

Statistica Sinica Preprint No: SS-2022-0040

Title	Grouped Network Poisson Autoregressive Model
Manuscript ID	SS-2022-0040
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0040
Complete List of Authors	Yuxin Tao, Dong Li and Xiaoyue Niu
Corresponding Authors	Xiaoyue Niu
E-mails	xiaoyue@psu.edu
Notice: Accepted version subject to English editing.	

Grouped Network Poisson Autoregressive Model

Yuxin Tao¹, Dong Li¹, and Xiaoyue Niu²

¹*Tsinghua University*, ²*The Pennsylvania State University*

Abstract: Multivariate Poisson autoregressive models are common ways to fit count time series data, while the statistical inference is quite challenging. The network Poisson autoregressive model (NPAR) reduces the inference complexity by incorporating network information into the dependence structure, where the response of each individual can be explained by its lagged values and the average effect of its neighbors. However, NPAR makes one strong assumption that all individuals are homogeneous and they share a common autoregressive coefficient. Here we propose a grouped network Poisson autoregressive model (GNPAR), where the individuals are classified into different groups with group-specific parameters to describe heterogeneous nodal behaviors. We present the stationarity and ergodicity of the GNPAR model and study the asymptotic properties of the maximum likelihood estimation. We develop an EM algorithm to estimate the unknown group labels and investigate the finite-sample performance of our estimation procedure using simulations. We analyze the Chicago Police Investigatory Stop Report data and find distinct dependence patterns in different neighborhoods of Chicago that could be potentially helpful for future crime prevention.

Key words and phrases: EM algorithm, Individual heterogeneity, Maximum likelihood estimation, Multivariate Poisson autoregression, Network data.

1. Introduction

Count time series data are often observed in practice. The monograph of Weiß (2018) summarizes the development of count time series analysis, and Davis, et al. (2021) gives a comprehensive methodological review. Count time series has its unique features, including integer-valued, over-dispersed, zero-inflated, time-dependent, and nonnegative autocorrelation. In the existing literature, most of earlier work on count time series modeling primarily focus on univariate cases. For example, Du and Li (1991) proposes an integer-valued AR model; Ferland, Latour and Oraichi (2006) introduces an integer-valued GARCH model, also called Poisson autoregression (PAR); Many others study variants of PAR with their statistical inference and applications (see, e.g., Fokianos, Rahbek and Tjøstheim (2009), Fokianos and Tjøstheim (2011, 2012), Neumann (2011), Wang, et al. (2014), Ahmad and Francq (2016), and Davis and Liu (2016)). On the other hand, multivariate count time series has not been studied much in the past, with only a few theoretical results available (see Latour (1997), Liu (2012), Pedeli and Karlis (2013), Andreassen (2013), Lee, Lee and Tjøstheim (2018)), although it has a range of important applications in many fields, such as environmental science, sociology, finance, marketing, and medicine, among others (Mahamunulu (1967), Aitchison and Ho (1989), Karlis and Meligkotsidou (2005, 2007), Weiß (2018), Fokianos, et al. (2020), Davis, et al. (2021)).

For multivariate PAR model, maximum likelihood based statistical inference is

quite challenging since the probability mass function of a multivariate Poisson random vector is usually of complicated functional form. To circumvent such difficulties, Fokianos, et al. (2020) presents a copula method and develops a novel conceptual framework to tackle multivariate count time series. To reduce the complex structure of multivariate count time series model, in the spirit of Zhu, et al. (2017), Armillotta and Fokianos (2021) proposes a network PAR (PNAR) model via incorporating network structure into multiple or high-dimensional PAR. Such an important technique is widely used to reduce model complexity in many settings, see, e.g., Zhu, et al. (2019a,b, 2020), Huang, et al. (2020), Zhou, et al. (2020), and Zhu, Cai and Ma (2021).

The PNAR model assumes that all individuals share a common dependence structure. This homogeneous assumption is often too stringent in practice. For example, if we consider the district-level crime cases in Chicago, we will find that there are districts where crimes frequently occur and there are relatively safer districts. So it is unreasonable to assume that the data-generating mechanism for all districts is the same. Further, the intensity process in PNAR models only regresses on the lagged observations.

In this paper, we propose a grouped NPAR model (GNPAR), where individuals are classified into different groups with group-specific parameters to describe heterogeneous nodal behaviors. Such an extension is of both theoretical and method-

ological importance, since it reduces the computational complexity from the general multivariate PAR models while providing more realistic and flexible setup and interpretation over PNAR models. Moreover, our model involves the lags of the intensity process in addition to lagged observations and the average effect of its neighbors, which allows more flexibility.

The remainder of the paper is organized as follows. Section 2 proposes a GNPARG model and gives its stationarity and ergodicity condition. Section 3 studies the maximum likelihood estimation of the GNPARG model with its asymptotics when prior information about the group label is known, and develops an EM algorithm to estimate the group ratio and labels when they are unknown. Section 4 reports numerical studies conducted to assess the finite-sample performance of our estimation procedure. We study the Chicago district-level crime data in Section 5.

Throughout the paper, we denote $\|\mathbf{x}\|_d = \left(\sum_{i=1}^p |x_i|^d\right)^{1/d}$ to be the ℓ^d -norm of a p -dimensional vector \mathbf{x} . For a $q \times p$ matrix $\mathbf{A} = (a_{ij})$, the generalized matrix norm is defined by $|||\mathbf{A}|||_d = \max_{\|\mathbf{x}\|_d=1} \|\mathbf{A}\mathbf{x}\|_d$. Particularly, $|||\mathbf{A}|||_1 = \max_{1 \leq j \leq p} \sum_{i=1}^q |a_{ij}|$ and $|||\mathbf{A}|||_2 = \sqrt{\rho(\mathbf{A}'\mathbf{A})}$, where $\rho(\cdot)$ denotes the spectral radius and $'$ denotes the transpose of a matrix or vector. $|||\mathbf{A}|||_2$ is in fact the operator norm of \mathbf{A} . The Frobenius norm of \mathbf{A} is denoted by $|||\mathbf{A}|||_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$.

2. Grouped Network Poisson Autoregressive Model

2.1 Previous Models

Following Lee, Lee and Tjøstheim (2018) and Fokianos, et al. (2020), we assume that $\{\mathbf{Y}_t = (Y_{1t}, Y_{2t}, \dots, Y_{Nt})', t \geq 1\}$ is an N -dimensional count time series, and $\{\boldsymbol{\lambda}_t = (\lambda_{1t}, \lambda_{2t}, \dots, \lambda_{Nt})', t \geq 1\}$ denotes the corresponding N -dimensional intensity process. Here N is fixed and finite. $\mathcal{F}_t^{\mathbf{Y}, \boldsymbol{\lambda}}$ is the σ -field generated by $\{\mathbf{Y}_t, \dots, \mathbf{Y}_0, \boldsymbol{\lambda}_0\}$ with $\boldsymbol{\lambda}_0$ being an initial value of $\{\boldsymbol{\lambda}_t\}$. The multivariate PAR model is defined as, for each $i = 1, 2, \dots, N$, and $t \geq 1$,

$$Y_{i,t} \mid \mathcal{F}_{t-1}^{\mathbf{Y}, \boldsymbol{\lambda}} \sim \text{Poisson}(\lambda_{i,t}), \quad \boldsymbol{\lambda}_t = \mathbf{d} + \mathbb{A}\boldsymbol{\lambda}_{t-1} + \mathbb{B}\mathbf{Y}_{t-1}, \quad (2.1)$$

where \mathbf{d} is an N -dimensional constant vector and \mathbb{A}, \mathbb{B} are $N \times N$ matrices. The elements of \mathbf{d}, \mathbb{A} and \mathbb{B} are assumed to be positive to ensure $\lambda_{i,t} > 0$ for all i and t .

In fact, for general \mathbb{A} and \mathbb{B} , statistical inference of model (2.1) is quite challenging when N is large. To reduce the complexity of model (2.1), in the spirit of Zhu, et al. (2017) and Zhou, et al. (2020), we introduce a network structure on the observed counts into model (2.1). Assume a known adjacency matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N \times N}$ is defined as $a_{ij} = 1$ if there is a directed edge from object i to object j , and $a_{ij} = 0$ otherwise. Let $a_{ii} = 0$ for $1 \leq i \leq N$. An NPAR model assumes that object i is only affected by the objects that it follows. It has the following form: for each

$i = 1, 2, \dots, N$, and $t \geq 1$,

$$Y_{i,t} \mid \mathcal{F}_{t-1}^{\mathbf{Y}, \boldsymbol{\lambda}} \sim \text{Poisson}(\lambda_{i,t}), \quad (2.2)$$

$$\lambda_{i,t} = \omega_0 + \alpha_0 Y_{i,t-1} + \rho_0 d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} + \beta_0 \lambda_{i,t-1},$$

where $\sum_{j \neq i}$ means $\sum_{j=1, j \neq i}^N$, and $d_i = \sum_{j=1}^N a_{ij}$ is the out-degree of i , which is the total number of objects that i points to. If there is no edge starting from object i , i.e., $d_i = 0$, we define that $d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} = 0$. α_0 measures the dependence on previous count, ρ_0 measures the dependence on the network structure, i.e. the average effect that neighbors have on each object, and β_0 measures the dependence on the previous intensity. The network structure reduces the inference complexity and makes the model more interpretable. Model (2.2) is different from the PNAR model proposed by Armillotta and Fokianos (2021) in that the lags of the intensity process $\boldsymbol{\lambda}_t$ are involved.

In model (2.2), however, all individuals are treated homogeneously since they share the same regression coefficients. This assumption is unrealistic in practice. For instance, the coefficient ρ_0 implies that all individuals are affected by their neighbors to the same extent, while in social networks the celebrities are less likely to be influenced by others than normal people.

2.2 Grouped Network Poisson Autoregressive Model

To relax the homogeneous assumption, in the spirit of Zhu and Pan (2020), we assume that all individuals can be classified into K groups, and each group is characterized by a specific set of parameters $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)' \in \mathbb{R}^4$, for $1 \leq k \leq K$, with each parameter being positive. Define a latent variable $z_{ik} \in \{0, 1\}$ for each object i , where $z_{ik} = 1$ if object i is from the k -th group, and $z_{ik} = 0$ otherwise. Assume $\{(z_{i1}, \dots, z_{iK})', 1 \leq i \leq N\}$ is a sequence of independent and identically distributed (i.i.d.) multinomial random vectors with number of events $n = 1$ and probability $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)'$. Here γ_k represents the group proportion satisfying $\gamma_k \geq 0$ and $\sum_{k=1}^K \gamma_k = 1$. A grouped network PAR model (GNPAR) can be constructed as

$$\begin{aligned}
 Y_{i,t} \mid \mathcal{F}_{t-1}^{\mathbf{Y}, \boldsymbol{\lambda}} &\sim \text{Poisson}(\lambda_{i,t}), \\
 \lambda_{i,t} &= \sum_{k=1}^K z_{ik} \left(\omega_k + \alpha_k Y_{i,t-1} + \rho_k d_i^{-1} \sum_{j \neq i} a_{ij} Y_{j,t-1} + \beta_k \lambda_{i,t-1} \right),
 \end{aligned} \tag{2.3}$$

for each $i = 1, \dots, N$, and $t \geq 1$. Following the NPAR model, the parameters $\omega_k, \alpha_k, \rho_k, \beta_k$ represent the group-specific baseline effect, regression coefficient on past observations, network effect, and regression coefficient on past intensity process, respectively. Note that we assume the adjacency matrix \mathbf{A} is asymmetric, which covers the special case of symmetric networks.

2.3 Stationarity and Ergodicity

In this subsection, we proceed to give a stationarity and ergodicity solution to model (2.3). Here dimension N is fixed throughout the article. Let $\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{N,t})'$, $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \dots, \lambda_{N,t})'$, $\mathbf{D} = \text{diag}(d_1, \dots, d_N)$, and $\mathbf{Z}_k = \text{diag}(z_{ik} : 1 \leq i \leq N)$. Furthermore, define $\mathcal{B}_0 = \sum_{k=1}^K \omega_k \mathbf{Z}_k \mathbf{1}_N$, $\mathcal{B}_1 = \sum_{k=1}^K \alpha_k \mathbf{Z}_k$, $\mathcal{B}_2 = \sum_{k=1}^K \rho_k \mathbf{Z}_k$, and $\mathcal{B}_3 = \sum_{k=1}^K \beta_k \mathbf{Z}_k$, where $\mathbf{1}_N = (1, \dots, 1)'$. Following Fokianos, et al. (2020), model (2.3) can be rewritten as

$$\mathbf{Y}_t = \mathbf{N}_t(\boldsymbol{\lambda}_t), \quad \boldsymbol{\lambda}_t = \mathcal{B}_0 + (\mathcal{B}_1 + \mathcal{B}_2 \mathbf{D}^{-1} \mathbf{A}) \mathbf{Y}_{t-1} + \mathcal{B}_3 \boldsymbol{\lambda}_{t-1}, \quad (2.4)$$

where $\{\mathbf{N}_t(\cdot)\}$ is a sequence of independent N -dimensional copula-Poisson processes.

See Subsection 4.1 for more details on the generation of $\mathbf{N}_t(\cdot)$.

Since \mathbf{Y}_t is integer-valued, ergodicity for model (2.4) is not sufficient to obtain the asymptotics of the parameter estimation, as discussed in Fokianos, Rahbek and Tjøstheim (2009), Fokianos and Tjøstheim (2011), and Tjøstheim (2012). To fix such an issue, ergodicity should be strengthened to geometric ergodicity. However, it is very difficult to establish geometric ergodicity, particularly, ϕ -irreducibility of (2.4) directly. To obtain ϕ -irreducibility, a perturbation method is helpful, i.e., adding a continuous component into the innovation, see Chapter 6 of Meyn and Tweedie (1993). Thus, in line with Fokianos, Rahbek and Tjøstheim (2009), we define the

perturbed model as

$$\mathbf{Y}_t^m = \mathbf{N}_t(\boldsymbol{\lambda}_t^m), \quad \boldsymbol{\lambda}_t^m = \mathcal{B}_0 + (\mathcal{B}_1 + \mathcal{B}_2 \mathbf{D}^{-1} \mathbf{A}) \mathbf{Y}_{t-1}^m + \mathcal{B}_3 \boldsymbol{\lambda}_{t-1}^m + \boldsymbol{\epsilon}_t^m, \quad (2.5)$$

where $\boldsymbol{\epsilon}_t^m = c_m \mathbf{V}_t$, the sequence c_m is strictly positive and tends to zero, as $m \rightarrow \infty$, and \mathbf{V}_t is an N -dimensional vector which consists of independent positive random variables, each of which having a bounded support of the form $[0, M]$, for some $M > 0$. From Lemma 1 in the Supplementary Materials, we can see that the difference between unperturbed model (2.4) and perturbed model (2.5) can be arbitrarily small in some sense.

The following proposition gives a sufficient condition on geometric ergodicity of model (2.5), together with a stationary and ergodic condition for model (2.4). The proof of Proposition 1 is postponed to the Supplementary Materials.

Proposition 1. (i). *The process $\{\boldsymbol{\lambda}_t^m, t > 0\}$ is a geometrically ergodic Markov chain with finite r th moments for any $r > 0$, if $\| \max_{1 \leq k \leq K} (\alpha_k + \beta_k) \mathbf{I}_N + \max_{1 \leq k \leq K} \rho_k \mathbf{D}^{-1} \mathbf{A} \|_2 < 1$. Moreover, the process $\{(\mathbf{Y}_t^m, \boldsymbol{\lambda}_t^m, \boldsymbol{\epsilon}_t), t > 0\}$ is $V_{\mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\epsilon}}$ -geometrically ergodic Markov chain with $V_{\mathbf{Y}, \boldsymbol{\lambda}, \boldsymbol{\epsilon}} = 1 + \|\mathbf{Y}\|_2^r + \|\boldsymbol{\lambda}\|_2^r + \|\boldsymbol{\epsilon}\|_2^r$, $r > 0$.*

(ii). *If $\| \left(\max_{1 \leq k \leq K} \alpha_k \right) \mathbf{I}_N + \left(\max_{1 \leq k \leq K} \rho_k \right) \mathbf{D}^{-1} \mathbf{A} \|_1 + \max_{1 \leq k \leq K} \beta_k < 1$, then there exists a unique stationary and ergodic solution $\{(\mathbf{Y}_t, \boldsymbol{\lambda}_t)\}$ to model (2.4) which is non-anticipative and satisfies $E\|\mathbf{Y}_t\|_r^r < \infty$ and $E\|\boldsymbol{\lambda}_t\|_r^r < \infty$ for any $r > 0$.*

Remark 1. In Proposition 1, (i) is developed via the perturbation technique, while

(ii) is based on the notion of weak dependence. The latter does not require an introduction of a perturbed model, but the obtained sufficient conditions are much stronger. In what follows, we prefer the sufficient stationarity and ergodicity condition (i) for the perturbed process and use the closeness between the perturbed model and the unperturbed one to obtain asymptotic normality of the maximum likelihood estimation (MLE) of model (2.4). It should be noted that geometric ergodicity of the perturbed process makes it possible to employ classical statistical inference theory similar to GARCH models.

Remark 2. Proposition 1 is constructed with a fixed N , which does not necessarily hold if N is diverging, since no stationarity and ergodicity conditions are available when $\min\{N, T\} \rightarrow \infty$. In fact, how to generally define stationarity of time series with diverging dimension still remains open. Moreover, the ergodicity conditions in Proposition 1 are different from those of the PNAR model proposed by Armillotta and Fokianos (2021), in that the latter does not contain the lags of the intensity process λ_t .

3. Parameter Estimation

This section studies the MLE of the GNPARG model and establishes its asymptotics. Since there exists a latent variable z_{ik} as group labels, parameter estimation and group classification need to be conducted simultaneously. We first study the MLE

of the model parameter when group labels are known, and then develop an EM algorithm for estimating the group labels when they are unknown. The former is useful if we have prior information for group classification, and the latter is more practical when there is little prior information available.

3.1 MLE when group labels are known

Suppose z_{ik} is known, define $\mathcal{G}_k = \{i \leq N : z_{ik} = 1\}$ and $N_k = |\mathcal{G}_k|$ for $1 \leq k \leq K$, denoting the group member and group size, respectively. Assume that the observations $\{\mathbf{Y}_t, t = 1, \dots, T\}$ are from model (2.4) with true parameter $\boldsymbol{\theta}_0 = (\omega_{k0}, \alpha_{k0}, \rho_{k0}, \beta_{k0} : 1 \leq k \leq K)' \in \mathbb{R}_+^{4K}$, where $\mathbb{R}_+ = (0, \infty)$. Let $\mathbf{Y}_t^{(k)} = (Y_{i,t} : i \in \mathcal{G}_k)' \in \mathbb{R}^{N_k}$, $t = 1, \dots, T$, be in the k th group. Define $\boldsymbol{\lambda}_t^{(k)} = (\lambda_{i,t} : i \in \mathcal{G}_k)' \in \mathbb{R}^{N_k}$, $\mathbf{D}^{(k)} = \text{diag}(d_i : i \in \mathcal{G}_k) \in \mathbb{R}^{N_k \times N_k}$, $\mathbf{A}^{(k)} = (a_{ij} : i \in \mathcal{G}_k, 1 \leq j \leq N) \in \mathbb{R}^{N_k \times N}$. Then, the GNPARG model (2.4) can be rewritten as

$$\mathbf{Y}_t^{(k)} = \mathbf{N}_t(\boldsymbol{\lambda}_t^{(k)}), \quad \boldsymbol{\lambda}_t^{(k)} = \omega_{k0} + \alpha_{k0} \mathbf{Y}_{t-1}^{(k)} + \rho_{k0} (\mathbf{D}^{(k)})^{-1} \mathbf{A}^{(k)} \mathbf{Y}_{t-1} + \beta_{k0} \boldsymbol{\lambda}_{t-1}^{(k)} \quad (3.6)$$

for $1 \leq k \leq K$. Under this setting, the true parameter $\boldsymbol{\theta}_{k0} = (\omega_{k0}, \alpha_{k0}, \rho_{k0}, \beta_{k0})'$ can be estimated separately for each group. Without loss of generality, we consider the MLE for the k th group hereinafter.

Let $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)' \in \mathbb{R}_+^4$ be the parameter. The conditional likelihood

function, given $\boldsymbol{\lambda}_0$, is given by

$$L(\boldsymbol{\theta}_k) = \prod_{t=1}^T \prod_{i \in \mathcal{G}_k} \left\{ \frac{\lambda_{i,t}^{Y_{i,t}}(\boldsymbol{\theta}_k) \exp(-\lambda_{i,t}(\boldsymbol{\theta}_k))}{Y_{i,t}!} \right\} \quad (3.7)$$

and the log-likelihood function (ignoring the constant) is

$$l(\boldsymbol{\theta}_k) = \frac{1}{T} \sum_{t=1}^T l_t(\boldsymbol{\theta}_k), \quad l_t(\boldsymbol{\theta}_k) = \frac{1}{N_k} \sum_{i \in \mathcal{G}_k} (Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)). \quad (3.8)$$

The MLE of $\boldsymbol{\theta}_{k0}$ is defined as

$$\widehat{\boldsymbol{\theta}}_k = (\widehat{\omega}_k, \widehat{\alpha}_k, \widehat{\rho}_k, \widehat{\beta}_k)' = \arg \max_{\boldsymbol{\theta}_k \in \Theta_k} l(\boldsymbol{\theta}_k). \quad (3.9)$$

Let $\Theta := \Theta_1 \times \cdots \times \Theta_K \subset \mathbb{R}_+^{4K}$ be the parameter space and $\boldsymbol{\theta} \in \Theta$. Before we study the asymptotics of $\widehat{\boldsymbol{\theta}}_k$, we first give two assumptions.

Assumption 1. *The parameter space Θ is a compact set of \mathbb{R}_+^{4K} , and the true parameter $\boldsymbol{\theta}_0$ is an interior point of Θ .*

Assumption 2. *$\boldsymbol{\theta}_0$ satisfies $\| \max_{1 \leq k \leq K} (\alpha_{k0} + \beta_{k0}) \mathbf{I}_N + \max_{1 \leq k \leq K} \rho_{k0} \mathbf{D}^{-1} \mathbf{A} \|_2 < 1$.*

The following theorem states the strong consistency and asymptotic normality of the MLE $\widehat{\boldsymbol{\theta}}_k$.

Theorem 1. *If Assumptions 1-2 hold, then there exists an open neighborhood, say, $O(\boldsymbol{\theta}_{k0}) = \{\boldsymbol{\theta}_k : \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k0}\|_2 < \delta\}$, of $\boldsymbol{\theta}_{k0}$ such that with probability tending to 1 as $T \rightarrow \infty$, the equation $S_T(\boldsymbol{\theta}_k) = 0$ has a unique solution, denoted by $\widehat{\boldsymbol{\theta}}_k$. Furthermore, $\widehat{\boldsymbol{\theta}}_k$ is strongly consistent, i.e., $\widehat{\boldsymbol{\theta}}_k \rightarrow \boldsymbol{\theta}_{k0}$ a.s., and is asymptotically normal, i.e.,*

$\sqrt{N_k T}(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{k0}) \xrightarrow{d} N(0, \mathbf{H}^{-1} \mathbf{G} \mathbf{H}^{-1})$, as $T \rightarrow \infty$, where “ \xrightarrow{d} ” stands for convergence in distribution, the matrices $\mathbf{G} := \mathbf{G}(\boldsymbol{\theta}_{k0})$ and $\mathbf{H} := \mathbf{H}(\boldsymbol{\theta}_{k0})$ are defined by

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta}_{k0}) &= \frac{1}{N_k} \sum_{i \in \mathcal{G}_k} \sum_{j \in \mathcal{G}_k} \mathbb{E} \left\{ \frac{1}{\lambda_{i,t}(\boldsymbol{\theta}_{k0}) \lambda_{j,t}(\boldsymbol{\theta}_{k0})} \Sigma_{ij,t}^{(k)}(\boldsymbol{\theta}_{k0}) \frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}_k} \frac{\partial \lambda_{j,t}(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}'_k} \right\}, \\ \mathbf{H}(\boldsymbol{\theta}_{k0}) &= \frac{1}{N_k} \sum_{i \in \mathcal{G}_k} \mathbb{E} \left\{ \frac{1}{\lambda_{i,t}(\boldsymbol{\theta}_{k0})} \frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}_k} \frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}'_k} \right\}, \\ \frac{\partial \boldsymbol{\lambda}_t^{(k)' }(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}_k} &= \left(\mathbf{1}_{N_k}, \mathbf{Y}_{t-1}^{(k)}, (\mathbf{D}^{(k)})^{-1} \mathbf{A}^{(k)} \mathbf{Y}_{t-1}, \boldsymbol{\lambda}_{t-1}^{(k)} \right)' + \beta_{k0} \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)' }(\boldsymbol{\theta}_{k0})}{\partial \boldsymbol{\theta}_k}, \end{aligned}$$

with $\Sigma_t^{(k)}(\boldsymbol{\theta}_{k0})$ being covariance matrix of $\mathbf{Y}_t^{(k)}$, $\Sigma_{ij,t}^{(k)}(\cdot)$ being the $(i, j)^{th}$ entry of $\Sigma_t^{(k)}(\cdot)$, and the expectation is taken with respect to the invariant stationary distribution of $\{\mathbf{Y}_t^{(k)}\}$.

The proof of Theorem 1 is postponed to the Supplementary Materials. Note that when the components of the process $\{\mathbf{Y}_t^{(k)}\}$ are uncorrelated, we have $\mathbf{G} = \mathbf{H}$, thus the asymptotic covariance matrix reduces to the standard one for ordinary MLE. In practice, all the above quantities can be consistently estimated by their respective sample counterparts, e.g.,

$$\widehat{\mathbf{H}} = \frac{1}{N_k T} \sum_{i \in \mathcal{G}_k} \sum_{t=1}^T \left\{ \frac{1}{\lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)} \frac{\partial \lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}_k} \frac{\partial \lambda_{i,t}(\widehat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}'_k} \right\}.$$

Remark 3. From Theorem 1, we can see that the convergence rate depends on both N_k and T , although N_k is fixed. The network structure characterized by \mathbf{A} and the number of group K are fixed in our model setting. There is no extra assumption on the network structure. Since parameters $\boldsymbol{\theta}_k$ from different groups are uncorrelated in

the asymptotic covariance matrix, the MLE of $\boldsymbol{\theta}_{k0}$ for each group can be conducted separately. Thus, the fixed group number K will not affect the convergence rate in each group's estimation, while larger one may consume more computational time, as more unknown parameters are involved. Further, if N_k is diverging, Theorem 1 will possibly break. As one referee points out, in that case, some connectivity and uniformity assumptions on the network structure should be imposed, as well as some regularity assumptions on the structure of dependence among the error, see, e.g., Zhu, et al. (2017) and Armillotta and Fokianos (2021). This needs to be further explored and we leave such an issue for future research.

3.2 Estimation with unknown group labels

When the group labels are unknown, the estimation includes the latent variables. A common method to deal with such mixture models is the EM algorithm. Recall that $z_{ik} \in \{0, 1\}$ indicates whether object i belongs to the k -th group. The full likelihood function is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \prod_{k=1}^K \left\{ \gamma_k \prod_{t=1}^T \frac{\lambda_{i,t}^{Y_{i,t}}(\boldsymbol{\theta}_k) \exp(-\lambda_{i,t}(\boldsymbol{\theta}_k))}{Y_{i,t}!} \right\}^{z_{ik}}. \quad (3.10)$$

The EM algorithm consists of two steps: the expectation step and the maximization step. First, we set an initial value for the parameter $\hat{\boldsymbol{\theta}}^{(0)}$ and $\hat{\boldsymbol{\gamma}}^{(0)}$, and follow the procedure described below. Specifically, in the m -th ($m \geq 1$) iteration, the estimation procedure is as follows:

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

$$z_{ik}^{(m)} = \mathbb{E}(z_{ik} | \hat{\boldsymbol{\theta}}^{(m-1)}) = \frac{\hat{\gamma}_k^{(m-1)} \prod_{t=1}^T \hat{\Delta}_{it,k}^{(m-1)}}{\sum_{j=1}^K \hat{\gamma}_j^{(m-1)} \prod_{t=1}^T \hat{\Delta}_{it,j}^{(m-1)}}, \quad (3.11)$$

where $\hat{\Delta}_{it,k}^{(m-1)} = \lambda_{i,t}^{Y_{i,t}}(\hat{\boldsymbol{\theta}}_k^{(m-1)}) \exp(-\lambda_{i,t}(\hat{\boldsymbol{\theta}}_k^{(m-1)}))$ (omitting the constant term), and $\hat{\boldsymbol{\theta}}_k^{(m-1)}$ is an estimate in the $(m-1)$ -th iteration.

- M-STEP. Given an estimate $z_{ik}^{(m)}$, we maximize Q -function below with respect to $\boldsymbol{\theta}_k$ and γ_k (ignoring the constant term):

$$\begin{aligned} Q(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(m-1)}) &= \mathbb{E} \left\{ \log L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{z}) | \mathbf{Y}, \hat{\boldsymbol{\theta}}^{(m-1)} \right\} \\ &= \mathbb{E} \left\{ \sum_{i=1}^N \sum_{k=1}^K z_{ik} \left(\log \gamma_k + \sum_{t=1}^T (Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)) \right) | \mathbf{Y}, \hat{\boldsymbol{\theta}}^{(m-1)} \right\} \\ &= \sum_{i=1}^N \sum_{k=1}^K z_{ik}^{(m)} \left(\log \gamma_k + \sum_{t=1}^T (Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)) \right). \end{aligned}$$

Thus, we have

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k^{(m)} &= \arg \max_{\boldsymbol{\theta}_k \in \Theta_k} \left\{ \sum_{i=1}^N z_{ik}^{(m)} \sum_{t=1}^T (Y_{i,t} \log \lambda_{i,t}(\boldsymbol{\theta}_k) - \lambda_{i,t}(\boldsymbol{\theta}_k)) \right\}, \\ \hat{\gamma}_k^{(m)} &= \frac{1}{N} \sum_{i=1}^N z_{ik}^{(m)}. \end{aligned} \quad (3.12)$$

Repeat the steps (3.11)-(3.12) until the EM algorithm converges and we can get the EM estimates $\hat{\boldsymbol{\theta}}_k$ and $\hat{\gamma}_k$ for $1 \leq k \leq K$. Note that the EM estimator $\hat{\boldsymbol{\theta}}_k$ given in (3.12) can be viewed as a weighted MLE estimator in (3.9), with the latent group variables z_{ik} as the weights.

Remark 4. In practice, the computation of the E-Step (3.11) may be unstable and quite sensitive to initial values, especially when the sample size T is large, which leads to unsatisfactory performance of the estimator $\widehat{\gamma}$. To address this problem, we adopt the TS estimation method introduced by Zhu and Pan (2020) to set the initial value. More specifically, we first estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level and obtain N sets of ML estimates $\widehat{\boldsymbol{\theta}}_k$, $1 \leq k \leq N$. Next, we apply some cluster algorithm (e.g., k -means clustering) to partition these N sets of estimates into K groups. Let $\widehat{\mathcal{G}}_k$ denote the corresponding members in group k , and $\widehat{N}_k := |\widehat{\mathcal{G}}_k|$ be the cardinality. Then the initial value of group proportion γ_k can be estimated as $\widehat{\gamma}_k^{(0)} = \widehat{N}_k/N$, for all $1 \leq k \leq K$. Finally, given the group information $\widehat{\mathcal{G}}$, we estimate $\boldsymbol{\theta}$ by the MLE (3.9), and set the estimate to be the initial value of $\boldsymbol{\theta}$, i.e. $\widehat{\boldsymbol{\theta}}^{(0)}$.

Remark 5. How to select a reasonable number of groups K has been a long lasting topic. We here recommend two procedures to determine K .

One is from the perspective of model setting. Since the nodes in the same group are characterized by the same set of parameters $\boldsymbol{\theta}_k$, we can adopt the TS estimation method introduced in **Remark 4** and classify the estimated parameters using classical cluster algorithms. Specifically, we first estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level and obtain N sets of ML estimates $\widehat{\boldsymbol{\theta}}_k$, $1 \leq k \leq N$. Then, we apply k -means clustering to partition these N sets of estimates into K groups, and select an optimal K based on the Elbow plot, Silhouette coefficient, or gap statistic.

The other is from the perspective of model fitting. As will be discussed in Section 4.3, we can try model fitting with different numbers of groups, say $K = 1, \dots, 5$, then compare their out-of-sample predicted RMSEs in (4.14) among candidate models, and finally choose a reasonable K in RMSEs. Meanwhile, the in-sample fitted RMSEs in (4.13) can also be used as an auxiliary measure for K selection.

4. Simulation Studies

In this section, we investigate the finite-sample performance of the proposed model and estimation procedure. We first explore the performance of the MLE of $\boldsymbol{\theta}$ when group labels are unknown. Then, we evaluate the model estimation and prediction accuracy when the number of groups K is misspecified. The performance of the MLE of $\boldsymbol{\theta}$ when group labels are known is reported in Supplementary Materials S2.1.

4.1 Simulated data

We first generate the adjacency matrix \mathbf{A} with two different mechanisms: the Erdős-Rényi model and the stochastic blockmodel. These mechanisms are chosen to illustrate the performance of our model under different network structures \mathbf{A} , and are independent of the membership generating mechanism in what follows. Note that directed graphs are considered here which cover the cases of undirected graphs.

Case 1: The Erdős-Rényi model

The Erdős-Rényi model (Erdős and Rényi, 1960) is by far the most thoroughly studied network model in the literature. It assumes that given the number of vertices N_v , all edges are independent with a given probability $p \in (0, 1)$. The Erdős-Rényi model has the property that for large N_v , the degree distribution of the graph is approximately Poisson distributed with mean $p(N_v - 1)$. Here we set $p = 3/N_v$. Loops are not allowed. A visualization of the network structure and the histogram of degree distributions of one realization are shown in Fig. 1, with $N_v = 50$.

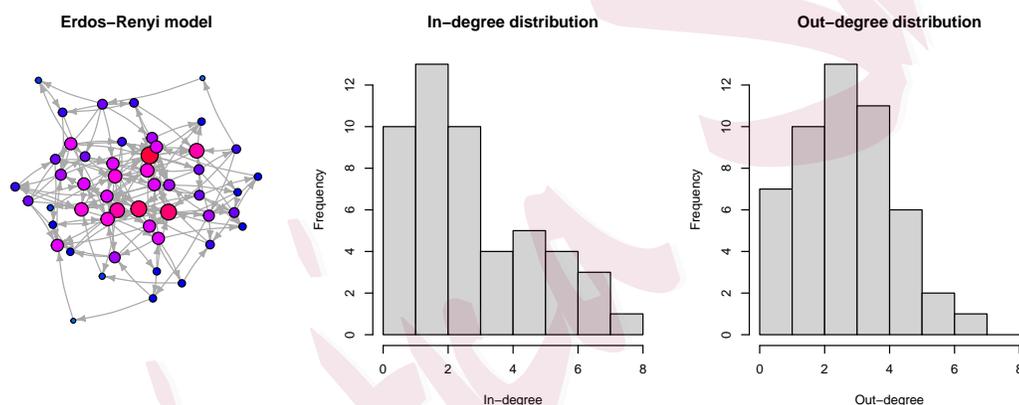


Figure 1: The network visualization and histogram of degree for the Erdős-Rényi model, with $N_v = 50$ and $p = 0.06$.

Case 2: Stochastic Blockmodel

The stochastic blockmodel (Holland, Laskey and Leinhardt, 1983) is another popular network topology in the literature. It assumes that the nodes in the same

block are more likely to be connected to each other than those from different blocks. Here we set $K_v = \{3, 5, 10\}$ as the total number of blocks, and $N_v = \{20, 50, 100\}$ as the total number of nodes, with each block having N_v/K_v nodes. We assume there is a directed edge to every pair of vertices with probability $3K_v/N_v$ if they belong to the same community, and $0.3/N_v$ for those in different communities. A network visualization and histogram of degree with $N_v = 50$ and $K_v = 5$ are shown in Fig. 2.

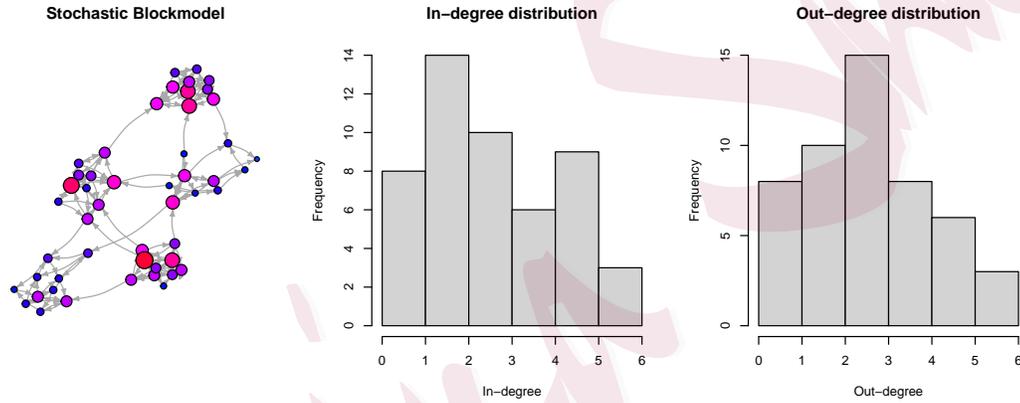


Figure 2: The network visualization and histogram of degree for stochastic block-model, with $N_v = 50$ vertices and $K_v = 5$ blocks.

Once the adjacency matrix \mathbf{A} is generated, it is fixed throughout the remaining simulation studies. We set the number of groups to be $K = 3$. To characterize different nodal behaviors, we set different parameters $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)'$ for each group, as listed in Table 1. Group 1 has relatively low regression coefficients on

past observations and past intensity process (i.e., α and β), while a relatively high network effect (i.e., ρ), implying the behavior of most objects are affected by the objects they follow. On the contrary, Group 3 has a lower ρ , and higher ω , α and β , representing the influential nodes which are more likely self-driven and less rely on others. They only take up 20% of whole objects. Group 2 has medium level of coefficients and group size. Note that the parameters satisfy Assumptions 1-2. We randomly assign each node to the K groups according to the group proportion γ_k .

Table 1: True parameters in model (2.3) for each group with $K = 3$.

	ω	α	ρ	β	γ
Group 1	0.2	0.1	0.3	0.2	0.5
Group 2	0.5	0.2	0.2	0.3	0.3
Group 3	1	0.3	0.1	0.4	0.2

Given an initial value $\boldsymbol{\lambda}_0 = \mathbf{4}$, the observed time series \mathbf{Y}_0 are generated as an N -dimensional count time series with intensity $\boldsymbol{\lambda}_0$, and $\{\mathbf{Y}_t, \boldsymbol{\lambda}_t; t \geq 1\}$ are simulated consecutively conditioned on the previous information, according to the GNPARG model (2.3). The first 50 samples are discarded to eliminate the effect of initial values. As mentioned before, to establish a well-defined joint distribution of multivariate count data with marginal Poisson distribution, we use the copula-based data generating process introduced in Fokianos, et al. (2020). It has the advantage that

copula is defined uniquely for continuous multivariate random variables, and it imposes arbitrary dependence among the marginal Poisson components. Denote $\{\mathbf{N}_t\}$ to be a sequence of independent N -dimensional copula-Poisson processes. Specifically, the data generating process is given below. Suppose that $\boldsymbol{\lambda}_0 = (\lambda_{1,0}, \dots, \lambda_{p,0})'$ is an initial value.

- (1) Let $\mathbf{U}_l = (U_{1,l}, \dots, U_{N,l})'$, $l = 1, \dots, S$, be a sample from an N -dimensional copula $C(u_1, \dots, u_N)$, where $U_{i,l}$, $l = 1, \dots, S$, follows the uniform distribution $U(0, 1)$ for $i = 1, \dots, N$.
- (2) Consider the transformation $X_{i,l} = -(\log U_{i,l})/\lambda_{i,0}$, $i = 1, \dots, N$. Then, $X_{i,l}$, $l = 1, \dots, S$, follows an exponential distribution with parameter $\lambda_{i,0}$, $i = 1, \dots, N$.
- (3) Define $Y_{i,0} = \max \{0 \leq s \leq S : \sum_{l=1}^s X_{i,l} \leq 1\}$, $i = 1, \dots, N$, by taking S large enough. Then $\mathbf{Y}_0 = (Y_{1,0}, \dots, Y_{N,0})'$ is a set of marginal Poisson process with parameter $\boldsymbol{\lambda}_0$.
- (4) Use model (2.3) to obtain $\boldsymbol{\lambda}_1$, return back to step (1) to obtain \mathbf{Y}_1 , and so on.

In practice, the sample size S should be large, say $S = 1000$. The copula $C(\cdot)$ can be chosen as the Gaussian or the Clayton copula, and the unknown parameter of the copula, say ϕ , needs to be determined based on the contemporaneous correlation among random variables. A parametric bootstrap based algorithm can be used to identify the copula structure and unknown parameter, see S-7 in Fokianos, et al.

(2020). In this section, we employ the Gaussian copula with parameter $\phi = 0.5$, allowing for arbitrary dependence among the marginal Poisson components.

4.2 Simulation results when group labels are unknown

To assess the finite-sample performance of the MLE when group labels are unknown, we apply the EM algorithm to estimate θ_0 and γ_0 simultaneously. The initial value is set by the TS estimation method described in **Remark 4**. Two different types of network structures are considered, each with combinations of network size (i.e., $N = 20, 50, 100$) and sample size (i.e., $T = 100, 200, 400$). Each case is randomly simulated with $R = 1000$ replicates. Denote the estimates obtained in the r th simulation to be $\hat{\theta}^{(r)} = (\hat{\omega}^{(r)}, \hat{\alpha}^{(r)}, \hat{\rho}^{(r)}, \hat{\beta}^{(r)})'$ and $\hat{\gamma}^{(r)}$, where $1 \leq r \leq R$. Moreover, the group label for each node is estimated as $\hat{z}_i^{(r)} = \arg \max_k \{\hat{z}_{ik}^{(r)}\}$. The simulation results are summarized in Tables 2-3 for the Erdős-Rényi model and stochastic blockmodel, respectively.

First, the RMSE is calculated for each estimator. Here we report the RMSE taken average of all groups. For example, for the network effect coefficient ρ , $\text{RMSE}_\rho = \{(KR)^{-1} \sum_{k=1}^K \sum_{r=1}^R (\hat{\rho}_k^{(r)} - \rho_k)^2\}^{1/2}$. For the group ratio γ , $\text{RMSE}_\gamma = \{(KR)^{-1} \sum_{k=1}^K \sum_{r=1}^R (\hat{\gamma}_k^{(r)} - \gamma_k)^2\}^{1/2}$. Next, we employ the misclassification rate (MCR) to evaluate the accuracy of estimated group label. Specifically, $\text{MCR} = (NR)^{-1} \sum_{r=1}^R \sum_{i=1}^N I(\hat{z}_i^{(r)} \neq z_i)$, where z_i is the true group label of object i . The last

column calculates the network density, which is defined as $\{N(N-1)\}^{-1} \sum_{i,j} a_{ij}$.

Table 2: Simulation results for the Erdős-Rényi model. The RMSEs ($\times 10^2$) are reported with the misclassification rate (%) and the network density (%).

N	T	ω	α	ρ	β	γ	MCR	Network Density
20	100	39.09	7.24	9.70	18.19	5.28	7.39	13.16%
	200	23.93	5.14	7.00	11.33	3.90	3.37	
	400	16.98	4.00	5.73	8.12	3.33	2.95	
50	100	27.44	4.72	6.51	14.99	2.85	2.49	7.10%
	200	19.59	3.37	4.72	8.87	2.51	0.96	
	400	14.53	2.63	3.21	6.25	2.72	1.07	
100	100	21.86	3.96	4.53	12.66	2.45	1.45	3.10%
	200	14.84	2.60	2.90	7.56	1.80	0.37	
	400	9.13	1.81	2.05	5.20	0.96	0.19	

Tables 2-3 show that the RMSEs are all very small for estimators $\hat{\alpha}$, $\hat{\rho}$, $\hat{\beta}$ and $\hat{\gamma}$. For the baseline effect estimator $\hat{\omega}_k$, the RMSEs are relatively large. As the network dimension N and sample size T increase, the RMSEs of $\hat{\theta}$ and $\hat{\gamma}$ decrease toward zero, which implies more accurate estimates and smaller standard deviations. Moreover, the misclassification rates of group labels are also quite small, and decrease rapidly as network size and sample size increase. These facts indicate the good performance

Table 3: Simulation results for the stochastic blockmodel. The RMSEs ($\times 10^2$) for each estimator are reported with the misclassification rate (%) and the network density (%).

N	T	ω	α	ρ	β	γ	MCR	Network Density
20	100	43.03	7.00	7.22	16.89	5.42	8.98	13.16%
	200	28.10	4.60	4.76	10.16	4.27	3.03	
	400	15.32	3.39	3.61	6.84	2.87	2.02	
50	100	26.56	5.12	6.69	15.12	3.46	3.30	6.24%
	200	15.80	3.11	4.39	8.36	2.29	0.60	
	400	11.70	2.49	3.22	6.15	2.20	0.73	
100	100	22.86	4.03	4.38	12.55	2.20	1.31	2.92%
	200	15.63	2.74	2.78	7.59	2.39	0.62	
	400	10.58	1.91	2.03	5.40	1.58	0.25	

of ML estimators and the effectiveness of the EM algorithm.

4.3 Model performance when number of groups K is misspecified

So far we set the group number as $K = 3$, while in reality the true number of groups is unknown and could be incorrectly specified. In this subsection, we study the impact of such misspecification on model estimation and prediction accuracy.

The data is generated under the stochastic blockmodel, and the true number of groups is $K = 3$ with the same parameters in Table 1. We choose $K = 1, 2, 4, 5$ as the misspecified number of groups. The network size is $N = 20, 50, 100$ and sample size is $T = 100, 200, 400$, each with $R = 1000$ replicates. The total time period of generated data is $T + 20$, where the first T periods are used for parameter estimation, and the remaining 20 periods are used for prediction.

For each selected number of groups K , denote $\hat{\mathbf{Y}}_t$ as the fitting response for $t = 1, \dots, T$ and the predicted value for $t = T + 1, \dots, T + 20$. Since the parameter estimation error cannot be naturally defined when the number of groups is incorrect, we employ the estimation error of the response instead to compare the performance of model with different K . The in-sample RMSE for the fitted value is defined as

$$\text{RMSE}_{esti} = \left\{ (NT)^{-1} \sum_{t=1}^T \|\hat{\mathbf{Y}}_t - \mathbf{E}(\mathbf{Y}_t | \mathcal{F}_{t-1}, \mathbf{Z})\|^2 \right\}^{1/2}, \quad (4.13)$$

where $\mathbf{E}(\mathbf{Y}_t | \mathcal{F}_{t-1}, \mathbf{Z})$ is the conditional expectation of response \mathbf{Y}_t based on the historical and group information, which is equal to $\boldsymbol{\lambda}_t$ in our model. The out-of-

sample predictive RMSE is defined as

$$\text{RMSE}_{pred} = \left\{ (20N)^{-1} \sum_{t=T+1}^{T+20} \|\widehat{\mathbf{Y}}_t - \mathbf{Y}_t\|^2 \right\}^{1/2}. \quad (4.14)$$

The mean values of these statistics are summarized in Table 4. We can see that both the estimation errors and the prediction errors shrink sharply from $K \leq 2$ to the true value $K = 3$ in all scenarios, and decrease smoothly as K increases. Especially the prediction errors remain steady for $K \geq 3$. Therefore, in practice we could try model fitting with different number of groups, say $K = 1, \dots, 5$, compare prediction errors among candidate models, and then select a reasonable number of groups K . This confirms the effectiveness of the second method of K -selection in **Remark 5**. The performance of the first method of K -selection has also been verified via simulations, which is included in Supplementary Materials S2.2.

5. Case study: Chicago Police Department Investigatory Stop Report (ISR) data

5.1 Data description

To illustrate the above methodology, we consider an application on crime data in Chicago. Chicago is one of the most racially and socio-economically segregated cities in America, and its crime rate remains at top level even in the world. We use the data from Chicago Police Department Investigatory Stop Report (ISR). An investigatory

Table 4: Simulation results for different number of groups K in stochastic block-models with 500 replicates. The $\text{RMSE}_{esti}(\times 10^2)$ and the RMSE_{pred} are reported.

N	T	Estimation					Prediction				
		K=1	K=2	K=3	K=4	K=5	K=1	K=2	K=3	K=4	K=5
20	100	50.68	31.30	19.81	20.04	20.97	1.50	1.45	1.43	1.43	1.44
	200	49.94	29.08	13.49	14.18	14.97	1.49	1.44	1.42	1.42	1.42
	400	49.54	28.02	10.56	10.69	11.34	1.50	1.45	1.42	1.42	1.42
50	100	45.19	27.87	14.33	14.44	15.21	1.38	1.34	1.32	1.32	1.32
	200	44.51	26.04	9.56	10.00	10.60	1.39	1.34	1.32	1.32	1.32
	400	44.22	25.39	6.68	6.80	7.21	1.39	1.34	1.32	1.32	1.32
100	100	46.55	26.78	12.79	13.15	13.70	1.42	1.37	1.35	1.35	1.35
	200	45.90	26.68	8.52	8.96	9.41	1.40	1.36	1.34	1.34	1.34
	400	45.59	25.52	5.94	6.21	6.52	1.42	1.37	1.35	1.35	1.35

stop could be made by a police officer if there are specific and articulable facts to suspect that a criminal activity is afoot. Thus, the number of investigatory stops in which any enforcement action was taken can be viewed as a measure of crime index in one area. It is of interest to study the dynamic and spatial patterns of investigatory stops, and how the crime numbers from different districts interact with each other, which will be very helpful in crime prevention and policy making.

Here we consider the number of daily investigatory stops that involve an enforcement action (arrest, personal service citation, etc.) in each district in 2019 ($T = 365$) as the response Y_{it} . The dataset is from the public data of Chicago Police Department named “ISR Data 2019” (<https://home.chicagopolice.org/statistics-data/isr-data/>). Chicago Police Department divides the whole city into $N = 22$ districts, as shown in Fig. 3 (a). The figure is in fact the criminal homicide distribution by district in 2019, where darker colors represent districts with relatively more criminal homicide cases. We see that the crime rate is high in the middle and southern part of Chicago, and relatively low in the northeast areas. Fig. 3 (b) shows the construction of the symmetric adjacency matrix, which is based on the spatial distribution of districts, i.e., there is an edge between district i and j if the two districts share a border. The network density is 19.9%. The larger nodes indicate more investigatory stops in these districts, and smaller nodes denote fewer stops. We can see that the distributions of investigatory stops and criminal homicide are very similar.

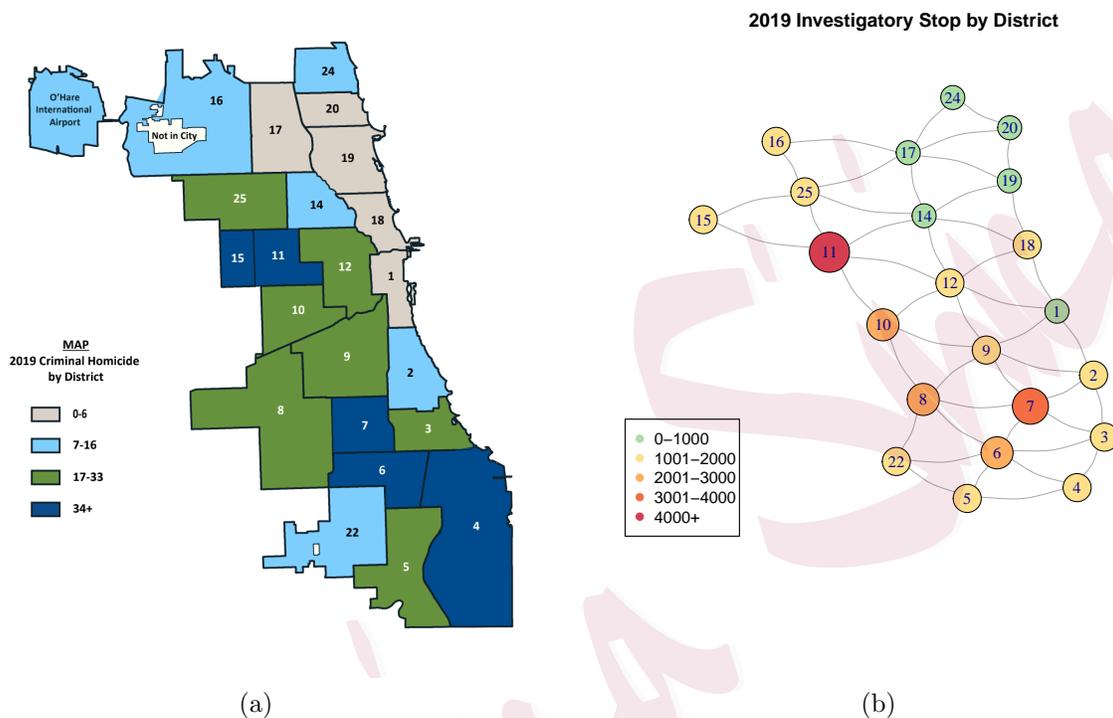


Figure 3: (a) District division in Chicago (2019 Criminal Homicide by District). Note that No.13, 21, 23 are unused. The figure is from the 2019 annual report of Chicago Police Department (<https://home.chicagopolice.org/wp-content/uploads/2020/09/19AR.pdf>). (b) The constructed network structure, where the node's color and size denote the level of 2019 yearly investigatory stops (involving enforcement action). The bigger nodes with deeper colors represent more investigatory stops occurred in this district.

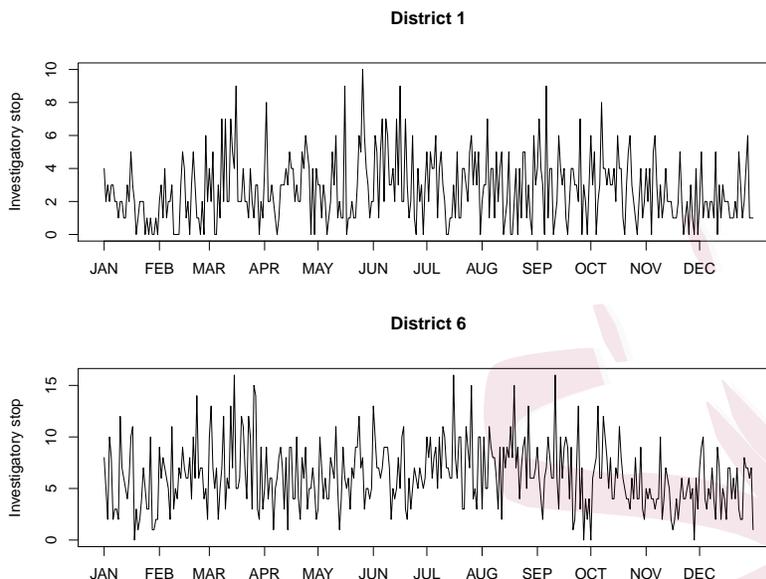


Figure 4: Number of daily investigatory stops in District 1 and 6.

The time series of the number of daily investigatory stops for District 1 and 6 are plotted in Fig. 4 for illustration. There exists dependency within the individual series. The average variance-to-mean ratio for each district is 1.85, and the overall variance-to-mean ratio is 3.16, which imply over-dispersion in the data.

5.2 Model estimation and interpretation

To determine the number of groups, we follow the two approaches described in **Remark 5**. First, we fit the data with different number of groups, say $K = 1, \dots, 5$, and calculate the RMSEs. The first 11 months are employed for model training,

and the last month is used for prediction evaluation. The in-sample RMSEs, defined as $\text{RMSE}_{esti} = \{(N(T - 31))^{-1} \sum_{t=1}^{T-31} \|\widehat{\mathbf{Y}}_t - \mathbf{Y}_t\|^2\}^{1/2}$, are 2.81, 2.78, 2.77, 2.77, 2.76 for each K , respectively, and the out-of-sample RMSEs, defined as $\text{RMSE}_{pred} = \{(31N)^{-1} \sum_{t=T-30}^T \|\widehat{\mathbf{Y}}_t - \mathbf{Y}_t\|^2\}^{1/2}$, are 2.50, 2.51, 2.51, 2.51, 2.50, respectively. It appears that 3, 4, 5 seems to be reasonable candidates of K . We then try the clustering method, i.e., estimate the coefficient parameter $\boldsymbol{\theta}$ at the nodal level and apply k -means clustering to partition these N sets of estimates into K groups. Fig. 5 below illustrates the selection of the optimal number of groups based on the Elbow plot, Silhouette coefficient, and gap statistic, respectively. We can see that all measures recommend $K = 3$. Thus, $K = 3$ is chosen in the following analysis.

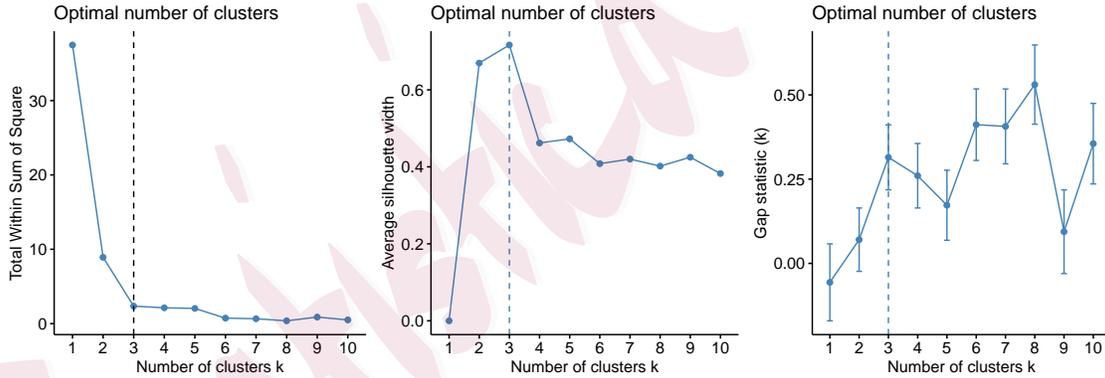


Figure 5: The selection of optimal number of groups based on the Elbow plot, Silhouette coefficient, and gap statistic, respectively.

We fit the GNPARG model (2.3) to the dataset with $K = 3$. The results are summarized in Table 5. We can see that for all groups, the estimated regression coefficient on past intensity process $\hat{\beta}$, appears to be much bigger than the regression coefficient on past observations $\hat{\alpha}$ and the network effect $\hat{\rho}$, implying that districts with large (small) number of investigatory stops are more likely to have large (small) number of investigatory stops in the future.

Table 5: Estimation results for ISR data using model (2.3) with $K = 3$.

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\beta}$	$\hat{\gamma}$
Group 1	0.0503	0.0931	0.0140	0.8650	0.4125
Group 2	1.4464	0.1109	0.1124	0.3281	0.2689
Group 3	0.0527	0.0659	1.9e-04	0.9259	0.3186

Fig. 6 (a) plots the estimated group labels for each district, and Fig. 6 (b) displays a box-plot for the number of daily investigatory stops Y_{it} in a grouped manner. The proportion of districts in each group is 0.41, 0.27, and 0.32, respectively.

The three groups show distinct numbers of stops and patterns of dependence. The districts in Group 3 are mainly in the southwest part of the city, which coincide with areas with the highest level of crime risk in Fig. 3. Group 3 has the highest number of stops, and the intensity of the count does not depend on its surroundings but mostly on its past intensity. Group 1 contains the safest areas and it also

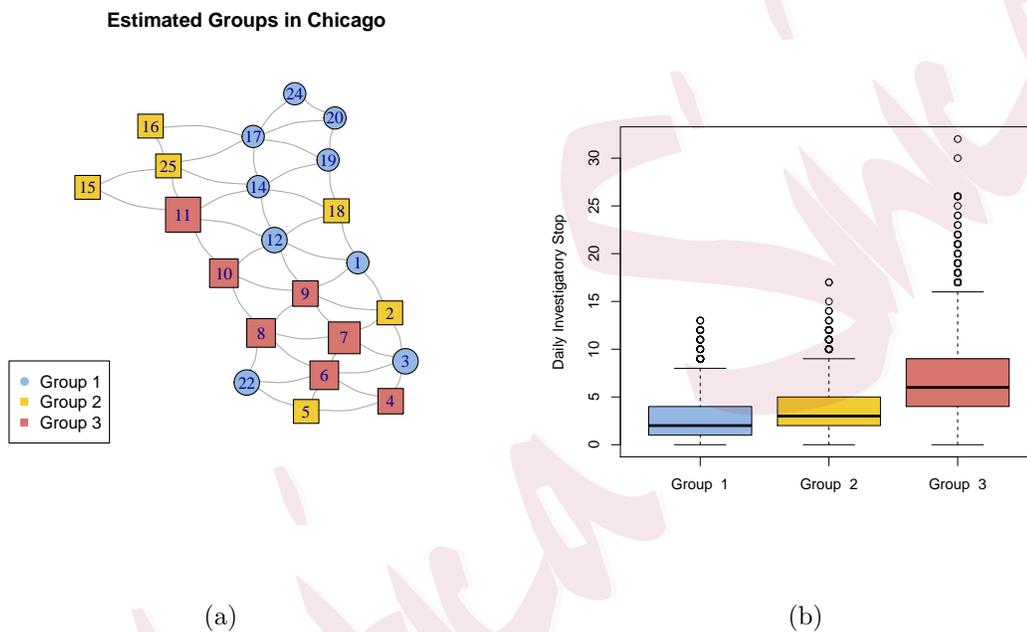


Figure 6: (a) Estimated group label for each district marked in different colors and shapes. The size of each node denotes the level of yearly investigatory stops, with bigger nodes representing more crimes occurred in that district. (b) Box plot for the number of daily investigatory stops in a grouped manner.

has a very small network effect, indicating that it is less likely to be affected by the surrounding areas. Both Group 1 and Group 3' future stops could be reliably predicted by their own historical information. On the contrary, Group 2 has moderate crime level and the districts are surrounding the high-risk areas. Group 2 has a relatively large network effect, which implies that the intensity of districts in Group 2 tend to be affected by events in their neighborhood.

The above observations also imply that although the network structure is symmetric, the mutual network effects between each pair of nodes could be different. For the districts in Group 3 that are connected to Group 2, Group 3 is hardly influenced by its neighbors, while Group 2 is more likely to be affected by events in their neighborhood. We guess that the latent flow network of population in Chicago is directed and asymmetric, but more data are needed to verify this conjecture in future study.

Further, we fit the NPAR model (2.2) that does not involve group structure on the same dataset for comparison. The estimation results are summarized in Table 6. For each estimator, the standard deviations are computed by $\mathbf{H}_T(\hat{\theta})^{-1}\mathbf{G}_T(\hat{\theta})\mathbf{H}_T(\hat{\theta})^{-1}$, where \mathbf{H}_T and \mathbf{G}_T are given in (S1.3) and (S1.4) in the Supplementary Materials. All estimates are statistically significant at 1% level. Still the momentum effect is much greater than the network effect. We can see that the group-wise information provided by the GNPARG model can shed more light on the real data and have more interpretability. The values of AIC for the GNPARG and NPAR model are 37340.57

Table 6: Estimation results for ISR data using NPAR model (2.2). The estimates, estimated standard deviations and p -values for each estimator are summarized.

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\rho}$	$\hat{\beta}$
Estimate	0.0128	0.0770	0.0033	0.9159
\widehat{SE}	0.0014	0.0015	7.1e-05	0.0017
p -value	< 0.001	< 0.001	< 0.001	< 0.001

and 37500.59, respectively, which suggests that GNPARG model fits the data better.

To sum up, in this case study, we have divided the districts in Chicago into three groups, and each group has its own feature of spatial and dynamic pattern of investigatory stops. We find that the spatial distribution of investigatory stops with enforcement action taken largely agrees with that of actual crime occurred, confirming the efficiency of the Investigatory Stop System of Chicago Police.

6. Conclusion and discussion

In this paper, we propose the grouped network Poisson autoregressive model. Compared with the traditional multivariate Poisson autoregressive model, our model has the following merits: (i). It incorporates the network information to reduce the number of unknown parameters and the computational complexity. (ii). The individual heterogeneity is introduced to describe different nodal behaviors for different groups,

which makes the model more flexible and realistic. (iii). The estimated group information and network effect can shed light on real social problems and lead to more practical interpretations.

Our model can be generalized in many ways. Firstly, we consider the linear form of Poisson autoregression, while log-linear form of Poisson autoregression is also very common in applications and could be easily extended to grouped case. Secondly, we assume the network structure to be fixed, but in practice the nodes may drop in and out of the model, and the association between nodes may change over time. Therefore, a time-varying network structure is worth studying. Thirdly, more covariates of the nodes or network structure information could be incorporated in the model for better fitting and group estimation. Lastly, in the existing methods, the network dimension N is fixed, and we study the asymptotic properties with increasing time sample size T . If N is diverging, the stationarity and ergodicity of count time series are unavailable under current methods, and the estimation could be problematic as the parameters grow quickly with the dimension of the matrix. It is still an open problem to be further studied.

Supplementary Materials

The online Supplementary Materials contain the technical proofs of Proposition 1, Theorem 1, several useful lemmas, as well as further simulation results when group

label is known, and the performance of the first K -selection method in **Remark 5**.

Acknowledgements

Tao and Li are supported in part by NSFC(No.71973077 and No.11771239). Niu is supported in part by NIH/NIAID R01 AI136664.

References

- Ahmad, A. and Francq, C. (2016). Poisson QMLE of count time series models. *Journal of Time Series Analysis* **37**, 291–314.
- Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653.
- Andreassen, C. M. (2013). *Models and inference for correlated count data*. Ph. D. thesis, Aarhus University, Denmark.
- Armillotta, M. and Fokianos, K. (2022). Poisson network autoregression. arXiv:2104.06296v3.
- Davis, R.A., Fokianos, K., Holan, S.H., Joe, H., Livsey, J., Lund, R., Pipiras, V. and Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association* **116**, 1533–1547.
- Davis, R.A. and Liu, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica* **26**, 1673–1707.

- Du, J.-G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis* **12**, 129–142.
- Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5**, 17–61.
- Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH processes. *Journal of Time Series Analysis* **27**, 923–942.
- Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association* **104**, 1430–1439.
- Fokianos, K., Støve, B., Tjøstheim, D. and Doukhan, P. (2020). Multivariate count autoregressions. *Bernoulli* **26**, 471–499.
- Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.
- Fokianos, K. and Tjøstheim, D. (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics* **64**, 1205–1225.
- Holland, P., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5**, 109–137.
- Huang, D., Wang, F., Zhu, X. and Wang, H. (2020). Two-mode network autoregressive model for large-scale networks. *Journal of Econometrics* **216**, 203–219.
- Karlis, D. and Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistics*

and Computing **15**, 255–265.

Karlis, D. and Meligkotsidou, L. (2007). Finite mixtures of multivariate Poisson regression with application.

Journal of Statistical Planning and Inference **137**, 1942–1960.

Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability* **29**, 228–248.

Lee, Y., Lee, S. and Tjøstheim, D. (2018). Asymptotic normality and parameter change test for bivariate

Poisson INGARCH models. *TEST* **27**, 52–69.

Liu, H. (2012). *Some models for time series of counts*. Ph. D. thesis, Columbia University, USA.

Mahamunulu, D. (1967). A note on regression in the multivariate Poisson distribution. *Journal of the*

American Statistical Association **62**, 251–258.

Meyn, S.P., Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*. Springer, London.

Neumann, M. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17**, 1268–

1284.

Tjøstheim, D. (2012). Some recent theory for autoregressive count time series. *TEST* **21**, 413–438.

Pedeli, X. and Karlis, D. (2013). On composite likelihood estimation of a multivariate INAR(1) model.

Journal of Time Series Analysis **34**, 206–220.

Wang, C., Liu, H., Yao, J.-F., Davis, R.A. and Li, W.K. (2014). Self-excited threshold Poisson autoregres-

sion. *Journal of the American Statistical Association* **109**, 777–787.

Wei, C.H. (2018). *An Introduction to Discrete-Valued Time Series*. Wiley, Hoboken.

Zhou, J., Li, D., Pan, R. and Wang, H. (2020). Network GARCH model. *Statistica Sinica* **30**, 1–18.

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

Zhu, X., Chang, X., Li, R. and Wang, H. (2019). Portal nodes screening for large scale social networks. *Journal of Econometrics* **209**, 145–157.

Zhu, X., Wang, W., Wang, H. and Härdle, W.K. (2019). Network quantile autoregression. *Journal of Econometrics* **212**, 345–358.

Zhu, X., Huang, D., Pan, R. and Wang, H. (2020). Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics* **215**, 591–606.

Zhu X. and Pan R. (2020). Grouped network vector autoregression. *Statistica Sinica* **30**, 1437–1462.

Zhu, X., Cai, Z. and Ma, Y. (2021). Network functional varying coefficient model. *Journal of the American Statistical Association*. (In press)

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

E-mail: taoyx19@mails.tsinghua.edu.cn

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

E-mail: malidong@tsinghua.edu.cn

Department of Statistics, Pennsylvania State University, University Park, PA, USA

E-mail: xiaoyue@psu.edu