Grouped Network Poisson Autoregressive Model

Yuxin Tao[1], Dong Li[1], and Xiaoyue Niu[2]

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

Supplementary Material

Here, we present technical proofs of Proposition 1 and Theorem 1, as well as several useful lemmas, as well as further simulation results when a group label is known and the performance of the first K-selection method in **Remark 5**.

S1. Technical Proofs

In this section, the notation C refers to a generic constant and may take different values in different places, and ν_m denotes a positive sequence with $\nu_m \to 0$ as $m \to \infty$. Before the proofs, we first give some expressions that will be used in the proofs.

Recall that the log-likelihood function (ignoring the constant) is

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

Then, the score function is

$$\begin{aligned} S_T(\boldsymbol{\theta}_k) &= \frac{1}{N_kT}\sum_{t=1}^{T}\sum_{i\in\mathcal{G}_k}\Big(\frac{Y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta}_k)} - 1\Big)\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}_k} \\ &= \frac{1}{N_kT}\sum_{t=1}^{T}\frac{\partial\boldsymbol{\lambda}_t^{(k)\prime}(\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}_k}\mathbf{C}_t^{-1}(\boldsymbol{\theta}_k)\big(\mathbf{Y}_t^{(k)} - \boldsymbol{\lambda}_t^{(k)}(\boldsymbol{\theta}_k)\big) := \frac{1}{N_kT}\sum_{t=1}^{T}s_t(\boldsymbol{\theta}_k), \end{aligned} \tag{S1.1}$$

where $\mathbf{C}_t(\boldsymbol{\theta}_k) = \mathrm{diag}(\boldsymbol{\lambda}_t^{(k)}(\boldsymbol{\theta}_k))$ is an $N_k \times N_k$ diagonal matrix, $\partial \boldsymbol{\lambda}_t^{(k)}/\partial \boldsymbol{\theta}_k'$ is an $N_k \times 4$ matrix, and

$$\begin{aligned} \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \omega_k} &= \mathbf{1}_{N_k} + \beta_k \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \omega_k}; \\ \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \alpha_k} &= \mathbf{Y}_{t-1}^{(k)} + \beta_k \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \alpha_k}; \\ \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \rho_k} &= (\mathbf{D}^{(k)})^{-1}\mathbf{A}^{(k)}\mathbf{Y}_{t-1} + \beta_k \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \rho_k}; \\ \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \beta_k} &= \boldsymbol{\lambda}_{t-1}^{(k)} + \beta_k \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \beta_k}. \end{aligned} \tag{S1.2}$$

The Hessian matrix is defined as

$$\mathbf{H}_T(\boldsymbol{\theta}_k) = \frac{1}{N_k T}\sum_{t=1}^{T}\sum_{i\in\mathcal{G}_k} \frac{Y_{i;t}}{\lambda_{i,t}^2(\boldsymbol{\theta}_k)} \frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}\frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k'} - \frac{1}{N_k T}\sum_{t=1}^{T}\sum_{i\in\mathcal{G}_k}\left(\frac{Y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta}_k)} - 1\right)\frac{\partial^2 \lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'}. \tag{S1.3}$$

Thus, the conditional information matrix is given by

$$\begin{aligned} \mathbf{G}_T(\boldsymbol{\theta}_k) &= \frac{1}{N_k T}\sum_{t=1}^{T}\sum_{i\in\mathcal{G}_k}\sum_{j\in\mathcal{G}_k} \frac{1}{\lambda_{i,t}(\boldsymbol{\theta}_k)\lambda_{j,t}(\boldsymbol{\theta}_k)} \boldsymbol{\Sigma}_{ij,t}^{(k)}(\boldsymbol{\theta}_k) \frac{\partial \lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}\frac{\partial \lambda_{j,t}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k'} \\ &= \frac{1}{N_k T}\sum_{t=1}^{T} \frac{\partial \boldsymbol{\lambda}_t^{(k)\prime}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k} \mathbf{C}_t^{-1}(\boldsymbol{\theta}_k)\boldsymbol{\Sigma}_t^{(k)}(\boldsymbol{\theta}_k)\mathbf{C}_t^{-1}(\boldsymbol{\theta}_k)\frac{\partial \boldsymbol{\lambda}_t^{(k)}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k'}, \end{aligned} \tag{S1.4}$$

where $\boldsymbol{\Sigma}_t^{(k)}(\cdot)$ denotes the true covariance matrix of $\mathbf{Y}_t^{(k)}$, and $\boldsymbol{\Sigma}_{ij,t}^{(k)}(\cdot)$ is the $(i,j)^{th}$ entry of $\boldsymbol{\Sigma}_t^{(k)}(\cdot)$. Particularly, when the components of the process $\{\mathbf{Y}_t^{(k)}\}$ are uncorrelated, then $\boldsymbol{\Sigma}_t^{(k)}(\boldsymbol{\theta}_k) = \mathbf{C}_t(\boldsymbol{\theta}_k)$.

The third order partial derivative of the log-likelihood $\partial^3 l_t(\boldsymbol{\theta}_k)/\partial\theta_{k,1}\partial\theta_{k,2}\partial\theta_{k,3}$ is given by

$$\begin{aligned} \frac{\partial^3 l_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}\partial\theta_{k,2}\partial\theta_{k,3}} = &-\frac{1}{N_k}\Big(\frac{Y_{i,t}}{\lambda_{i,t}^2(\boldsymbol{\theta}_k)}\Big)\left(\frac{\partial^2\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}\partial\theta_{k,2}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,3}} + \frac{\partial^2\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}\partial\theta_{k,3}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,2}} + \frac{\partial^2\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,2}\partial\theta_{k,3}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}}\right) \\ &+\frac{2}{N_k}\Big(\frac{Y_{i,t}}{\lambda_{i,t}^3(\boldsymbol{\theta}_k)}\Big)\left(\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,2}}\frac{\partial\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,3}}\right) + \frac{1}{N_k}\Big(\frac{Y_{i,t}}{\lambda_{i,t}(\boldsymbol{\theta}_k)} - 1\Big)\frac{\partial^3\lambda_{i,t}(\boldsymbol{\theta}_k)}{\partial\theta_{k,1}\partial\theta_{k,2}\partial\theta_{k,3}}, \end{aligned}$$

where $\theta_{k,1}, \theta_{k,2}, \theta_{k,3} \in \{\omega_k, \alpha_k, \rho_k, \beta_k\}$.

As discussed in Section 2.3, the main problem is that the sufficient condition on stationarity and ergodicity for the unperturbed model is useless to obtain the asymptotics of $\widehat{\boldsymbol{\theta}}_k$ (see, e.g., Fokianos, Rahbek, and Tjøstheim (2009), Fokianos and Tjøstheim (2011) for detailed discussion). Thus we borrow the corresponding condition for the perturbed model and then show that the two models are "close" in some sense. We define analogously $S_T^m(\boldsymbol{\theta}_k), \mathbf{H}_T^m(\boldsymbol{\theta}_k), \mathbf{G}_T^m(\boldsymbol{\theta}_k)$ to be the corresponding score function, Hessian matrix, and conditional information matrix for the perturbed model (2.5) with $\left(\mathbf{Y}_t^{(k)}, \boldsymbol{\lambda}_t^{(k)}\right)$ being replaced by $\left((\mathbf{Y}_t^m)^{(k)}, (\boldsymbol{\lambda}_t^m)^{(k)}\right)$, $1 \le k \le K$. Finally, Theorem 1 follows immediately from Lemmas 2–4 below, which verify the conditions of Theorem 3.2.23 in Taniguchi and Kakizawa (2000, Chap. 3).

S1.1 Some Lemmas

To prove the asymptotics of the MLE, we first give several lemmas.

Lemma 1. *For models (2.4)–(2.5), if* $|||\max_{1\le k\le K}(\alpha_k+\beta_k)\mathbf{I}_N + \max_{1\le k\le K}\rho_k\mathbf{D}^{-1}\mathbf{A}|||_2 < 1$ *holds, then*

(i). $\left\|\mathrm{E}\left(\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t\right)\right\|_2 = \left\|\mathrm{E}\left(\mathbf{Y}_t^m - \mathbf{Y}_t\right)\right\|_2 \le \delta_{1,m}$;

(ii). $\mathrm{E}\left\|\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t\right\|_2^2 \le \delta_{2,m}$;

(iii). $\mathrm{E}\left\|\mathbf{Y}_t^m - \mathbf{Y}_t\right\|_2^2 \le \delta_{3,m}$,

where $\delta_{i,m} \to 0$, $i = 1,2,3$, *as* $m \to \infty$. *In addition, for any* $\delta > 0$, $\left\|\boldsymbol{\lambda}_t^m - \boldsymbol{\lambda}_t\right\|_2 \le \delta$ *and* $\left\|\mathbf{Y}_t^m - \mathbf{Y}_t\right\|_2 \le \delta$ *a.s. for sufficiently large* m.

Proof. The proof is similar to that of Lemma 3.1 in Fokianos et al. (2020) and it is thus omitted. □

Lemma 2. *Let* $\mathbf{G}^m(\boldsymbol{\theta}_k) = N_k^{-1}\mathrm{E}\big(s_t^m(\boldsymbol{\theta}_k)s_t^m(\boldsymbol{\theta}_k)'\big)$ *and* $\mathbf{G}(\boldsymbol{\theta}_k) = N_k^{-1}\mathrm{E}\big(s_t(\boldsymbol{\theta}_k)s_t(\boldsymbol{\theta}_k)'\big)$. *If Assumptions 1–2 hold, then* $\mathbf{G}^m(\boldsymbol{\theta}_{k0}) \to \mathbf{G}(\boldsymbol{\theta}_{k0})$ *as* $m \to \infty$, *for each* $1 \le k \le K$.

Proof. Since all quantities are evaluated at the true value $\boldsymbol{\theta}_{k0}$, we suppress the notation that depends on $\boldsymbol{\theta}_k$ for simplicity. Similar to the proof of Lemma 4.1 in Fokianos et al. (2020) with $\boldsymbol{\lambda}_t^{m(k)}$, $\boldsymbol{\lambda}_t^{(k)}$ replacing their $\boldsymbol{\lambda}_t^m$, $\boldsymbol{\lambda}_t$, respectively, we have

$$\left\|\frac{\partial\boldsymbol{\lambda}_t^{m(k)}}{\partial\xi} - \frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\xi}\right\|_2 \le \nu_m, \quad \text{a.s.,} \quad \xi \in \{\omega_k, \alpha_k, \rho_k, \beta_k\}.$$

Next, we consider the operator norm of the matrix difference

$$\left|\left|\left|s_t^m\left(s_t^m\right)' - s_t s_t'\right|\right|\right|_2 \le \left\|s_t^m - s_t\right\|_2 \left\|\left(s_t^m\right)'\right\|_2 + \left\|s_t\right\|_2 \left\|\left(s_t^m - s_t\right)'\right\|_2.$$

Note that

$$\begin{aligned} s_t^m - s_t = & \left\{ \left(\frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} \right)' - \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \right\} (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \\ & + \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \left\{ (\mathbf{C}_t^m)^{-1} - (\mathbf{C}_t)^{-1} \right\} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \\ & + \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \mathbf{C}_t^{-1} \left\{ \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) - \left(\mathbf{Y}_t^{(k)} - \boldsymbol{\lambda}_t^{(k)} \right) \right\} \\ := & I_1 + I_2 + I_3. \end{aligned}$$

For I_1, it follows that

$$\|I_1\|_2 \leq \left\|\left| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right|\right\|_2 \left\|\left| (\mathbf{C}_t^m)^{-1} \right|\right\|_2 \left\| \mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right\|_2$$

with

$$\begin{aligned} \left\|\left| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} - \frac{\partial \boldsymbol{\lambda}_t{}^{(k)}}{\partial \boldsymbol{\theta}_k'} \right|\right\|_2^2 \leq & \left\|\left| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right|\right\|_F^2 \\ = & \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \omega_k} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \omega_k} \right\|_2^2 + \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \alpha_k} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \alpha_k} \right\|_2^2 + \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \rho_k} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \rho_k} \right\|_2^2 + \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \beta_k} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \beta_k} \right\|_2^2. \end{aligned}$$

It is not hard to prove that

$$\mathrm{E} \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \xi} - \frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \xi} \right\|_2^2 \leq \nu_m \to 0, \quad \xi \in \{\omega_k, \alpha_k, \rho_k, \beta_k\}.$$

As an example we here only prove the case $\xi = \rho_k$ and the other three cases can be proved similarly. Clearly, by Lemma 1 and $0 \leq \beta_k < 1$, it follows that

$$\begin{aligned} \mathrm{E} \left\| \frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \rho_k} - \frac{\partial \boldsymbol{\lambda}_t{}^{(k)}}{\partial \rho_k} \right\|_2^2 \leq & \left\|\left| \mathbf{D}^{(k)^{-1}} \mathbf{A}^{(k)} \right|\right\|_2^2 \mathrm{E} \left\| \mathbf{Y}_{t-1}^m - \mathbf{Y}_{t-1} \right\|_2^2 + \beta_k^2 \mathrm{E} \left\| \frac{\partial \boldsymbol{\lambda}_{t-1}^{m}{}^{(k)}}{\partial \rho_k} - \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \rho_k} \right\|_2^2 \\ & + 2 \left\|\left| \mathbf{D}^{(k)^{-1}} \mathbf{A}^{(k)} \right|\right\|_2 \beta_k \mathrm{E} \left\{ \left\| \mathbf{Y}_{t-1}^m - \mathbf{Y}_{t-1} \right\|_2 \left\| \frac{\partial \boldsymbol{\lambda}_{t-1}^{m}{}^{(k)}}{\partial \rho_k} - \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \rho_k} \right\|_2 \right\} \\ \leq & \left\|\left| \mathbf{D}^{(k)^{-1}} \mathbf{A}^{(k)} \right|\right\|_2^2 \delta_{3,m} + \beta_k^2 \mathrm{E} \left\| \frac{\partial \boldsymbol{\lambda}_{t-1}^{m}{}^{(k)}}{\partial \rho_k} - \frac{\partial \boldsymbol{\lambda}_{t-1}^{(k)}}{\partial \rho_k} \right\|_2^2 + 2\beta_k \left\|\left| \mathbf{D}^{(k)^{-1}} \mathbf{A}^{(k)} \right|\right\|_2 \nu_m \sqrt{\delta_{3,m}} \\ \leq & \nu_m. \end{aligned}$$

For the second term of I_1, it follows that

$$\left\|\left|(\mathbf{C}_t^m)^{-1}\right|\right\|_2 \le \max_{1\le i\le N} \frac{1}{\lambda_{i,t}} \le \frac{1}{\omega} \le C.$$

In addition,

$$\mathrm{E}\left\|\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}\right\|_2^2 = \sum_{i\in\mathcal{G}_k} \mathrm{E}\left(Y_{i,t}^m - \lambda_{i,t}^m\right)^2 = \sum_{i\in\mathcal{G}_k} \mathrm{E}\left\{\mathrm{E}\left(Y_{i,t}^m - \lambda_{i,t}^m\right)^2 \middle| \lambda_{i,t}\right\} = \sum_{i\in\mathcal{G}_k} \mathrm{E}\left(\lambda_{i,t}^m\right) < C.$$

Combining the above results and using the Cauchy-Schwartz inequality, we can get

$$\mathrm{E}\left\{\left\|\left|\frac{\partial\boldsymbol{\lambda}_t^{m(k)}}{\partial\boldsymbol{\theta}_k'} - \frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\boldsymbol{\theta}_k'}\right|\right\|_2 \left\|\left|(\mathbf{C}_t^m)^{-1}\right|\right\|_2 \left\|\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}\right\|_2\right\} \to 0, \quad \text{as } m\to\infty.$$

For I_2, note that

$$\mathrm{E}\left\|\left|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\boldsymbol{\theta}_k'}\right|\right\|_2^4 \le \mathrm{E}\left\|\left|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\boldsymbol{\theta}_k'}\right|\right\|_F^4 = \mathrm{E}\left\{\left\|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\omega_k}\right\|_2^2 + \left\|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\alpha_k}\right\|_2^2 + \left\|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\rho_k}\right\|_2^2 + \left\|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\beta_k}\right\|_2^2\right\}^2 < C,$$

where the last inequality is proved by the finity of each summand. Further, by Proposition 1,

$$\begin{aligned}
\mathrm{E}\left\|\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}\right\|_2^4 &= \mathrm{E}\Big[\Big\{\sum_{i\in\mathcal{G}_k}\left(Y_{i,t}^m - \lambda_{i,t}^m\right)^2\Big\}^2\Big] \\
&= \mathrm{E}\Big\{\sum_{i\in\mathcal{G}_k}\left(Y_{i,t}^m - \lambda_{i,t}^m\right)^4 + 2\sum_{i,j\in\mathcal{G}_k, i\neq j}\left(Y_{i,t}^m - \lambda_{i,t}^m\right)^2\left(Y_{j,t}^m - \lambda_{j,t}^m\right)^2\Big\} \\
&\le \sum_{i\in\mathcal{G}_k}\sum_{j=1}^4 c_{ij}\mathrm{E}\left\{\left(\lambda_{i,t}^m\right)^j\right\} < C
\end{aligned}$$

for some finite positive constants (c_{ij}).

Clearly,

$$\left\|\left|(\mathbf{C}_t^m)^{-1} - (\mathbf{C}_t)^{-1}\right|\right\|_2^2 \le \left\|\left|(\mathbf{C}_t^m)^{-1} - (\mathbf{C}_t)^{-1}\right|\right\|_F^2 \le C\left\|\boldsymbol{\lambda}_t^{m(k)} - \boldsymbol{\lambda}_t^{(k)}\right\|_2^2,$$

thus its expectation tends to zero by Lemma 1. Based on these results, we can get

$$\mathrm{E}\left\{\left\|\left|\frac{\partial\boldsymbol{\lambda}_t^{(k)}}{\partial\boldsymbol{\theta}_k'}\right|\right\|_2 \left\|\left|(\mathbf{C}_t^m)^{-1} - (\mathbf{C}_t)^{-1}\right|\right\|_2 \left\|\left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}\right)\right\|_2\right\} \to 0.$$

For I_3, by Lemma 1, the Cauchy-Schwartz inequality and the preceding results, we can similarly prove

$$\mathrm{E}\left\{\left\|\left|\frac{\partial\boldsymbol{\lambda}_t{}^{(k)}}{\partial\boldsymbol{\theta}_k'}\right|\right\|_2 \left\|\left|\mathbf{C}_t^{-1}\right|\right\|_2 \left\|\left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}\right) - \left(\mathbf{Y}_t^{(k)} - \boldsymbol{\lambda}_t^{(k)}\right)\right\|_2\right\} \to 0 \quad \text{as } m\to\infty.$$

Thus, $\mathrm{E}\left\|s_t^m - s_t\right\|_2 \to 0$ as $m \to \infty$.

Further, it is not hard to show that $\mathrm{E}\left\|s_t\right\|_2^2 < \infty$ and $\mathrm{E}\left\|(s_t^m)'\right\|_2^2 < \infty$. Thus, it follows that

$$\frac{1}{N_k}\mathrm{E}\left|\left\|s_t^m (s_t^m)' - s_t s_t'\right\|\right|_2 \to 0 \quad \text{as } m \to \infty.$$

The proof is complete. □

Lemma 3. *If Assumptions 1–2 hold, then the score functions for the unperturbed model (2.4) and perturbed one (2.5) evaluated at the true value $\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k0}$ satisfy, for each $1 \le k \le K$,*

(i). $S_T^m \xrightarrow{\text{a.s}} \mathbf{0}$, *as* $T \to \infty$*;*

(ii). $\sqrt{N_k T} S_T^m \xrightarrow{\text{d}} S^m \sim N\left(\mathbf{0}, \mathbf{G}^m\right)$, *as* $T \to \infty$*;*

(iii). $S^m \xrightarrow{\text{d}} N(\mathbf{0}, \mathbf{G})$, *as* $m \to \infty$*;*

(iv). $\lim_{m\to\infty} \limsup_{T\to\infty} P(\sqrt{N_k T}\left\|S_T^m - S_T\right\|_2 > \epsilon) = 0, \quad \forall \epsilon > 0.$

Proof. Recall that

$$s_t^m(\boldsymbol{\theta}_k) = \frac{\partial \boldsymbol{\lambda}_t^{m(k)'}(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}\mathbf{C}_t^{-1}(\boldsymbol{\theta}_k)\big(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)}(\boldsymbol{\theta}_k)\big),$$

we have $\mathrm{E}\big(s_t^m(\boldsymbol{\theta}_{k0})\big|\mathcal{F}_{t-1,m}^{\mathbf{Y},\boldsymbol{\lambda}}\big) = \mathbf{0}$, where $\mathcal{F}_{t-1,m}^{\mathbf{Y},\boldsymbol{\lambda}}$ denotes the σ-field generated by $\{\mathbf{Y}_{t-1}^{m}{}^{(k)}, \ldots, \mathbf{Y}_0^{m(k)}$, $\boldsymbol{\epsilon}_{t-1}^m, \ldots, \boldsymbol{\epsilon}_0^m\}$. Then $\{T S_T^m\}_{T\ge 1}$ for the perturbed model is a zero mean, square integrable martingale sequence, with $(s_t^m)_{t\ge 1}$ a martingale difference sequence. Note that $\mathrm{E}\left\|s_t^m\right\|_2^2 < \infty$, then, by the strong law of large numbers, $S_T^m \to \mathbf{0}$ a.s. as $T \to \infty$. Thus, (i) holds.

For (ii), by the Central Limit Theorem for martingale difference, it is easy to show (ii) holds by verifying the following conditions

$$\sum_{t=1}^T \mathrm{E}\left\{\left\|\frac{s_t^m}{\sqrt{N_k T}}\right\|_2^2 I\left(\left\|s_t^m\right\|_2 > \sqrt{N_k T}\delta\right)\Big|\mathcal{F}_{t-1,m}^{\mathbf{Y},\boldsymbol{\lambda}}\right\} < \frac{1}{N_k^2 T^2 \delta^2}\sum_{t=1}^T \mathrm{E}\left\{\left\|s_t^m\right\|_2^4 \Big|\mathcal{F}_{t-1,m}^{\mathbf{Y},\boldsymbol{\lambda}}\right\} \to 0$$

and

$$\frac{1}{N_k T}\sum_{t=1}^T \mathrm{Var}\left(s_t^m\Big|\mathcal{F}_{t-1,m}^{\mathbf{Y},\boldsymbol{\lambda}}\right) \xrightarrow{a.s.} \frac{1}{N_k}\mathrm{E}\left(s_t^m s_t^{m\prime}\right) := \mathbf{G}^m.$$

For (iii), it can easily be proved by Lemma 2 and Proposition 6.3.9 in Brockwell and Davis (1991) via the characteristic function procedure, and it is thus omitted.

For (iv), a simple algebraic calculation gives that

$$\begin{aligned}\sqrt{N_k T}(S_T^m - S_T) =& \frac{1}{\sqrt{N_k T}} \sum_{t=1}^{T} (s_t^m - s_t) \\ =& \frac{1}{\sqrt{N_k T}} \sum_{t=1}^{T} \left\{ \left(\frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} \right)' - \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \right\} (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \\ &+ \frac{1}{\sqrt{N_k T}} \sum_{t=1}^{T} \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \left\{ (\mathbf{C}_t^m)^{-1} - (\mathbf{C}_t)^{-1} \right\} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \\ &+ \frac{1}{\sqrt{N_k T}} \sum_{t=1}^{T} \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \mathbf{C}_t^{-1} \left\{ \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) - \left(\mathbf{Y}_t^{(k)} - \boldsymbol{\lambda}_t^{(k)} \right) \right\}.\end{aligned}$$

Note that

$$\begin{aligned}& P\left(\left\| \frac{1}{\sqrt{N_k T}} \sum_{t=1}^{T} \left\{ \left(\frac{\partial \boldsymbol{\lambda}_t^{m(k)}}{\partial \boldsymbol{\theta}_k'} \right)' - \left(\frac{\partial \boldsymbol{\lambda}_t^{(k)}}{\partial \boldsymbol{\theta}_k'} \right)' \right\} (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \right\|_2 > \epsilon \right) \\ \le & P\left(\frac{\nu_m}{\sqrt{N_k T}} \left\| \sum_{t=1}^{T} (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \right\|_2 > \epsilon \right) \\ \le & \frac{\nu_m^2}{\epsilon^2 N_k T} \mathrm{E} \left\| \sum_{t=1}^{T} (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \right\|_2^2 \\ = & \frac{\nu_m^2}{\epsilon^2 N_k T} \sum_{t=1}^{T} \mathrm{E} \left\| (\mathbf{C}_t^m)^{-1} \left(\mathbf{Y}_t^{m(k)} - \boldsymbol{\lambda}_t^{m(k)} \right) \right\|_2^2 \le C \nu_m^2 \to 0.\end{aligned}$$

The other two summands can be similarly proved to converge to zero in probability.

Thus, (iv) holds and the proof is complete. □

Lemma 4. *If Assumptions 1–2 hold, then the Hessian matrix for the unperturbed model (2.4) and perturbed one (2.5) evaluated at the true value $\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k0}$ satisfy, for each $1 \le k \le K$,*

(i). $\mathbf{H}_T^m \xrightarrow{\mathrm{p}} \mathbf{H}^m$ *as $T \to \infty$;*

(ii). $\lim_{m\to\infty} \limsup_{T\to\infty} P\left(N_k \, |||\mathbf{H}_T^m - \mathbf{H}_T|||_2 > \epsilon \right) = 0, \quad \forall \epsilon > 0.$

where $\mathbf{H}_T$ is defined in (S1.3) and analogously for $\mathbf{H}_T^m$, and $\mathbf{H}^m$ is defined analogous to $\mathbf{H}$ in (3.10).

Proof. The proof is similar to that of Lemma 3.3 in Fokianos, Rahbek, and Tjøstheim (2009) using decomposition technique and Theorem 1 in Jensen and Rahbek (2007), and it is thus omitted. □

S1.2 Proof of Proposition 1

For (i), by Propositions 3.1 in Fokianos et al. (2020), it suffices to prove that $|||\mathcal{B}_1 + \mathcal{B}_2\mathbf{D}^{-1}\mathbf{A} + \mathcal{B}_3|||_2 < 1$, i.e., $|||\sum_{k=1}^{K}(\alpha_k + \beta_k)\mathbf{Z}_k + \sum_{k=1}^{K}\rho_k\mathbf{Z}_k\mathbf{D}^{-1}\mathbf{A}|||_2 < 1$. Suppose $\mathbf{M}_1$, $\mathbf{M}_2$ are two general nonnegative matrix satisfying $\mathbf{0} \leq \mathbf{M}_1 \leq \mathbf{M}_2$. For each $m = 1, 2, \ldots$, we have $\mathbf{0} \leq \mathbf{M}_1^m \leq \mathbf{M}_2^m$, then $|||\mathbf{M}_1^m|||_F \leq |||\mathbf{M}_2^m|||_F$. By the Gelfand formula for the spectral radius of a matrix $\mathbf{M}$, i.e., $\rho(\mathbf{M}) = \lim_{m\to\infty} |||\mathbf{M}^m|||^{1/m}$, where $|||\cdot|||$ is any matrix norm, we can get that $\rho(\mathbf{M}_1) \leq \rho(\mathbf{M}_2)$ as $m \to \infty$. Since $|||\mathbf{M}|||_2 = \sqrt{\rho(\mathbf{M}'\mathbf{M})}$ and $\mathbf{M}_2'\mathbf{M}_2 - \mathbf{M}_1'\mathbf{M}_1 = \mathbf{M}_2'(\mathbf{M}_2 - \mathbf{M}_1) + (\mathbf{M}_2' - \mathbf{M}_1')\mathbf{M}_1 \geq 0$, it follows that $|||\mathbf{M}_1|||_2 \leq |||\mathbf{M}_2|||_2$. See also 8.1.P8 (page 524) of Horn and Johnson (2013, Chap. 8). Note that $\sum_{k=1}^{K}(\alpha_k + \beta_k)\mathbf{Z}_k + \sum_{k=1}^{K}\rho_k\mathbf{Z}_k\mathbf{D}^{-1}\mathbf{A}$ is nonnegative and $\boldsymbol{\theta}_k = (\omega_k, \alpha_k, \rho_k, \beta_k)'$ is assumed to be positive. Hence, we obtain that $|||\sum_{k=1}^{K}(\alpha_k + \beta_k)\mathbf{Z}_k + \sum_{k=1}^{K}\rho_k\mathbf{Z}_k\mathbf{D}^{-1}\mathbf{A}|||_2 \leq |||\max_{1\leq k\leq K}(\alpha_k + \beta_k)\mathbf{I}_N + \max_{1\leq k\leq K}\rho_k\mathbf{D}^{-1}\mathbf{A}|||_2 < 1$, which completes the proof.

For (ii), we can easily get that $|||\sum_{k=1}^{K}\alpha_k\mathbf{Z}_k + \sum_{k=1}^{K}\rho_k\mathbf{Z}_k\mathbf{D}^{-1}\mathbf{A}|||_1 + |||\sum_{k=1}^{K}\beta_k\mathbf{Z}_k|||_1 \leq |||\big(\max_{1\leq k\leq K}\alpha_k\big)\mathbf{I}_N + \big(\max_{1\leq k\leq K}\rho_k\big)\mathbf{D}^{-1}\mathbf{A}|||_1 + \max_{1\leq k\leq K}\beta_k < 1$. Thus, the result holds by Propositions 3.2 in Fokianos et al. (2020).

□

S1.3 Proof of Theorem 1

We here sketch the proof. By similar arguments in Lemma 3.4 in Fokianos, Rahbek, and Tjøstheim (2009), it is not hard to prove that all third order partial derivatives of the log-likelihood function of the perturbed model (2.5) are uniformly bounded in the neighborhood $O(\boldsymbol{\theta}_{k0})$ of the true parameter. Further, all third order partial derivatives of the log-likelihood of the perturbed model tend to their counterparts of the unperturbed model, which implies that the latter are uniformly bounded.

By Lemmas 3–4, the boundedness of all third order partial derivatives of the log-likelihood function of the unperturbed model (2.4), and Proposition 6.3.9 in Brockwell and Davis (1991), the conditions (A.1)–(A.3) of Lemma 1 in Jensen and Rahbek (2004) are verified. Thus, Theorem 1 holds. □

S2. Further Simulation Results

S2.1 Simulation results when a group label is known

To assess the finite-sample performance of the MLE when a group label is known, we consider two different types of network structure, 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 rth simulation to be $\widehat{\boldsymbol{\theta}}^{(r)} = (\widehat{\boldsymbol{\omega}}^{(r)}, \widehat{\boldsymbol{\alpha}}^{(r)}, \widehat{\boldsymbol{\rho}}^{(r)}, \widehat{\boldsymbol{\beta}}^{(r)})'$, where $1 \le r \le R$. The simulation results are summarized in Tables 1–2 for the Erdös–Rényi model and stochastic blockmodel, respectively.

We summarize the root mean square error (RMSE) and the coverage rate of the confidence interval for each estimator. Specifically, for the network effect coefficient ρ_1 in Group 1, the RMSE is calculated as $\text{RMSE}_{\rho_1} = \{R^{-1}\sum_{r=1}^{R}(\widehat{\rho}_1^{(r)} - \rho_1)\}^{1/2}$. The 95% confidence intervals for ρ_1 is $\text{CI}_{\rho_1}^{(r)} = \big(\widehat{\rho}_1^{(r)} - z_{0.975}\widehat{\text{SD}}_{\rho_1}, \widehat{\rho}_1^{(r)} + z_{0.975}\widehat{\text{SD}}_{\rho_1}\big)$, where $\widehat{\text{SD}}_{\rho_1}$ is the standard deviation of $\widehat{\rho}_1$, and z_α is the αth quantile of a standard normal distribution. Then, the coverage probability is defined as $\text{CP}_{\rho_1} = R^{-1}\sum_{r=1}^{R} I\big(\rho_1 \in \text{CI}_{\rho_1}^{(r)}\big)$, where $I(\cdot)$ is an indicator function. The performance of other estimators are evaluated similarly.

From Tables 1–2 for the two network structures, we find that the RMSEs are all very small for all estimators $\widehat{\boldsymbol{\theta}}_k = (\widehat{\omega}_k, \widehat{\alpha}_k, \widehat{\rho}_k, \widehat{\beta}_k)', 1 \le k \le K$. As the network dimension N and sample size T increase, the estimators perform better with smaller RMSEs, which implies more accurate estimates and smaller standard deviations. These results largely agree with the theoretical ones. Note that estimators in Group 1 always perform better than the other two groups, because it has the largest proportion of nodes. Moreover, the coverage probabilities are close to the nominal level 95%. It can be seen that the estimated standard deviations $\widehat{\text{SD}}$ approximate the true ones well.

To see the overall performance of $\widehat{\boldsymbol{\theta}}$, Fig. 1 plots the histograms of standardized $\widehat{\boldsymbol{\theta}}_1$ with $N = 100$ and $T = 400$ for the stochastic block network structure. From Fig. 1, we can see that the empirical densities of each estimator are very close to normal ones. The results strongly support the consistency and asymptotic

Table 1: Simulation results for the Erdös–Rényi model. The RMSEs ($\times 10^2$) for each estimator are reported with their coverage rates (%) in the parentheses.

N	T	ω_1	α_1	ρ_1	β_1	ω_2	α_2	ρ_2	β_2	ω_3	α_3	ρ_3	β_3
20	100	7.48	4.16	4.42	10.76	18.60	5.24	5.35	14.31	43.71	5.75	7.56	14.16
		(90.2)	(95.1)	(95.6)	(97.3)	(91.2)	(94.3)	(94.7)	(94.2)	(92.1)	(94.9)	(95.6)	(93.4)
	200	5.44	2.99	3.21	7.71	13.26	3.50	3.55	9.68	27.44	4.25	5.18	9.43
		(90.4)	(95.3)	(94.3)	(94.4)	(90.8)	(94.7)	(94.5)	(95.3)	(93.2)	(95.0)	(92.9)	(94.0)
	400	4.26	2.15	2.24	5.45	9.48	2.45	2.51	6.77	19.34	2.83	3.92	6.52
		(87.6)	(95.3)	(94.2)	(95.4)	(90.5)	(95.0)	(94.7)	(94.3)	(92.9)	(95.0)	(92.4)	(95.0)
50	100	6.10	3.29	3.81	10.24	18.16	3.66	5.10	13.8	33.57	4.28	5.34	10.95
		(93.6)	(95.1)	(95.3)	(97.9)	(92.3)	(94.7)	(95.1)	(91.7)	(93.7)	(94.4)	(96.2)	(93.6)
	200	4.58	2.28	2.71	7.46	13.10	2.56	3.58	9.94	23.1	3.10	3.65	7.48
		(94.6)	(95.3)	(95.2)	(95.5)	(94.3)	(94.4)	(94.7)	(94.8)	(93.7)	(94.6)	(95.2)	(94.5)
	400	3.08	1.68	1.98	5.35	8.55	1.81	2.61	6.72	15.65	2.18	2.62	5.29
		(95.2)	(95.0)	(94.9)	(95.0)	(94.7)	(94.6)	(94.7)	(95.1)	(94.5)	(95.2)	(94.5)	(95.1)
100	100	7.47	2.89	3.41	9.60	14.37	3.40	3.72	11.62	29.75	3.92	3.69	9.69
		(93.6)	(93.9)	(95.2)	(94.0)	(94.2)	(93.9)	(94.4)	(94.2)	(93.0)	(94.6)	(95.0)	(94.4)
	200	5.08	2.02	2.50	7.10	9.88	2.46	2.61	8.27	19.68	2.89	2.56	6.82
		(93.9)	(94.5)	(94.8)	(94.9)	(95.0)	(94.6)	(94.8)	(94.5)	(94.0)	(95.1)	(95.2)	(94.9)
	400	3.76	1.43	1.76	5.05	7.17	1.70	1.80	5.83	13.79	1.96	1.74	4.71
		(94.1)	(95.3)	(94.7)	(94.3)	(94.0)	(95.8)	(95.0)	(94.3)	(94.2)	(95.3)	(95.1)	(94.9)

Table 2: Simulation results for the stochastic blockmodel. The RMSEs ($\times 10^2$) for each estimator are reported with their coverage rates (%) in the parentheses.

N	T	ω_1	α_1	ρ_1	β_1	ω_2	α_2	ρ_2	β_2	ω_3	α_3	ρ_3	β_3
20	100	8.07	4.12	4.07	10.06	17.76	4.90	5.69	14.72	48.72	5.90	5.73	13.39
		(91.7)	(94.8)	(95.4)	(94.8)	(92.3)	(94.3)	(95.0)	(94.1)	(92.5)	(94.3)	(94.9)	(93.3)
	200	5.90	2.82	2.97	7.38	11.94	3.65	4.20	10.23	30.34	4.02	3.88	9.01
		(89.9)	(94.7)	(95.3)	(95.5)	(94.5)	(94.8)	(94.4)	(94.5)	(91.7)	(95.4)	(94.3)	(93.9)
	400	4.71	2.09	2.21	5.29	8.75	2.41	2.94	7.14	21.12	2.90	2.76	6.33
		(87.1)	(95.2)	(93.5)	(94.6)	(94.1)	(94.8)	(93.0)	(94.7)	(91.4)	(94.7)	(95.4)	(95.5)
50	100	6.21	3.29	3.95	10.44	15.34	3.82	4.35	12.40	33.18	4.42	6.15	10.83
		(94.0)	(94.2)	(96.2)	(98.1)	(93.7)	(94.8)	(95.1)	(93.5)	(93.4)	(94.9)	(97.3)	(94.4)
	200	4.34	2.31	2.91	7.61	10.35	2.69	3.12	8.78	21.99	3.17	4.34	7.43
		(94.9)	(94.3)	(94.5)	(95.3)	(94.4)	(95.0)	(94.6)	(95.5)	(93.9)	(94.6)	(95.0)	(94.7)
	400	2.93	1.69	1.97	5.26	7.42	1.86	2.10	6.14	14.83	2.21	2.92	5.07
		(95.4)	(94.9)	(94.7)	(94.9)	(93.7)	(95.5)	(95.0)	(94.4)	(95.0)	(94.7)	(94.4)	(94.4)
100	100	5.69	2.84	3.38	9.41	13.60	3.36	3.79	11.78	28.25	4.05	3.67	9.45
		(93.9)	(95.0)	(94.8)	(94.1)	(93.8)	(94.9)	(94.8)	(94.0)	(93.9)	(95.0)	(94.4)	(94.2)
	200	4.09	2.08	2.55	6.93	9.70	2.37	2.82	8.58	18.39	2.73	2.44	6.37
		(94.4)	(94.5)	(94.6)	(94.4)	(95.3)	(94.5)	(95.4)	(95.3)	(95.2)	(94.8)	(94.6)	(94.6)
	400	2.82	1.52	1.76	4.77	6.58	1.70	1.93	5.85	13.51	1.96	1.76	4.61
		(95.1)	(94.7)	(95.7)	(95.0)	(94.9)	(95.8)	(95.8)	(95.8)	(94.9)	(95.3)	(94.5)	(95.0)

normality of the MLE for our proposed model again.

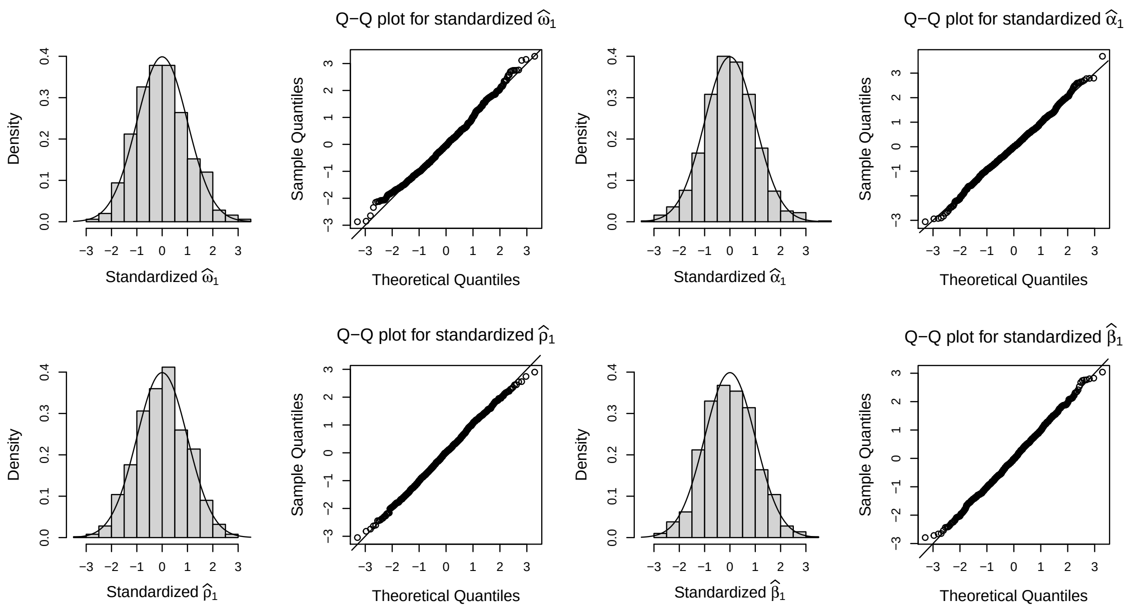

Figure 1: The histograms and Q-Q plots of standardized $\widehat{\boldsymbol{\theta}}_1$ with network dimension $N = 100$ and sampling size $T = 400$ under the stochastic block network structure. The curve in the histogram denotes the density of standard normal distribution.

S2.2 The performance of the clustering method of K-selection

In this subsection, we study the performance of the first method of selecting K proposed in **Remark 5**, 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. The optimal number of groups is chosen based on classical statistics in clustering. The true number of groups is $K = 3$ with parameters in Table 3. The data is generated under the stochastic blockmodel, and the network size is $N = 20, 50, 100$ and sample size is $T = 100, 200, 400$, each with $R = 1000$ replicates.

Here we use the **NbClust** package in R, which provides up to 30 indices for determining the number

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

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

of clusters, including the Silhouette coefficient, gap statistic, etc (see Charrad et al. (2014)). Table 4 reports the number of indices which recommend K as the optimal number of groups, taking average on 1000 replicates, where $1 \leq K \leq 10$. We can see that $K = 2$ and $K = 3$ are recommended the most in all scenarios, with more than 6 indices. When $T = 400$, the number of group K is correctly estimated as 3, while when $T = 100$ or 200, $K = 2$ has slightly more recommendations than $K = 3$. Thus, this method of selecting K has better performance with larger T and N in practice. When T or N is small, we could combine this method with the model fitting criterion (the second method in **Remark 5**) to select a reasonable number of groups.

References

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*. 2nd Edn. Springer, New York.

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* **61**, 1–36.

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.

Table 4: The number of indices which recommend K as the optimal number of groups, taking average on 1000 replicates, where $1 \leq K \leq 10$. (The most recommended K are marked in bold.)

N	T	$K=1$	2	3	4	5	6	7	8	9	10
20	100	0.26	**8.63**	6.20	2.22	1.27	1.14	0.93	1.11	1.50	4.73
	200	0.31	**8.56**	6.61	1.92	1.42	1.17	0.96	1.21	1.51	4.33
	400	0.40	7.70	**8.34**	1.54	1.29	1.29	1.12	1.00	1.38	3.95
50	100	0.14	**9.79**	6.81	2.66	1.29	1.08	0.80	1.34	0.73	3.37
	200	0.21	**9.84**	7.31	1.96	0.95	1.17	0.68	1.66	0.59	3.64
	400	0.29	8.62	**11.03**	1.07	0.48	0.87	0.41	1.80	0.24	3.19
100	100	0.12	**9.79**	7.03	2.93	1.77	1.10	1.35	0.85	1.08	1.99
	200	0.18	**9.81**	7.84	2.06	1.51	1.30	1.78	0.76	0.92	1.83
	400	0.36	8.00	**12.34**	1.15	0.61	0.96	2.24	0.47	0.57	1.30

Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression. *Journal of Multivariate Analysis* **102**, 563–578.

Horn, R.A. and Johnson, C.R. (2013). *Matrix Analysis.* 2nd Edn. Cambridge University Press, Cambridge.

Jensen, S.T. and Rahbek, A. (2004). Asymptotic inference for nonstationary GARCH. *Econometric Theory* **20**, 1203–1226.

Jensen, S.T. and Rahbek, A. (2007). On the law of large numbers for (geometrically) ergodic Markov chains. *Econometric Theory* **23**, 761–766.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer, New York.