37	6.90	12.48	30.03	1.1	4.7
	TS	4.49	12.33	7.20	11.57	28.34	1.1	4.8
500	EM	1.04	8.82	3.19	6.76	13.95	0.4	1.1
	TS	1.42	3.84	1.42	2.31	7.16	0.4	0.3

Table 3: Simulation Results with 1000 Replications for the power-law model. The RMSE ($\times 10^2$) are reported for the EM and TS estimation respectively. The network density (ND) and the misclassification rate (MCR) is also reported in percent (%).

N	Est.	α	β_0	β_1	β_2	γ	ND	MCR
				Scenario 1. $T = N/2$				
100	EM	3.21	28.42	9.69	12.75	43.40	2.3	9.4
	TS	14.22	72.19	39.86	35.14	116.84	2.3	32.0
200	EM	1.74	13.15	5.67	9.86	23.44	1.2	3.5
	TS	12.08	34.17	27.13	27.83	64.49	1.2	18.0
500	EM	0.78	5.94	2.67	5.55	11.00	0.5	0.8
	TS	7.15	15.46	12.04	13.17	32.13	0.5	4.5
				Scenario 2. $T = 2N$				
100	EM	3.79	36.09	11.19	16.27	50.06	2.3	12.0
	TS	6.15	14.07	10.01	13.95	30.63	2.3	4.4
200	EM	2.33	17.64	6.65	11.67	27.50	1.2	4.7
	TS	2.99	6.20	4.00	6.14	14.08	1.2	0.9
500	EM	0.74	5.70	2.42	4.92	10.37	0.5	0.7
	TS	0.02	0.35	0.12	0.39	0.64	0.5	0.0

Table 4: Simulation Results with 500 Replications with different Ks (number of groups) for the power-law distribution network. The true number of groups is setted to be $K = 3$. The median values of $\mathrm{Err}_{est}^{(K)}$ $(\times 10^2)$ and $\mathrm{Err}_{pred}^{(K)}$ are reported respectively.

		Estimation					Prediction				
N	Est.	$K = 1$	$K = 2$	$K = 3$	$K = 5$	$K = 7$	$K = 1$	$K = 2$	$K = 3$	$K = 5$	$K = 7$
Scenario 1. $T = N/2$											
100	EM	147.7	111.4	69.1	22.1	25.4	2.48	2.29	2.10	2.02	2.02
	TS	147.7	136.7	129.1	118.7	109.6	2.48	2.42	2.37	2.32	2.28
200	EM	148.0	112.3	8.3	10.0	11.1	2.49	2.29	2.01	2.00	2.00
	TS	148.0	122.2	109.8	95.3	86.6	2.49	2.34	2.29	2.22	2.18
500	EM	148.4	113.4	2.8	3.4	3.9	2.49	2.30	2.00	2.00	2.00
	TS	148.4	105.2	50.1	41.6	38.3	2.49	2.26	2.06	2.04	2.03
Scenario 2. $T = 2N$											
100	EM	147.8	112.2	86.5	10.9	11.3	2.48	2.29	2.16	2.03	2.01
	TS	147.8	104.0	54.8	36.6	26.5	2.48	2.25	2.07	2.04	2.02
200	EM	148.2	112.5	3.8	4.4	4.8	2.49	2.29	2.01	2.01	2.00
	TS	148.2	103.8	3.8	5.5	7.0	2.49	2.25	2.01	2.00	2.00
500	EM	148.2	113.2	1.4	1.7	2.3	2.49	2.29	2.00	2.01	2.02
	TS	148.2	104.1	1.4	2.2	2.9	2.49	2.25	2.00	2.00	2.00

Table 5: The detailed GNAR analysis results for the Sina Weibo dataset.

Regression coefficient	Group 1	Group 2	Group 3
Group Ratio (α)	0.447	0.361	0.192
Baseline Effect (β_0)	0.857	1.681	0.236
Network Effect (β_1)	0.031	0.026	0.002
Momentum Effect (β_2)	0.765	0.396	0.958
Gender (γ_1)	0.077	0.155	0.009
Number of Labels (γ_2)	0.006	0.018	0.002

Table 6: The prediction RMSE for $\mathrm{PM}_{2.5}$ dataset using GNAR model (with EM and TS estimation respectively), NAR model, AR model.

	GNAR (EM)	GNAR (TS)	NAR	AR
Spring	0.375	0.387	0.388	0.739
Summer	0.328	0.328	0.330	0.941
Autumn	0.439	0.439	0.441	1.122
Winter	0.546	0.565	0.561	0.955

Table 7: Estimation results of the $\mathrm{PM}_{2.5}$ dataset by EM algorithm. Two groups are set for spring, summer, and autumn. While in winter, the number of groups is chosen to be $K = 3$.

	Spring		Summer		Autumn		Winter		
Group	1	2	1	2	1	2	1	2	3
Group Ratio (α)	0.61	0.39	0.67	0.33	0.53	0.47	0.30	0.12	0.58
Baseline Effect (β_0)	1.26	0.77	0.46	0.55	0.25	0.41	1.80	1.39	0.20
Network Effect (β_1)	0.14	0.11	0.20	-0.04	0.32	0.11	0.16	0.05	0.20
Momentum Effect (β_2)	0.55	0.65	0.67	0.87	0.62	0.76	0.43	0.57	0.74