Semiparametric spatio-temporal models with unknown and banded autoregressive coefficient matrices

Hongxia Wang[a], Xuehong Luo[a], Long Ling[b*]

([a]School of Statistics and Mathematics, Nanjing Audit University, Nanjing 211815, China [b]Department of Management Science and Engineering, Nanjing Normal University, Nanjing 210023, China)

Abstract

We consider a new class of semiparametric spatio-temporal models with unknown and banded autoregressive coefficient matrices. The setting represents a type of sparse structure in order to include as many panels as possible. We apply the local linear method and least squares method for Yule-Walker equation to estimate trend function and spatio-temporal autoregressive coefficient matrices respectively. We also balance the over-determined and under-determined phenomena in part by adjusting the order of extracting sample information. Both the asymptotic normality and convergence rates of the proposed estimators are established. The proposed methods are further illustrated using both simulation and case studies, the results also show that our estimator is stable among different sample size, and it performs better than the traditional method with known spatial weight matrices.

Keywords: Spatio-temporal autoregression; Unknown and banded coefficient matrices; Local linear estimation; Yule – Walker equation

1 Introduction

The demand of spatio-temporal prediction arises from panel studies of economics, air pollution analysis, epidemic phenomena, and various other fields. For example, we analyze the monthly air quality index (AQI) for Beijing-Tianjin-Hebei urban agglomeration of China in the period of Jan 2014 – Nov 2019. The detailed analysis for this data set will be presented

*Corresponding author. Email:linglong0206@126.com.

in Section 5. Fig. 1 depicts the AQI in a map for consecutive months, it's noticeable that the AQI level is analogous in region and each month's AQI level is similar to that of the last month, those fully show the spatial effect and dynamic effect. Correlation and heterogeneity are

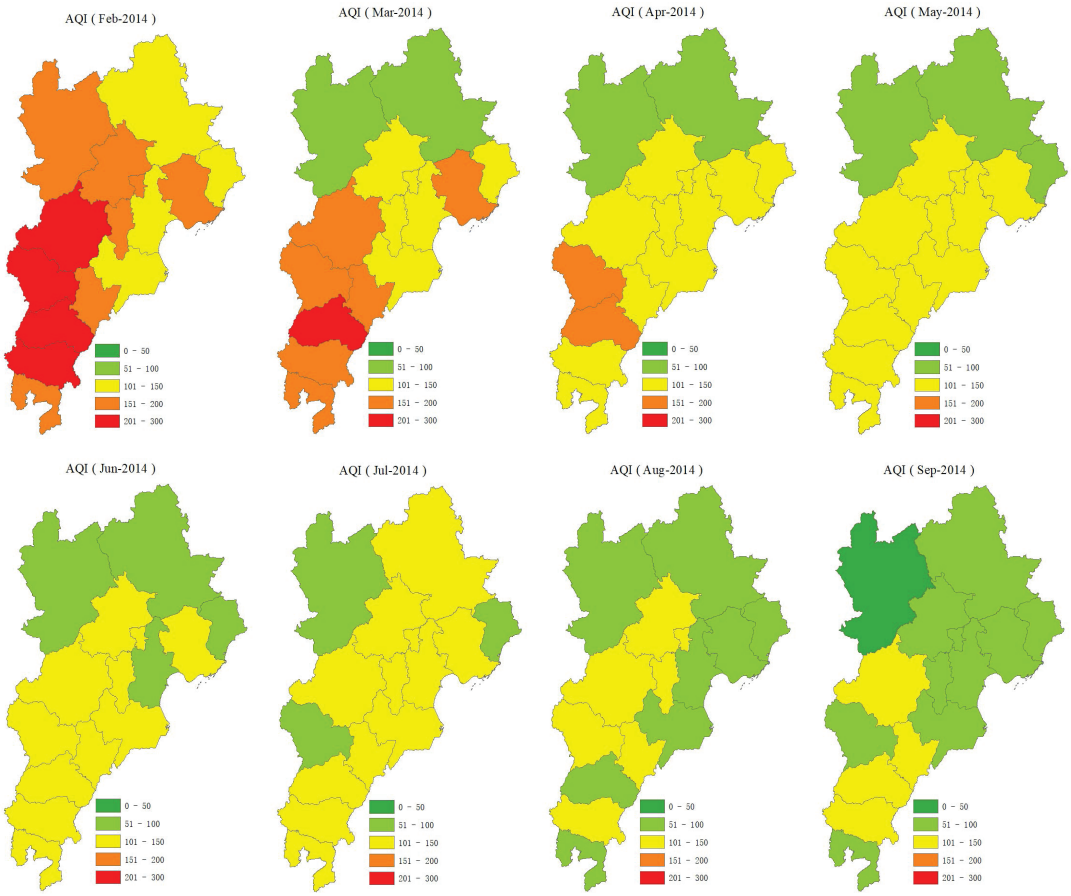

Figure 1: Maps of the monthly AQI in the period of Feb-Sep 2014 for Beijing-Tianjin-Hebei urban agglomeration of China.

two significant features of spatio-temporal data. The class of spatio-temporal autoregression (STAR) models is the most common product to model cross sectional correlation among different locations, it was first introduced by Cliff and Ord (1973) and has been applied to various domains. For example, Deutsch and Ramos (1986) examined river flows with STAR model. Szummer and Picard (1996) used a STAR model to aid in the synthesis of images of phenomena such as moving water, fire, or other evolving textures. For the effect of spatial heterogeneity, Lin and Lee (2010) proposed the robust generalized method of moments (GMM) estimators for STAR model in the presence of heteroskedastic disturbances, and efficiency of estimator can be improved by constructing the optimal weighted estimation. Kelejian and Prucha (2010) defined instrumental variable estimators and gave some general theories concerning the joint asymptotic distribution of those estimators and the GMM estimator in STAR models with autoregressive and heteroskedastic disturbances. To tackle the panel data with time dependence, Fu and Li (2020) explored the association between socioeconomic indicators and global PM2.5 using a spatial econometric model coupled with a temporal weighted regression based on the hybrid

method.

In practice the spans of dependence among locations exists, Guo et al. (2016) showed that it is rare to collect the information enough from neighbour variables, and information from farther locations may become redundant. Hence it is pertinent to consider the setting with infinite location bands in the coefficient matrices. Notably, Dou et al. (2016) and Gao et al. (2019) investigated the feasibility of estimation when the pertinent spatial weight matrix is unknown, and showed that generalized Yule-Walker equations are viable and efficient. Due to the limits of parametric models, more recent developments extend the class of STAR models to nonparametric and semiparametric spatio-temporal autoregressive models. For example, Biau and Cadre (2004) and Hallin et al. (2004) proposed the local linear method to model spatial heterogeneity. Robinson (2010) applied the adaptive estimation method to estimate the semiparametric spatial autoregressive model with non-normal innovations. Ulteriorly, Wang et al. (2012) considered the following spatio-temporal models:

$$
\begin{aligned}
Y_{\boldsymbol{i},t} &= m(\boldsymbol{X}_{\boldsymbol{i},t}) + R_{\boldsymbol{i},t}, \\
\boldsymbol{R} &= \boldsymbol{\rho}\mathbf{W}\boldsymbol{R} + \boldsymbol{\varepsilon},
\end{aligned}
$$

where $\boldsymbol{X}_{\boldsymbol{i},t} \in \mathbb{R}^l, Y_{\boldsymbol{i},t} \in \mathbb{R}, \boldsymbol{i} \in \Lambda_{\boldsymbol{n}} = \{(1,\cdots,n_1)\times(1,\cdots,n_2)\times\cdots\times(1,\cdots,n_d)\}, t \in T_{\boldsymbol{n}} = (1,\cdots,n_0), \boldsymbol{n} = (n_0,n_1,\cdots,n_d)$, and $m(\cdot)$ is an unknown function, $\boldsymbol{\rho}$ is the coefficient need to be estimated, error term $\boldsymbol{\varepsilon}$ is independently identically normal distribution, the vector $\boldsymbol{R}$ is consists of $R_{\boldsymbol{i},t}$ which are positively correlated in a lexicographical order. $\mathbf{W}$ is the $n^*\times n^*$ weight matrix which measures the dependence and $n^* = \prod_{j=0}^{d} n_j$. The common practice in spatial econometrics to assume the main diagonal elements of known $\mathbf{W}$ are zero, see Moran (1948)and Wang et al. (2018). Motivated by the evidence in some practical cases, we extend the semiparametric spatio-temporal models with autoregressive errors above to following models by allowing spatial weight matrices are completely unknown but are assumed to be sparse, and the similar sparse structure can be found in Guo et al. (2016):

$$
\begin{aligned}
Y_{\boldsymbol{i},t} &= m(\boldsymbol{X}_{\boldsymbol{i},t}) + R_{\boldsymbol{i},t}, && (1)\\
R_{\boldsymbol{i},t} &= (\boldsymbol{d}_{\boldsymbol{i}}^0)^{\mathsf{T}}\boldsymbol{R}_t + (\boldsymbol{d}_{\boldsymbol{i}}^1)^{\mathsf{T}}\boldsymbol{R}_{t-1} + \cdots + (\boldsymbol{d}_{\boldsymbol{i}}^p)^{\mathsf{T}}\boldsymbol{R}_{t-p} + \varepsilon_{\boldsymbol{i},t}, && (2)
\end{aligned}
$$

where $\boldsymbol{X}_{\boldsymbol{i},t} \in \mathbb{R}^l$ and $Y_{\boldsymbol{i},t} \in \mathbb{R}$ represents the observations collected from every location we studied at time t, $\boldsymbol{X}_{\boldsymbol{i},t}$ has finite dimension l, $\boldsymbol{R}_t$ is consists of $R_{\boldsymbol{i},t}$ in a lexicographical order of $\boldsymbol{i}$, $\varepsilon_{\boldsymbol{i},t}$ is the independent and identically distributed innovation at time t of location $\boldsymbol{i}$, and $\boldsymbol{\varepsilon}_t$ denotes the vector consists of the corresponding $\varepsilon_{\boldsymbol{i},t}$, it has a zero mean and satisfies the condition $\mathrm{Cov}(\boldsymbol{R}_{t-j},\boldsymbol{\varepsilon}_t) = 0$ for all $j \geq 1$, and the unknown positive definite matrix $\boldsymbol{\Sigma}_\varepsilon$ denotes the variance of $\boldsymbol{\varepsilon}_t$. Furthermore we assume that $\boldsymbol{D}_j \equiv (\boldsymbol{d}_1^j,\cdots,\boldsymbol{d}_{\bar{n}}^j)^{\mathsf{T}} \equiv (d_{i_0,j_0}^j)$ are $\bar{n}\times\bar{n}$ unknown banded coefficient matrices and $\bar{n} = n^*/n_0$. The coefficient matrix $\boldsymbol{D}_0$ captures the pure spatial effect among different locations, other coefficient matrices $\boldsymbol{D}_j$ capture dynamic effect for $j \neq 0$. To the banded matrices we have $d_{i_0,j_0}^j = 0$ for all $1 \leq k_0 < |i_0 - j_0| < \bar{n}$, the bandwidth k_0 is an unknown integer which should be repeatedly determined in different cases, and $d_{i_0,i_0}^0 = 0$ for

$1 \leq i_0 \leq \bar{n}$, the latter is equal to the common practice of spatial weight matrix in econometrics. For simplicity, we assume that the autoregressive order p is known, as the order-determination problem has already been thoroughly studied, see, e.g., Chapter 4 of Lütkepohl (2007). Note that the condition $d^j_{i_0,j_0} = 0$ does not mean $\mathrm{Cov}(R_{i_0,t_1}, R_{j_0,t_2}) = 0$, and we don't require $\mathbf{\Sigma}_\varepsilon$ to be banded.

This study seeks to make contributions in the following three respects. Firstly, we consider the more flexible and reasonable semiparametric models, the unknown and banded setting in models comes with the decrease of the number of estimated parameters and weaken the curse of dimension, it also avoids the unaccurate issue of constructing spatial weight matrices. Secondly, we try to avoid over-determined and under-determined by adjusting the order of extracting the information from sample data, and adopt a version of marginal Bayesian information criterion to identify the true bandwidth under the order of autoregression p is known. Finally, both asymptotic distribution and convergence rate of estimators are established in the new class of models under mild conditions. We develop the asymptotic normality of nonparametric predictors and the estimated coefficients under a general setting for stationary and α-mixing processes, and the convergence rates of the estimators are same with $1/\sqrt{n_0}$. The analysis results of simulation and real data sets indicate that the proposed methodology performs well.

The rest of the paper is organized as follows. Section 2 is the main part of the paper, where the new class of models and estimation method are specified. The asymptotic properties are stated in Section 3. Numerical results on simulated data and real data are reported in Section 4 and 5 respectively. In the Appendix we present all the technical proofs.

To guarantee the desired results, we make some notations here. For a $n \times n$ matrix $\mathbf{H}$ is consist of $h_{i,j}$, $\lambda_{\max}(\mathbf{H})$ and $\lambda_{\min}(\mathbf{H})$ denote the largest eigenvalue and the smallest eigenvalue of matrix $\mathbf{H}$ respectively, $\|\mathbf{H}\|_1 = \max_j \sum_i |h_{i,j}|$ and $\|\mathbf{H}\|_2 = \sqrt{\lambda_{\max}(\mathbf{H}^\mathsf{T}\mathbf{H})}$ are the L_1 norm and Euclidean L_2 norm respectively, $\mathbf{H}_{\max}$ is the largest value of matrix $\mathbf{H}$. For subset $S \subset \{1, \cdots, n\}$, let $\mathbf{H}_S$ be a sub-matrix consisting of the columns of $\mathbf{H}$ in S and $|S|$ be the cardinality of S. The letter c and C are used to denote constants whose values are unimportant and may vary from line to line.

2 Main methodology

2.1 Local linear estimator

The object is to estimate coefficient matrices $\boldsymbol{D}_j$ for all $j \in \{0, 1, \cdots, p\}$, and $R_{\boldsymbol{i},t}$ is not observed in process of collecting data, we propose below a new two-step estimation method which combines the local linear fitting method and the generalized Yule-Walker equation: first, getting the estimator $\widehat{R}_{\boldsymbol{i},t}$ of $R_{\boldsymbol{i},t}$, then obtaining all coefficient matrices with the least squares method to Yule-Walker equation.

We use Taylor expansion of $m(\boldsymbol{X}_{\boldsymbol{i},t})$ around $\boldsymbol{x}_0$ and local polynomial fitting method to

estimate $m(\boldsymbol{x}_0)$ in model (2):

$$\min \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}} \sum_{t\in T_{\boldsymbol{n}}} \left\{ Y_{\boldsymbol{i},t} - m(\boldsymbol{x}_0) - [m^{'}(\boldsymbol{x}_0)]^{\mathsf{T}}(\boldsymbol{X}_{\boldsymbol{i},t} - \boldsymbol{x}_0) \right\}^2 \boldsymbol{K}_{\boldsymbol{i},t},$$

where $\boldsymbol{K}_{\boldsymbol{i},t} = (1/(h_1 \cdots h_l))\boldsymbol{K}((\boldsymbol{X}_{\boldsymbol{i},t} - \boldsymbol{x}_0)/\boldsymbol{h})$ is the non-negative weight function on $\mathbb{R}^l$, the bandwidth of nonparametric method $\boldsymbol{h}$ is equal to $(h_1, \cdots, h_l)^{\mathsf{T}}$, and $(\boldsymbol{X}_{\boldsymbol{i},t} - \boldsymbol{x}_0)/\boldsymbol{h}$ is a vector representation of the difference between $\boldsymbol{X}_{\boldsymbol{i},t}(j)$ and $\boldsymbol{x}_0(j)$ for $j = 1, \cdots, l$, which are the element of $\boldsymbol{X}_{\boldsymbol{i},t}$ and $\boldsymbol{x}_0$ respectively, it has the short form of following formula:

$$\frac{\boldsymbol{X}_{\boldsymbol{i},t} - \boldsymbol{x}_0}{\boldsymbol{h}} = \left(\frac{\boldsymbol{X}_{\boldsymbol{i},t}(1) - \boldsymbol{x}_0(1)}{h_1}, \cdots, \frac{\boldsymbol{X}_{\boldsymbol{i},t}(l) - \boldsymbol{x}_0(l)}{h_l} \right)^{\mathsf{T}}.$$

Based on the results of the weighted least square method, we can obtain the estimator

$$\widehat{m}(\boldsymbol{x}_0) = \boldsymbol{e}_1^{\mathsf{T}} \left\{ \frac{1}{n^*} \boldsymbol{\mathcal{X}}^{\mathsf{T}} \mathbf{W}_0 \boldsymbol{\mathcal{X}} \right\}^{-1} \left\{ \frac{1}{n^*} \boldsymbol{\mathcal{X}}^{\mathsf{T}} \mathbf{W}_0 \boldsymbol{\mathcal{Y}} \right\} \equiv \boldsymbol{e}_1^{\mathsf{T}} \mathbf{U}_{\boldsymbol{n}}^{-1} \mathbf{J}_{\boldsymbol{n}}, \tag{3}$$

where

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} 1 & \left(\frac{\boldsymbol{X}_{1,1} - \boldsymbol{x}_0}{\boldsymbol{h}}\right)^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \left(\frac{\boldsymbol{X}_{\bar{n},n_0} - \boldsymbol{x}_0}{\boldsymbol{h}}\right)^{\mathsf{T}} \end{bmatrix}, \quad \boldsymbol{\mathcal{Y}} = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{\bar{n},n_0} \end{bmatrix},$$

and the weight matrix $\mathbf{W}_0$ =diag$(\boldsymbol{K}_{1,1}, \cdots, \boldsymbol{K}_{\bar{n},n_0})$, the $(l+1) \times 1$ unit vector $\boldsymbol{e}_1$ with 1 as its first element. On the other hand, applying the local polynomial fitting method to estimate error value of each observation location in model (1), the mechanism of weight function $\boldsymbol{K}_{\boldsymbol{i},t}$ will fail when $\boldsymbol{x}_0$ equal to $\boldsymbol{X}_{\boldsymbol{i},t}$, a common practice in practical studies is to use leave-one-out method to avoid that failure case, see, e.g., Linton and Xiao (2007). Then our objective here in practical cases is to solve the minimization problem

$$\min_{\boldsymbol{x}_0 \neq \boldsymbol{X}_{\boldsymbol{i},t}} \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}} \sum_{t\in T_{\boldsymbol{n}}} \left\{ Y_{\boldsymbol{i},t} - m(\boldsymbol{x}_0) - [m^{'}(\boldsymbol{x}_0)]^{\mathsf{T}}(\boldsymbol{X}_{\boldsymbol{i},t} - \boldsymbol{x}_0) \right\}^2 \boldsymbol{K}_{\boldsymbol{i},t}.$$

By Theorem 1 in Section 3, the resulting estimator of $R_{\boldsymbol{i},t}$ is a zero mean autoregressive random field which admits the following expression:

$$\widehat{R}_{\boldsymbol{i},t} = Y_{\boldsymbol{i},t} - \widehat{m}(\boldsymbol{X}_{\boldsymbol{i},t}).$$

In the first step, we use the methodology above to obtain $\hat{m}(\boldsymbol{X}_{\boldsymbol{i},t})$ in the process of estimate error term. Throughout this paper, $\boldsymbol{R}_t$ is referred to as a strictly stationary process defined by (2).

2.2 Generalized Yule-Walker estimator

With the banded condition of model above, the information from farther locations become redundant, though there may be non-zero correlations among all component series of $\boldsymbol{R}_t$. This reflects that bandwidth of banded model is the distance of dependency between different locations, and the optimal bandwidth parameter k_0 is case-dependent, we use a Bayesian information criterion method to estimate it in the following content.

Let $\boldsymbol{\Sigma}_j = \mathrm{Cov}(\boldsymbol{R}_{t+j}, \boldsymbol{R}_t)$ for any $j \geq 0$, and $E(\boldsymbol{R}_t) = 0$, then the Yule-Walker equations below follows from (2) and properties of stationary process that

$$\boldsymbol{\Sigma}_j = (\mathbf{I}_{\bar{n}} - \boldsymbol{D}_0)^{-1} \sum_{k=1}^{p} \boldsymbol{D}_k \boldsymbol{\Sigma}_{j+k}, \quad 0 \leq j < p,$$

$$\boldsymbol{\Sigma}_j = (\mathbf{I}_{\bar{n}} - \boldsymbol{D}_0)^{-1} \sum_{k=1}^{p} \boldsymbol{D}_k \boldsymbol{\Sigma}_{j-k}, \quad j \geq p,$$

which ensure the feasibility of second moment. Let $\mathbf{I}_{\bar{n}} - \boldsymbol{D}_0$ be invertible, where $\mathbf{I}_{\bar{n}}$ is a $\bar{n} \times \bar{n}$ identity matrix. To avoid the endogeneity and inconsistent estimators, we convert the original estimation problem to estimate the coefficient matrix of Yule-Walker equation of $\boldsymbol{\Sigma}_p$. Take the i_0th row of the $\boldsymbol{\Sigma}_p$,

$$\boldsymbol{\Sigma}_p^{\mathsf{T}} \boldsymbol{e}_{i_0} = \sum_{j=0}^{p} \boldsymbol{\Sigma}_{p-j}^{\mathsf{T}} \boldsymbol{D}_j^{\mathsf{T}} \boldsymbol{e}_{i_0} = \sum_{j=0}^{p} \boldsymbol{\Sigma}_{p-j}^{\mathsf{T}} \boldsymbol{d}_{i_0}^{j} \equiv \mathbf{V}_{i_0} \boldsymbol{\beta}_{i_0}, \quad i_0 = 1, \cdots, \bar{n}, \tag{4}$$

where $\boldsymbol{e}_{i_0}$ denotes the $\bar{n} \times 1$ unit vector with the i_0th element equal to 1. The $\tau_{i_0} \times 1$ vector $\boldsymbol{\beta}_{i_0}$ consists of $\boldsymbol{d}_{i_0}^{j}$ after removing the zero element for $j = 0, \cdots, p$, e.g., $\boldsymbol{\beta}_{i_0} = ((\boldsymbol{d}_{i_0}^{0'})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0}^{p'})^{\mathsf{T}})^{\mathsf{T}}$, in which there are some vectors $\boldsymbol{d}_{i_0}^{j'}$ with non-zero elements from $\boldsymbol{d}_{i_0}^{j}$. Similarly, we stack the corresponding columns of $\boldsymbol{\Sigma}_{p-j}^{\mathsf{T}}$ horizontally for $j = 0, \cdots, p$ to construct the $\bar{n} \times \tau_{i_0}$ matrix. For the banded area of a matrix, there are three regions divided by two boundaries, the number of elements is constant in the middle region, and the number of elements in the remaining two regions changes inversely with each other, then the actual formulas of τ_{i_0} can be written as

$$\tau_{i_0} \equiv \tau_{i_0}(k_0) = \begin{cases} (p+1)(k_0 + i_0) - 1 & 1 \leq i_0 \leq k_0, \\ (p+1)(2k_0 + 1) - 1 & k_0 < i_0 \leq \bar{n} - k_0, \\ (p+1)(k_0 + \bar{n} + 1 - i_0) - 1 & \bar{n} - k_0 < i_0 \leq \bar{n}. \end{cases} \tag{5}$$

Since the stationary process $\boldsymbol{R}_t$ is a zero mean field, we replace $\boldsymbol{\Sigma}_j$ by the sample (auto)covariance matrices

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p+j} \boldsymbol{R}_{t-p}^{\mathsf{T}}, \quad j = 0, \cdots, p, \tag{6}$$

which omit finite terms is reasonable to ensure the validity of (9) when $n_0 \to \infty$. Consequently, the least squares estimator of $\boldsymbol{\beta}_{i_0}$ based on (4) is

$$\widehat{\boldsymbol{\beta}}_{i_0} = (\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0})^{-1} \widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \widehat{\mathbf{z}}_{i_0}, \quad i_0 = 1, \cdots, \bar{n},$$

where $\widehat{\mathbf{z}}_{i_0} = \widehat{\boldsymbol{\Sigma}}_p^{\mathsf{T}} \boldsymbol{e}_{i_0}$, and the hat values denote the corresponding sample values. More explicitly,

$$\widehat{\mathbf{z}}_{i_0} = \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} R_{i_0,t} = \widehat{\mathbf{V}}_{i_0} \boldsymbol{\beta}_{i_0} + \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \boldsymbol{\varepsilon}_{i_0,t}, \tag{7}$$

$$\widehat{\mathbf{V}}_{i_0} = \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} (\boldsymbol{r}_T)^{\mathsf{T}}, \tag{8}$$

where the $1\times\tau_{i_0}$ vector $(\boldsymbol{r}_T)^{\mathsf{T}}=((\boldsymbol{r}_{T_p})^{\mathsf{T}},(\boldsymbol{r}_{T_{p-1}})^{\mathsf{T}},\cdots,(\boldsymbol{r}_{T_0})^{\mathsf{T}})$ and $\mathrm{T}_j=t-p+j$, the sub-vector $\boldsymbol{r}_{T_j}$ consists of R_{q,T_j} for $q\in S_{i_0}$ when $j=p$ and for $q\in S_{i_0}^{+}$ when $j\neq p$, where

$$S_{i_0}=\{q:1\leq q\leq\bar{n},1\leq|i_0-j_0|\leq k_0\}\quad\text{and}\quad S_{i_0}^{+}=\{q:1\leq q\leq\bar{n},|i_0-j_0|\leq k_0\}\,.$$

Then it holds that

$$\widehat{\boldsymbol{\beta}}_{i_0}-\boldsymbol{\beta}_{i_0}=\frac{1}{n_0}(\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0})^{-1}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t},\quad i_0=1,\cdots,\bar{n}. \tag{9}$$

In addition, the mean of corresponding residual sum of squares for the estimation of xth row of Yule-Walker equation is expressed in Euclidean norm as

$$\mathrm{RSS}_{i_0}(k_0)=\frac{1}{\bar{n}}\left\|\widehat{\mathbf{z}}_{i_0}-\widehat{\mathbf{V}}_{i_0}\widehat{\boldsymbol{\beta}}_{i_0}\right\|_2^2,\quad i_0=1,\cdots,\bar{n}. \tag{10}$$

The number of elements of whole coefficient matrices which we need to estimate is a function of the bandwidth k_0, we observe the specific expression of (5) imply that this is a under-determined case in the sense that the number of estimation equations is less than the number of estimation parameters, especially in the case of small sample size. Note that there are $\bar{n}+1$ parameters to be estimated with $\bar{n}$ equations in (4) when k_0 equals to $\frac{\bar{n}+2}{4}$ which is an extreme value and satisfies boundary condition ($<\bar{n}$) for $p=1$ and $i_0=k_0$, a similar under-determined situation can also be found in Section 2.2 of Dou et al. (2016) and Gao et al. (2019). In order to improve the estimation accuracy, we estimate parameters using the following r Yule-Walker equations:

$$\begin{bmatrix}\boldsymbol{\Sigma}_p^{\mathsf{T}}\\ \boldsymbol{\Sigma}_{p+1}^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{p+r-1}^{\mathsf{T}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\Sigma}_p^{\mathsf{T}}\\ \boldsymbol{\Sigma}_{p+1}^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{p+r-1}^{\mathsf{T}}\end{bmatrix}\boldsymbol{D}_0^{\mathsf{T}}+\begin{bmatrix}\boldsymbol{\Sigma}_{p-1}^{\mathsf{T}}\\ \boldsymbol{\Sigma}_p^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{p+r-2}^{\mathsf{T}}\end{bmatrix}\boldsymbol{D}_1^{\mathsf{T}}+\cdots+\begin{bmatrix}\boldsymbol{\Sigma}_0^{\mathsf{T}}\\ \boldsymbol{\Sigma}_1^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{r-1}^{\mathsf{T}}\end{bmatrix}\boldsymbol{D}_p^{\mathsf{T}}, \tag{11}$$

then the i_0th row of equation (11) implies

$$\mathbf{z}_{i_0}^{'}\equiv\begin{bmatrix}\boldsymbol{\Sigma}_p^{\mathsf{T}}\\ \boldsymbol{\Sigma}_{p+1}^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{p+r-1}^{\mathsf{T}}\end{bmatrix}\boldsymbol{e}_{i_0}=\sum_{j=0}^{p}\begin{bmatrix}\boldsymbol{\Sigma}_{p-j}^{\mathsf{T}}\\ \boldsymbol{\Sigma}_{p-j+1}^{\mathsf{T}}\\ \vdots\\ \boldsymbol{\Sigma}_{p-j+r-1}^{\mathsf{T}}\end{bmatrix}\boldsymbol{d}_{i_0}^{j}\equiv\mathbf{V}_{i_0}^{'}\boldsymbol{\beta}_{i_0}, \tag{12}$$

where the definition of $\mathbf{V}_{i_0}^{'}$ is similar to$\mathbf{V}_{i_0}$. For ease of calculation, we also replace $\boldsymbol{\Sigma}_j$ by the corresponding sample matrices, hence $\widehat{\mathbf{V}}_{i_0}^{'}$ could be written as

$$\widehat{\mathbf{V}}_{i_0}^{'}=\begin{bmatrix}\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}(\boldsymbol{r}_{(1),T})^{\mathsf{T}}\\ \frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}(\boldsymbol{r}_{(2),T})^{\mathsf{T}}\\ \vdots\\ \frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}(\boldsymbol{r}_{(r),T})^{\mathsf{T}}\end{bmatrix}.$$

where $(\boldsymbol{r}_{(k),T})^{\mathsf{T}} = ((\boldsymbol{r}_{(k),T_{p+k-1}})^{\mathsf{T}}, (\boldsymbol{r}_{(k),T_{p+k-2}})^{\mathsf{T}}, \cdots, (\boldsymbol{r}_{(k),T_{k-1}})^{\mathsf{T}})$, the sub-vector $\boldsymbol{r}_{(k),T_j}$ consists of R_{q,T_j} for $q \in S_{i_0}$ when $j = p+k-1$ and for $q \in S_{i_0}^{+}$ when $j \neq p+k-1$. Similarly, by the least squares method, we obtain the estimator

$$\widehat{\boldsymbol{\beta}}_{i_0}^{'} = \left((\widehat{\mathbf{V}}_{i_0}^{'})^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0}^{'} \right)^{-1} (\widehat{\mathbf{V}}_{i_0}^{'})^{\mathsf{T}} \widehat{\mathbf{z}}_{i_0}^{'}. \tag{13}$$

Denote a $\bar{n}r \times 1$ vector $\boldsymbol{f}_{\varepsilon_{i_0}} = (\frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p}^{\mathsf{T}} \varepsilon_{i_0,t}, \cdots, \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p}^{\mathsf{T}} \varepsilon_{i_0,t-r+1})^{\mathsf{T}}$, then it holds that for $i_0 = 1, \cdots, \bar{n}$,

$$\widehat{\boldsymbol{\beta}}_{i_0}^{'} - \boldsymbol{\beta}_{i_0} = \left((\widehat{\mathbf{V}}_{i_0}^{'})^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0}^{'} \right)^{-1} (\widehat{\mathbf{V}}_{i_0}^{'})^{\mathsf{T}} \boldsymbol{f}_{\varepsilon_{i_0}}.$$

The multiple generalized Yule-Walker estimator has the same properties as $n_0 \to \infty$ and $\bar{n} \to \infty$, and we usually take $r = 1$ for calculation convenience, we can increase r appropriately if there is lack of accuracy. But it is important that it performs better than generalized Yule-Walker estimator when there exists under-determined case.

2.3 A consistent estimator for high dimensional vector

When $\bar{n}/\sqrt{n_0} \to \infty$, the estimator (7) presents a convergence rate different from standard rate $1/\sqrt{n_0}$ by Theorem 3 in Section 3. For a high dimensional vector $\mathbf{z}_{i_0}$, we know the value of k_0 is finite, then the dimension of estimator is far less than $\bar{n}$, this is a traditional over-determined scenario. The probability of over-determined situation increases with the value of $r \times \bar{n}$, we have to decrease the numerical value $r \times \bar{n}$ for the objective accuracy of estimation parameters, that means we should omit some columns of (r) Yule-Walker equations. Borrowing the idea from Dou et al. (2016), we propose an alternative estimator to restore the standard $\sqrt{n_0}$-consistency and the asymptotic normality.

Since the lth row of $\widehat{\mathbf{V}}_{i_0}^{'}$ is $\boldsymbol{e}_l^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0}^{'}$, which is the sample covariance between $R_{l,t-p}$ and $\boldsymbol{r}_{(k),T}^{\mathsf{T}}$. Then define a non-negative parameter for $k = 1, \cdots, r$,

$$\delta_k^{(l)} = \left\| \boldsymbol{e}_l^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0}^{'} \right\|_1 = \left\| \frac{1}{n_0} \sum_{t=p+1}^{n_0} R_{l,t-p} (\boldsymbol{r}_{(k),T})^{\mathsf{T}} \right\|_1,$$

which can be rewritten as the lth row of the sum of all $|\widehat{\boldsymbol{\Sigma}}_j|$, the sum of $\delta_k^{(l)}$ presents the whole strength of correlation between $R_{l,t-p}$ and $\boldsymbol{r}_{(k),T_j}$ in this situation that more than one Yule-Walker equations. When $\delta^{(l)} = \sum_{k=1}^{r} \delta_k^{(l)}$ is close to 0, the lth row of all $|\widehat{\boldsymbol{\Sigma}}_j|$ is also close to 0, that implies parameters which we are estimating are meaningless, hence we may only keep the lth equation in (12) with the d_{i_0} largest $\delta^{(l)}$, and d_{i_0} is a prescribed number.

Let $\boldsymbol{m}_{(i_0),t-p}$ be the $d_{i_0} \times 1$ sub-vector of $\boldsymbol{R}_{t-p}$, and it consists of the corresponding d_{i_0} largest $\delta^{(l)}$, we replace $\boldsymbol{R}_{t-p}$ with the informative $\boldsymbol{m}_{(i_0),t-p}$ in the process of estimation here. Then the new estimator is defined as

$$\widetilde{\boldsymbol{\beta}}_{i_0}^{'} = (\widetilde{\mathbf{M}}_{i_0}^{\mathsf{T}} \widetilde{\mathbf{M}}_{i_0})^{-1} \widetilde{\mathbf{M}}_{i_0}^{\mathsf{T}} \widetilde{\mathbf{z}}_{i_0}, \quad i_0 = 1, \cdots, \bar{n},$$

where

$$\widetilde{\mathbf{M}}_{i_0} = \begin{bmatrix} \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}(\boldsymbol{r}_{(1),T})^{\mathsf{T}} \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}(\boldsymbol{r}_{(2),T})^{\mathsf{T}} \\ \vdots \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}(\boldsymbol{r}_{(r),T})^{\mathsf{T}} \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{z}}_{i_0} = \begin{bmatrix} \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}R_{i_0,t} \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}R_{i_0,t+1} \\ \vdots \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}R_{i_0,t+r-1} \end{bmatrix}. \quad (14)$$

Thus

$$\widetilde{\boldsymbol{\beta}}'_{i_0} - \boldsymbol{\beta}_{i_0} = (\widetilde{\mathbf{M}}_{i_0}^{\mathsf{T}}\widetilde{\mathbf{M}}_{i_0})^{-1}\widetilde{\mathbf{M}}_{i_0}^{\mathsf{T}}\widetilde{\boldsymbol{f}}_{\varepsilon_{i_0}},$$

where $\widetilde{\boldsymbol{f}}_{\varepsilon_{i_0}} = (\frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}^{\mathsf{T}}\varepsilon_{i_0,t}, \cdots, \frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{m}_{(i_0),t-p}^{\mathsf{T}}\varepsilon_{i_0,t-r+1})^{\mathsf{T}}$. Theorem 5 in Section 3 shows the asymptotic normality of the above estimator as long as the condition $d_{i_0} = o(\sqrt{n_0})$ holds uniformly for all i_0, and we have proved that the consistent estimator is effective in the section of simulation study. Compared with Gao et al. (2019), we adjusted the order of extracting sample information, and the over-determined and under-determined phenomena can be balanced in part by increasing the amount of information first and then restricting the estimation with convergence rate.

2.4 Determination of optimal bandwidth

The bandwidth is unknown and we need to estimate k_0 actually. We propose to determine the optimal bandwidth k_0 based on the marginal Bayesian information criterion,

$$BIC_{i_0}(k) = \log RSS_{i_0}(k) + \frac{1}{n_0}p\tau_{i_0}(k)C_{n_0}\log(\bar{n} \vee n_0), \qquad i_0 = 1, \cdots, \bar{n},$$

where $(\bar{n} \vee n_0) = \max(\bar{n}, n_0)$, and C_{n_0} is some positive constant which diverges with n_0, similar ideas can be found in Guo et al. (2016) and Lam and Yao (2012). We know the corresponding BIC is minimal for $k = k_0$ when all the sample covariance matrices $\widehat{\boldsymbol{\Sigma}}_j$ are replaced by the true $\boldsymbol{\Sigma}_j$. Then the hat value of optimal bandwidth k_0 is defined as

$$\widehat{k}_0 = \max_{1\le i_0\le \bar{n}} \arg\min_{1\le k\le K} BIC_{i_0}(k),$$

where $K \ge 1$ is a prescribed upper boundary of bandwidth, and our numerical study shows that it is insensitive to the choice of K provided that $K \ge k_0$. In practice, we often choose the upper boundary K to be $[\sqrt{\bar{n}}]$ which is integer. The ratio-based method was also used to determine optimal bandwidth in Gao et al. (2019), but the results of simulation show BIC method performs better in this model.

3 Theoretical properties

3.1 Regularity conditions

A strictly stationary of spatial process $R_{\boldsymbol{i},t'}$ is α-mixing at any fixed time $t^{'}$ if

$$\alpha(d(E_1,E_2)) = \sup_{A\in\mathscr{F}_{E_1},B\in\mathscr{F}_{E_2}} |P(A)P(B)-P(AB)| \leq \psi(|E_1|,|E_2|)\varphi(d(E_1,E_2)) \to 0,$$

as $d(E_1,E_2)\to\infty$, where $d(E_1,E_2)$ is the ordinary Euclidean distance, $\mathscr{F}_E$ denotes the σ-algebra generated by $\{R_{\boldsymbol{i},t'} : \boldsymbol{i}\in E\}$, and $\psi:\mathbb{N}^2\to\mathbb{R}^+$ is a symmetric positive function which is non-decreasing in each variable, the monotonically decreasing function $\varphi(x)\to 0$ as $x\to\infty$ and satisfies $\varphi(0)=1$.

Similarly, the strictly stationary of time process $R_{\boldsymbol{i}',t}$ is α-mixing at any fixed location $\boldsymbol{i}^{'}$ if

$$\alpha^{'}(d(E_1^{'},E_2^{'})) = \sup_{A\in\mathscr{F}_{E_1^{'}},B\in\mathscr{F}_{E_2^{'}}} |P(A)P(B)-P(AB)| \leq \psi^{'}(|E_1^{'}|,|E_2^{'}|)\varphi^{'}(d(E_1^{'},E_2^{'})) \to 0,$$

as $d(E_1^{'},E_2^{'})\to\infty$, where $\mathscr{F}_{E'}$ denotes the σ-algebra generated by $\{R_{\boldsymbol{i}',t} : t\in E^{'}\}$, $\psi^{'}$ and $\varphi^{'}$ are similar to ψ and φ above in setting of condition. Some regularity conditions are now in order.

A1. The random field $\{\boldsymbol{X}_{\boldsymbol{i},t}\}$ is strictly stationary. For all distinct $(\boldsymbol{i},t)$ and $(\boldsymbol{j},\tau)$ in $\Lambda_{\boldsymbol{n}}\times T_{\boldsymbol{n}}$, the $\{\boldsymbol{X}_{\boldsymbol{i},t}\}$ and $\{\boldsymbol{X}_{\boldsymbol{j},\tau}\}$ admit a joint density $f_{(\boldsymbol{i},t)(\boldsymbol{j},\tau)}$ and f denotes the marginal density of $\boldsymbol{X}_{\boldsymbol{i},t}$.

A2. The support of $\boldsymbol{K}(\cdot)$ is $[-c,c]^l$ as $c<\infty$ and $\boldsymbol{K}(\cdot)$ is a bounded function.

A3. $m(\boldsymbol{x})$ is twice differentiable, denoting by $m^{'}(\boldsymbol{x})$ and $m^{''}(\boldsymbol{x})$ its gradient and matrix of its second derivatives at $\boldsymbol{x}$ respectively, and $m^{'}(\boldsymbol{x})$ is continuous for $\boldsymbol{x}\in\mathbb{R}^l$.

A4. $\boldsymbol{h}\to 0$, and $\sqrt{n^* h_1\cdots h_l}\boldsymbol{h}_{\max}^2\to 0$ as $\boldsymbol{n}\to\infty$.

A5. (a) The innovations $\boldsymbol{\varepsilon}_t$ are independent and identically distributed satisfying $E(\varepsilon_{\boldsymbol{i},t})=0$ and $\mathrm{Cov}(\boldsymbol{R}_t,\boldsymbol{\varepsilon}_t)=0$ for $\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}$, $t\in T_{\boldsymbol{n}}$.

(b) The process $\boldsymbol{R}_t$ in models (1) and (2) is α-mixing satisfying $\sum_{k=1}^{\infty}\alpha(k)^{\frac{\gamma}{4+\gamma}}<\infty$ for some constant $\gamma>0$ and $E(R_{\boldsymbol{i},t})=0$ for $\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}$, $t\in T_{\boldsymbol{n}}$.

(c) For $\gamma>0$ in (b) above,

$$\sup_{1\leq j\leq\bar{n}} E\left|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{\Sigma}_k\boldsymbol{R}_{t-k}\right|^{4+\gamma}<\infty,\quad \sup_{1\leq j\leq\bar{n}} E\left|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-k}\right|^{4+\gamma}<\infty,\quad \sup_{1\leq i_0\leq\bar{n}} E\left|\varepsilon_{i_0,t}\right|^{4+\gamma}<\infty$$

for $k\in\{0,\cdots,p\}$.

A6. The matrix $\mathbf{I}-\boldsymbol{D}_0$ is invertible. And for each $j=(1,\cdots,p)$, $(i)\left\|(\mathbf{I}-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_j\right\|_2<1$, $(ii)d_{i_0,j_0}^{j}$ and the diagonal elements of $\mathbf{K}_{i_0},\mathbf{P}_{i_0}\mathbf{P}_{i_0}^{\mathsf{T}}$ and $\widetilde{\mathbf{P}}_{i_0}\widetilde{\mathbf{P}}_{i_0}^{\mathsf{T}}$ are all bounded uniformly, where $\mathbf{K}_{i_0},\mathbf{P}_{i_0}$ and $\widetilde{\mathbf{P}}_{i_0}$ are specified in (15), (16) and (17), respectively. $(iii)|d_{i_0,i_0-k_0}^{j}|$ or $|d_{i_0,i_0+k_0}^{j}|$ is greater than $\sqrt{C_{n_0}k_0n_0^{-1}\log(\bar{n}\vee n_0)}$, where $(\bar{n}C_{n_0}\log(n_0))/n_0\to 0$ and $(C_{n_0}^2\log n_0)/n_0\to\infty$ as $n_0\to\infty$.

A7. For any finite number of columns of $\mathbf{V}_{i_0}$, denoted by $\mathbf{F}_{i_0}$ and $\mathbf{H}_{i_0}$ in matrix form and $\mathbf{F}_{i_0}\neq\mathbf{H}_{i_0}$, $\lambda_1\leq\lambda_{\min}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_{i_0}(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\}\leq\lambda_{\max}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_{i_0}(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\}<\lambda_2$ for some positive constants $\lambda_1<\lambda_2$, and the rank of $\mathbf{V}_{i_0}$ is equal to τ_{i_0}.

Remark 1. Condition A1 is standard in this context; it has been used, for instance, by Masry (1986) in the serial case, and by Tran (1990) in the spatial context. Conditions A2-A4 are proposed in Wang et al. (2012) as standard for nonparametric estimation process, Condition A2 is just for the minor convenient process of proof, it can be extended to the infinite support set in practice as is the good case in Wang et al. (2012), and Condition A4 is required such that the bias goes to zero faster than the standard error, and thus is a suboptimal nonparametric bandwidth choice. Condition A5 limits the dependence across different spatial locations. It is implied by, for example, the conditions imposed in Yu et al. (2012). Condition A6 ensures that the bandwidth k_0 is asymptotically identifiable, as $\sqrt{n_0^{-1}\log(\bar{n}\vee n_0)}$ is the minimum order of a non-zero coefficient to be identifiable, see, e.g., Guo et al. (2016). Similar to the Condition A4 in Gao et al.(2019), Condition A7 guarantees that the boundaries of $RSS_{i_0}(k)$ are identifiable, and the rank is equal to the number of parameters we need to estimate, then ensure the validity of estimation with generalized Yule-Walker equations.

3.2 Asymptotic properties

We first state the asymptotic properties of hat value of the error term $\hat{R}_{x_0,t}$.

Theorem 1. Let Conditions A1-A5 hold, then

$$(i)E(\widehat{R}_{\boldsymbol{i},t}) = 0$$
$$(ii)\sqrt{n^* h_1\cdots h_l}(R_{\boldsymbol{i},t} - \widehat{R}_{\boldsymbol{i},t}) \to_d N(\mathbf{0}, \sigma^2 \boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}^{-1}\boldsymbol{\Sigma}(\mathbf{U}^{-1})^{\mathsf{T}}\boldsymbol{e}_1),$$

where

$$\mathbf{U} = \begin{bmatrix} f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} & f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} \\ f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} & f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} \end{bmatrix},$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{K}^2(\boldsymbol{z})d\boldsymbol{z} & f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{K}^2(\boldsymbol{z})d\boldsymbol{z} \\ f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}\boldsymbol{K}^2(\boldsymbol{z})d\boldsymbol{z} & f(\boldsymbol{x}_0)\int_{\mathbb{R}^l}\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{K}^2(\boldsymbol{z})d\boldsymbol{z} \end{bmatrix}.$$

Remark 2. Theorem 1 indicates that $\widehat{R}_{\boldsymbol{i},t}$ has the same expectation as $R_{\boldsymbol{i},t}$, and the nonparametric estimator of unobserved $R_{\boldsymbol{i},t}$ also performs well, then the analyses of simulation and case study are fairly reliable.

Theorem 2. Let conditions A1-A7 hold and $\bar{n} = o(n_0)$, then $P(\widehat{k}_0 = k_0) \to 1$ as $n_0 \to \infty$.

For $i_0 = 1, \cdots, \bar{n}$, let

$$\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(j) = \mathrm{Cov}(\boldsymbol{R}_{t-p+j}\varepsilon_{i_0,t+j}, \boldsymbol{R}_{t-p}\varepsilon_{i_0,t}), \quad j = 0,1,2,\cdots,$$
$$\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}} = \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(0) + \sum_{j=1}^{\infty}[\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(j) + \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}^{\mathsf{T}}(j)].$$

Meanwhile, define

$$\mathbf{U}_{i_0} \equiv \begin{bmatrix} \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_p^{\mathsf{T}}\mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_{p-1}^{\mathsf{T}}\mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_0\mathbf{I}_{S_{i_0}^+} \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_{p-1}\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_p^{\mathsf{T}}\mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_{p-1}\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_x}\boldsymbol{\Sigma}_{p-1}^{\mathsf{T}}\mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_{p-1}\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_0\mathbf{I}_{S_{i_0}^+} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_p^{\mathsf{T}}\mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_{p-1}^{\mathsf{T}}\mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}\boldsymbol{\Sigma}_0\mathbf{I}_{S_{i_0}^+} \end{bmatrix}$$

and

$$\mathbf{K}_{i_0} \equiv \begin{bmatrix} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_{p-1}^{\mathsf{T}} \mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_0 \mathbf{I}_{S_{i_0}^+} \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{p-1} \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{p-1} \boldsymbol{\Sigma}_{p-1}^{\mathsf{T}} \mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{p-1} \boldsymbol{\Sigma}_0 \mathbf{I}_{S_{i_0}^+} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_{p-1}^{\mathsf{T}} \mathbf{I}_{S_{i_0}^+} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_0 \mathbf{I}_{S_{i_0}^+} \end{bmatrix}, \tag{15}$$

where $\mathbf{I}_{S_{i_0}}$ is the $\bar{n} \times |S_{i_0}|$ matrix obtained by the stacking $\boldsymbol{e}_j$ for all $j \in S_{i_0}$ parallel, in other words, $\mathbf{I}_{S_{i_0}} = (\boldsymbol{e}_j, j \in S_{i_0})$. Similarly, $\mathbf{I}_{S_{i_0}^+}$ is consisting of $\boldsymbol{e}_j$ for all $j \in S_{i_0}^+$.

Theorem 3. Let conditions A1-A7 hold.

(*i*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = o(\sqrt{n_0})$. If k_0 is fixed, then

$$\sqrt{n_0}\mathbf{U}_{i_0}^{-\frac{1}{2}} \mathbf{K}_{i_0}(\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}) \to_d N(\mathbf{0}, \mathbf{I}_{\tau_{i_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 = O_p(\sqrt{\frac{k_0}{n_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

(*ii*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = O(\sqrt{n_0})$, and$\bar{n} = o(n_0)$. If k_0 is fixed, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 = O_p(\frac{\bar{n}}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 = O_p(\frac{\bar{n}}{n_0}\sqrt{k_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

Remark 3. Theorem 3 indicates that the standard convergence rate prevails as long as $\bar{n} = o(n_0)$. However the convergence rate may be slower when $\bar{n}$ is of higher orders than $\sqrt{n_0}$.

To present the convergence rate for the estimation errors, Theorem 4 consider the L_1 norm of estimation errors.

Theorem 4. Let conditions A1-A7 hold.

(*i*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = O(\sqrt{n_0})$. If k_0 is fixed, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_1 = O_p(\frac{\bar{n}}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_1 = O_p(\frac{\bar{n}k_0}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

(*ii*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = o(\sqrt{n_0})$. If k_0 is fixed, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_1 = O_p(\frac{1}{\sqrt{n_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_1 = O_p(\frac{k_0}{\sqrt{n_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

(*iii*) As $n_0 \to \infty, \bar{n} \to \infty$, $\frac{\bar{n}}{\sqrt{n_0}} \to \infty$, and $\bar{n} = o(n_0)$. If k_0 is fixed, then

$$\left\| \widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0} \right\|_1 = O_p(\frac{\bar{n}}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\| \widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0} \right\|_1 = O_p(\frac{\bar{n} k_0}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

Theorem 4 indicates that the L_1-norm of estimation errors tend to 0 when those conditions above hold, and the convergence rates for L_1-norm of estimation errors keep different correlation with $\bar{n}$, n_0 and k_0 under different setting conditions.

To derive the asymptotic properties of estimator $\widehat{\boldsymbol{\beta}}_{i_0}^{'}$, define the element

$$\begin{aligned} \mathbf{Q}_{i_0}(j_1, j_2) &= \mathrm{Cov}(\boldsymbol{R}_{t-p}\varepsilon_{i_0,t+1-j_1}, \boldsymbol{R}_{t-p}\varepsilon_{i_0,t+1-j_2}) \\ &+ \sum_{j=1}^{\infty} [\mathrm{Cov}(\boldsymbol{R}_{t-p+j}\varepsilon_{i_0,t+1+j-j_1}, \boldsymbol{R}_{t-p}\varepsilon_{i_0,t+1-j_2}) \\ &+ \mathrm{Cov}(\boldsymbol{R}_{t-p}\varepsilon_{i_0,t+1-j_1}, \boldsymbol{R}_{t-p+j}\varepsilon_{i_0,t+1+j-j_2})] \end{aligned}$$

of $r \times r$ matrix $\mathbf{Q}_{i_0}$ for $j_1, j_2 \in \{1, \cdots, r\}$. Furthermore,

$$\mathbf{P}_{i_0} = \begin{bmatrix} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p & \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_{p+1} & \cdots & \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_{p+r-1} \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{p-1} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_p & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{p+r-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_0 & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_1 & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_{r-1} \end{bmatrix}. \tag{16}$$

In addition, the Condition A7 need to be adapted under the multiple Yule-Walker equations, then Condition A8 was built to establish Theorem 5.

A8. For any finite number of columns of $\mathbf{V}_{i_0}^{'}$, denoted by $\mathbf{F}_{i_0}$ and $\mathbf{H}_{i_0}$ in matrix form and $\mathbf{F}_{i_0} \neq \mathbf{H}_{i_0}$, $\lambda_1 \leq \lambda_{\min}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I} - \mathbf{H}_{i_0}(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\} \leq \lambda_{\max}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I} - \mathbf{H}_{i_0}(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\} < \lambda_2$ for some positive constants $\lambda_1 < \lambda_2$, and the rank of $\mathbf{V}_{i_0}^{'}$ is equal to τ_{i_0}.

Theorem 5. Let conditions A1-A7 hold.

(*i*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = o(\sqrt{n_0})$. If k_0 is fixed, then

$$\sqrt{n_0}(\mathbf{P}_{i_0}\mathbf{Q}_{i_0}\mathbf{P}_{i_0}^{\mathsf{T}})^{-\frac{1}{2}}\mathbf{P}_{i_0}\mathbf{P}_{i_0}^{\mathsf{T}}(\widehat{\boldsymbol{\beta}}_{i_0}^{'} - \boldsymbol{\beta}_{i_0}) \to_d N(\mathbf{0}, \mathbf{I}_{\tau_{i_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\| \widehat{\boldsymbol{\beta}}_{i_0}^{'} - \boldsymbol{\beta}_{i_0} \right\|_2 = O_p(\sqrt{\frac{k_0}{n_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

(*ii*) As $n_0 \to \infty, \bar{n} \to \infty$, and $\bar{n} = O(\sqrt{n_0})$, and$\bar{n} = o(n_0)$. If k_0 is fixed, then

$$\left\| \widehat{\boldsymbol{\beta}}_{i_0}^{'} - \boldsymbol{\beta}_{i_0} \right\|_2 = O_p(\frac{\bar{n}}{n_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, then

$$\left\| \widehat{\boldsymbol{\beta}}_{i_0}^{'} - \boldsymbol{\beta}_{i_0} \right\|_2 = O_p(\frac{\bar{n}}{n_0}\sqrt{k_0}), \quad i_0 = 1, \cdots, \bar{n}.$$

To derive the asymptotic properties of estimator $\widetilde{\boldsymbol{\beta}}_{i_0}$, redefine the (auto)covariance matrices,

$$\boldsymbol{\Sigma}'_{(i_0),j} = \mathrm{Cov}(\boldsymbol{R}_t, \boldsymbol{m}_{(i_0),t-j}), \qquad \widehat{\boldsymbol{\Sigma}}'_{(i_0),j} = \mathrm{Cov}(\boldsymbol{R}_{t-p+j}, \boldsymbol{m}_{(i_0),t-p}).$$

By some similar notations as that of Theorem 5, define the element

$$\begin{aligned}\widetilde{\mathbf{Q}}_{i_0}(j_1, j_2) &= \mathrm{Cov}(\boldsymbol{m}_{(i_0),t-p}\varepsilon_{i_0,t+1-j_1}, \boldsymbol{m}_{(i_0),t-p}\varepsilon_{i_0,t+1-j_2}) \\ &+ \sum_{j=1}^{\infty}\big[\mathrm{Cov}(\boldsymbol{m}_{(i_0),t-p+j}\varepsilon_{i_0,t+1+j-j_1}, \boldsymbol{m}_{(i_0),t-p}\varepsilon_{i_0,t+1-j_2}) \\ &+ \mathrm{Cov}(\boldsymbol{m}_{(i_0),t-p}\varepsilon_{i_0,t+1-j_1}, \boldsymbol{m}_{(i_0),t-p+j}\varepsilon_{i_0,t+1+j-j_2})\big]\end{aligned}$$

of $r \times r$ matrix $\widetilde{\mathbf{Q}}_{i_0}$ for $j_1, j_2 \in \{1, \cdots, r\}$, and

$$\widetilde{\mathbf{P}}_{i_0} = \begin{bmatrix} \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p} & \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p+1} & \cdots & \mathbf{I}_{S_{i_0}}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p+r-1} \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p-1} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),p+r-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),0} & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),1} & \cdots & \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}}\boldsymbol{\Sigma}'_{(i_0),r-1} \end{bmatrix}. \tag{17}$$

Similarly, the Condition A5(c) and A7 just do not suit the new estimator, then we adapt them for Condition A8 in order to establish Theorem 6.

A9. For $\gamma > 0$ specified in A5(b),

$$\sup_{1\le j\le \bar{n}} E\left|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{\Sigma}'_k\boldsymbol{m}_{t-k}\right|^{4+\gamma} < \infty, \quad \sup_{1\le j\le \bar{n}} E\left|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-k}\right|^{4+\gamma} < \infty, \quad \sup_{1\le i_0\le \bar{n}} E\left|\varepsilon_{i_0,t}\right|^{4+\gamma} < \infty$$

for $k = (0, \cdots, p)$ and $i_0 = (1, \cdots, \bar{n})$.

A10. For any finite number of columns of $\widetilde{\mathbf{M}}_{i_0}$, denoted by $\mathbf{F}_{i_0}$ and $\mathbf{H}_{i_0}$ in matrix form and $\mathbf{F}_{i_0} \neq \mathbf{H}_{i_0}$, $\lambda_1 \le \lambda_{\min}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_{i_0}(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\} \le \lambda_{\max}\{\mathbf{F}_{i_0}^{\mathsf{T}}(\mathbf{I}-\mathbf{H}_x(\mathbf{H}_{i_0}^{\mathsf{T}}\mathbf{H}_{i_0})^{-1}\mathbf{H}_{i_0}^{\mathsf{T}})\mathbf{F}_{i_0}\} < \lambda_2$ for some positive constants $\lambda_1 < \lambda_2$, and the rank of $\widetilde{\mathbf{M}}_{i_0}$ is equal to τ_{i_0}.

Theorem 6. Let conditions A1-A5(a,b), A6 and A9-A10 hold. As $n_0 \to \infty, \bar{n} \to \infty$, and $d_{i_0} = o(\sqrt{n_0})$. If k_0 is fixed, then

$$\sqrt{n_0}(\widetilde{\mathbf{P}}_{i_0}\widetilde{\mathbf{Q}}_{i_0}\widetilde{\mathbf{P}}_{i_0}^{\mathsf{T}})^{-\frac{1}{2}}\widetilde{\mathbf{P}}_{i_0}\widetilde{\mathbf{P}}_{i_0}^{\mathsf{T}}(\widetilde{\boldsymbol{\beta}}'_{i_0} - \boldsymbol{\beta}_{i_0}) \to_d N(\mathbf{0}, \mathbf{I}_{\tau_{i_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

If $k_0 = o\left(C_{n_0}^{-1}n_0/\log(\bar{n} \vee n_0)\right)$, then

$$\left\|\widetilde{\boldsymbol{\beta}}'_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 = O_p(\sqrt{\frac{k_0}{n_0}}), \quad i_0 = 1, \cdots, \bar{n}.$$

Theorem 6 indicates that the estimator $\widetilde{\boldsymbol{\beta}}'_{i_0}$ are asymptotic normal with the standard rate as long as $d_{i_0} = o(\sqrt{n_0})$. If the positive integer r equals to 1, we can also achieve the standard convergence rate in Theorem 3.

4 Simulation study

In this section, we conduct two simulations as follows to evaluate the finite-sample properties of the proposed methods. For the sake of simplicity, we simulate $Y_{\boldsymbol{i},t}$ from the model

$$\begin{aligned} Y_{i,j,t} &= m(\boldsymbol{X}_{i,j,t}) + R_{i,j,t}, \\ \boldsymbol{R}_t &= \boldsymbol{D}_0\boldsymbol{R}_t + \boldsymbol{D}_1\boldsymbol{R}_{t-1} + \boldsymbol{\varepsilon}_t, \end{aligned}$$

with $d = 2, l = p = 1$, where innovations are *iid* from a standard normal distribution. To simulate the spatio-temporal process $\nu_{i,j,t}$, we follow the spectral method of Cressie (1993) that

$$\nu_{i,j,t} = \left(\frac{2}{M}\right)^{\frac{1}{2}} \sum_{k=1}^{M} \cos(i \cdot w(1,k) + j \cdot w(2,k) + t \cdot q(k) + r(k)), \qquad k = 1, \cdots, M,$$

where $w(i,k), i = 1, 2$, and $q(k)$ are iid from a standard normal distribution, independent of $r(k)$, which are iid uniform random variables on $[-\pi, \pi]$. We apply the data central processing method to $R_{i,j,t}$ and set the bandwidth of nonparametric estimation $h_1 = n^{*-\frac{1}{5}} s(X)$ where $s(X)$ is the sample standard deviation of X. For each nonparametric estimation process, we choose Epanechnikov kernel function $\boldsymbol{K}(t) = 0.75(1-t^2)\mathrm{I}(|t| \leq 1)$. Then we consider two settings for $m(\cdot)$ and coefficient matrices $\boldsymbol{D}_0 \equiv (d^0_{i_0,j_0})$ and $\boldsymbol{D}_1 \equiv (d^1_{i_0,j_0})$.

Scenario 1. $m(t)$ is $\sin(t)$ for spatio-temporal process, and $\{d^0_{i_0,j_0}, d^1_{i_0,j_0} : |i_0 - j_0| \leq k_0\}$ are generated independently from $U([-2.5,-1] \bigcup [1, 2.5])$, then rescale $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ to $\eta_0 \cdot \boldsymbol{D}_0 / \|\boldsymbol{D}_0\|_2$ and $\eta_1 \cdot \boldsymbol{D}_1 / \|\boldsymbol{D}_1\|_2$, where η_0 and η_1 are drawn independently from $U(0.4, 1)$.

Scenario 2. $m(t)$ is $(t+5)^{-\frac{1}{2}}$ for spatio-temporal process, and $\{d^0_{i_0,j_0}, d^1_{i_0,j_0} : |i_0 - j_0| = k_0\}$ are generated independently from $Bernoulli(0.5)$ on two points $\{-1.5, 1.5\}$, $\{d^0_{i_0,j_0}, d^1_{i_0,j_0} : |i_0 - j_0| < k_0\}$ are drawn independently from mixture distribution $\xi \cdot \chi^2(3) + (1-\xi) \cdot N(0,1)$ with $P(\xi = 1) = 0.4 = 1 - P(\xi = 0)$, then rescale $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ as in Scenario 1 above.

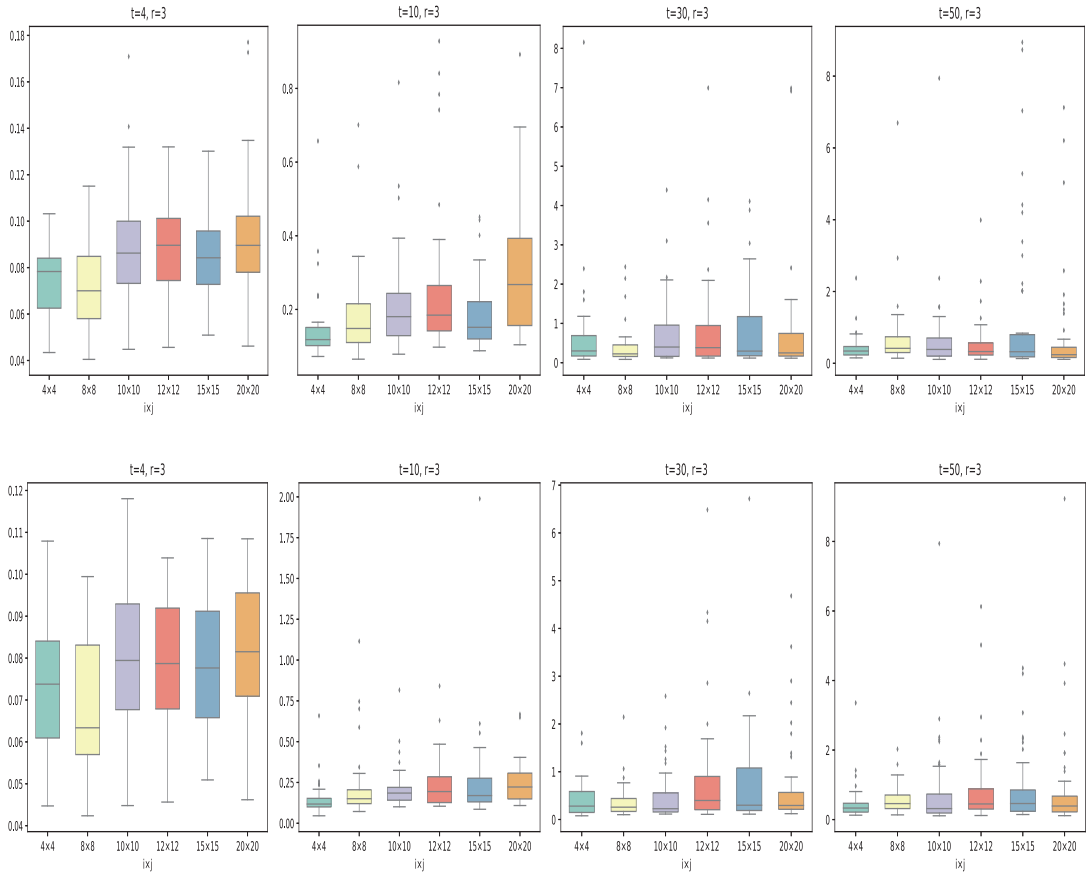

Figure 2: Boxplots of $\|\boldsymbol{D}_0 - \widehat{\boldsymbol{D}}_0\|_2$ for scenario 1.

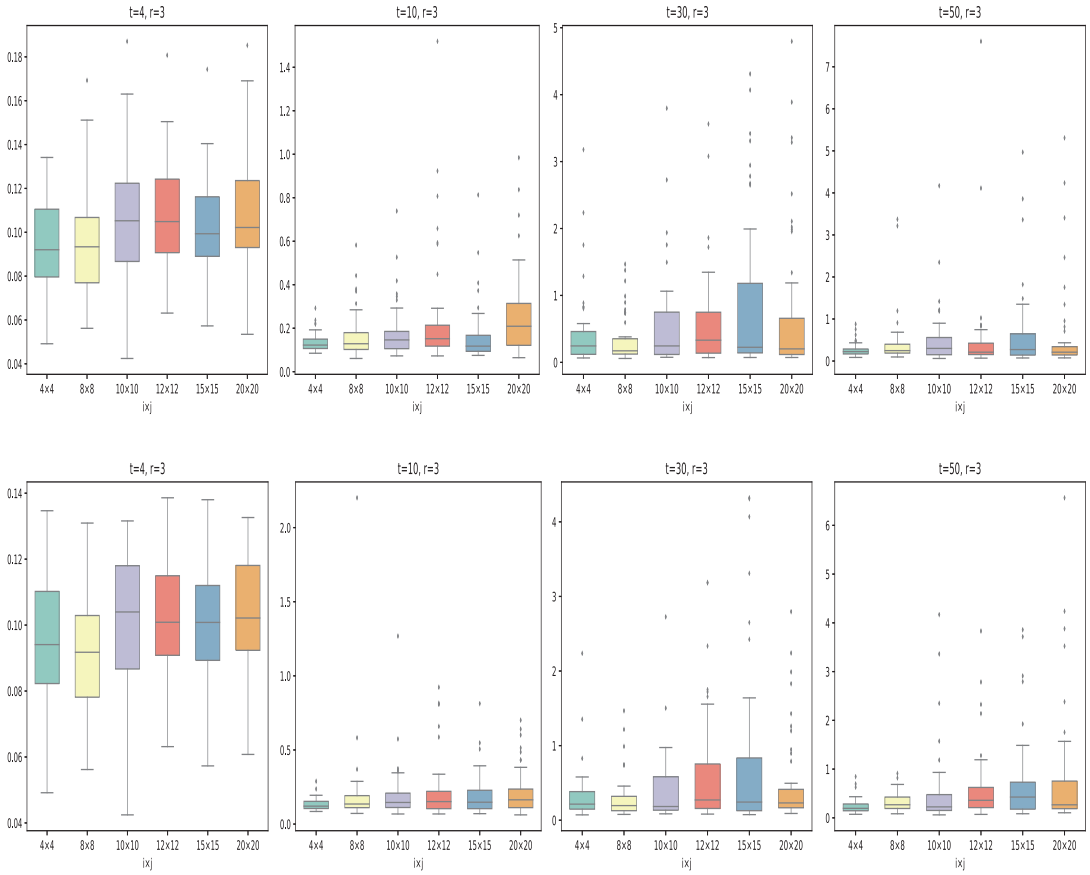

Figure 3: Boxplots of $\|\boldsymbol{D}_1 - \widehat{\boldsymbol{D}}_1\|_2$ for scenario 1.

For each setting, we replicate the experiment 100 times with $M = 1000$ and set sample pize with $\{i \times j \times t : i = j = 4, 8, 10, 12, 15, 20; t = 4, 10, 30, 50\}$, this leads to the 24 different estimator(i, j, t) combinations. In addition, we choose $d_x = \min(\bar{n}, n_0^{0.467})$ after many numerical simulation experiments and $k_0 = 3, K = 7$. Figure 2 and Figure 3 depict some boxplots of estimation errors $\|\boldsymbol{D}_0 - \widehat{\boldsymbol{D}}_0\|_2$ and $\|\boldsymbol{D}_1 - \widehat{\boldsymbol{D}}_1\|_2$ respectively for estimator(14) with $r = 1, 3$ under the setting of scenario 1. As indicated clearly in Figures 2 and 3, when the time size is small, the errors in estimating the coefficient matrices fluctuate with the location size; when the time size increases to a certain extent, the errors tend to be stable and ignore the change of location size. Those results above signify that the estimator (14) can balance over-determined and under-determined in part. Note that the errors in estimating the coefficient matrices based on $r = 3$ don't perform worse than the errors based on $r = 1$, this shows that multiple generalized Yule-Walker equations can provide more information for estimation process. Moreover, the accuracy of $\hat{k}$ increases with the sample size can be found in Table 1, and a dominant proportion of $\{\hat{k} = k_0\}$ and $\{\hat{k} > k_0\}$ usually produces more stable estimation errors. The ultimate goal of our model is prediction, then the most important index is relative frequencies of occurrence of the events $\{\hat{k} \geq k_0\}$, and the results of ratio-based method (Gao et al., 2019) are also reported in parentheses. Both results of two methods tend to 1 with the increase of sample size, and the probability value is gradually similar. On the other hand, the probability of occurrence of the

estimated optimal bandwidth equal to the real bandwidth under BIC method is significantly higher, which is also an important reason why the marginal Bayesian information criterion is finally used to determine the optimal bandwidth in this paper.

Table 1: Relative frequencies of occurrence (%) of the events $\{\hat{k} = k_0\}$ and $\{\hat{k} > k_0\}$ based on $r = 1$ by BIC method and ratio-based method (in parentheses).

		Scenario 1			Scenario 2		
$i \times j$	t	$\{\hat{k} = k_0\}$	$\{\hat{k} > k_0\}$	$\{\hat{k} \geq k_0\}$	$\{\hat{k} = k_0\}$	$\{\hat{k} > k_0\}$	$\{\hat{k} \geq k_0\}$
	4	8 (6)	14 (12)	22 (18)	12 (7)	2 (4)	14 (11)
4×4	10	17 (13)	5 (8)	22 (21)	21 (13)	2 (7)	23 (20)
	30	22 (14)	4 (13)	26 (27)	26 (17)	6 (10)	32 (27)
	50	30 (18)	8 (18)	38 (36)	28 (21)	10 (14)	38 (35)
	4	21 (17)	4 (7)	25 (24)	13 (13)	16 (15)	29 (28)
8×8	10	27 (16)	8 (16)	35 (32)	35 (19)	8 (15)	43 (34)
	30	32 (24)	14 (24)	46 (48)	38 (25)	10 (26)	48 (51)
	50	38 (26)	24 (27)	62 (53)	45 (34)	23 (31)	68 (65)
	4	25 (20)	11 (16)	36 (36)	27 (21)	16 (18)	43 (39)
10×10	10	30 (21)	18 (25)	48 (46)	35 (28)	21 (31)	56 (59)
	30	39 (26)	17 (29)	56 (55)	46 (31)	25 (37)	71 (68)
	50	43 (32)	32 (34)	75 (66)	46 (38)	42 (36)	88 (74)
	4	31 (24)	19 (26)	50 (50)	34 (23)	19 (30)	53 (53)
12×12	10	37 (24)	29 (41)	66 (65)	38 (25)	24 (35)	62 (60)
	30	36 (30)	36 (41)	72 (71)	40 (28)	31 (42)	71 (70)
	50	48 (34)	36 (46)	84 (80)	50 (33)	33 (46)	83 (79)
	4	41 (26)	29 (42)	70 (68)	37 (26)	22 (34)	59 (60)
15×15	10	47 (28)	34 (43)	81 (71)	40 (24)	22 (37)	62 (61)
	30	53 (33)	27 (48)	80 (81)	44 (30)	37 (49)	81 (79)
	50	54 (39)	37 (51)	91 (90)	49 (34)	35 (50)	84 (84)
	4	45 (30)	34 (48)	79 (78)	40 (29)	27 (37)	67 (68)
20×20	10	49 (29)	36 (50)	85 (79)	44 (33)	39 (48)	83 (81)
	30	58 (33)	36 (58)	94 (91)	57 (37)	39 (58)	96 (95)
	50	60 (42)	39 (57)	99 (99)	57 (41)	43 (57)	100 (98)

To show the estimation errors for scenario 2, we omit the worse results based on $r = 1$ to save space, and Figure 4 indicate a similar estimation results to scenario 1, so does Table 1. Overall, the estimator (14) is stable for the sample size.

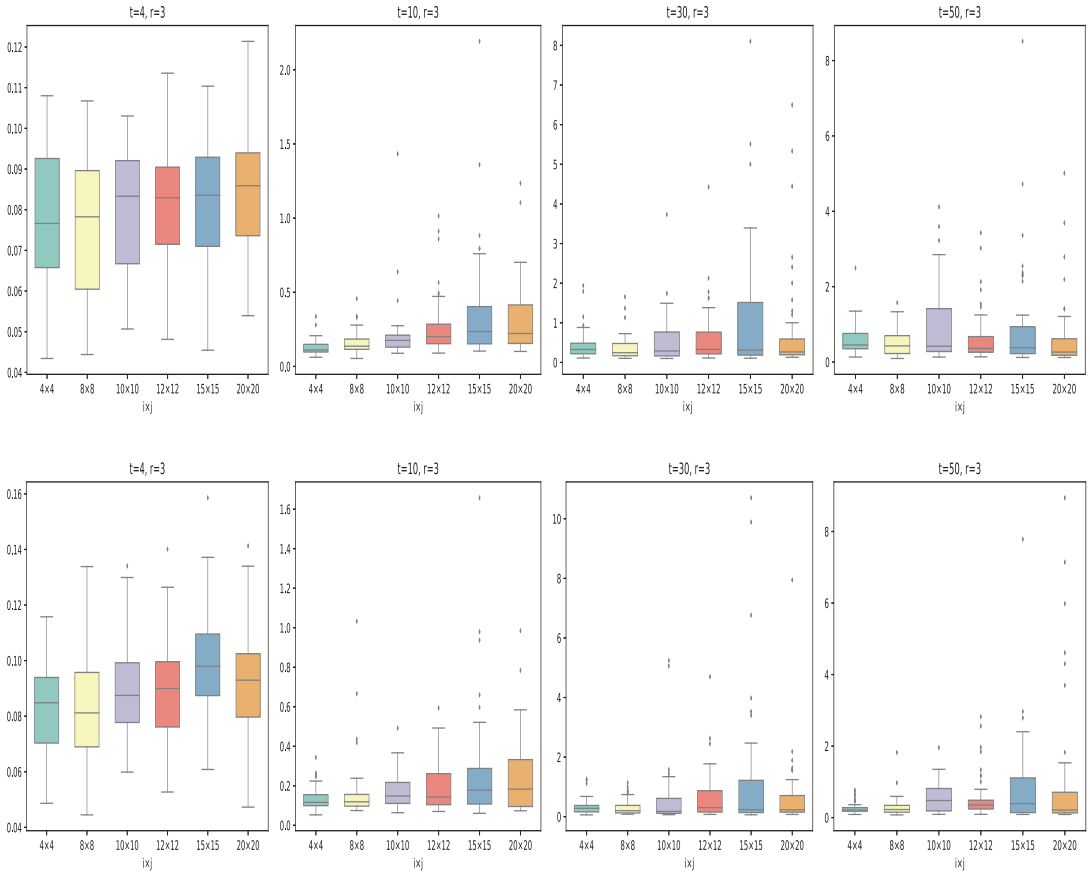

Figure 4: Boxplots of $\|\boldsymbol{D}_0 - \widehat{\boldsymbol{D}}_0\|_2$ (the top panels) and $\|\boldsymbol{D}_1 - \widehat{\boldsymbol{D}}_1\|_2$ (the bottom panels) for scenario 2.

5 Real data analysis

We now illustrate the proposed model via an application to two spatio-temporal data sets in this section, and the dimension of independent variable is small, then we don't need to solve the curse of dimensionality. The settings not mentioned below are same as section 4.

Case 1. We analyze here the monthly air quality index (AQI) for Beijing-Tianjin-Hebei urban agglomeration of China in the period of Jan 2014 – Nov 2019, and $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3, temperature, speed of the winds, rainfall of the corresponding period as independent variables. Data can be availabled from https://www.aqistudy.cn/historydata/ and http://hj.zc12369.com/. For this data set, location size $\bar{n} = 14$, time size $n_0 = 71$. Figure 5 presents the time series plots of the monthly estimated error terms at three cities Beijing, Chinwangtao and Kalgan. To fit the banded model with $p = 1, K = 7$ and $r = 3$, we might need to arrange the 14 cities in a certain order and here consider the operating frequency of trains to Beijing as the ordering.

According to the Air Quality Standard in China, the AQI is marked at 6 different levels, the higher AQI indicates the worse air quality, and Figure 1 depicts the AQI level for consecutive months. For general public the prediction for a specific AQI value is more significant than that

for the level, the estimated bandwidth $\hat{k}$ and the mean squared predictive errors of one-step ahead and two-step ahead predictions based on the spatio-temporal error models with known spatial weight matrix (STEM) and spatio-temporal models with autoregressive errors(STBEM) are reported in Table 2, and the known spatial matrix was set to spatial contiguity matrix. It is easy to see from Table 2 that the STBEM has the greater prediction accuracy than STEM.

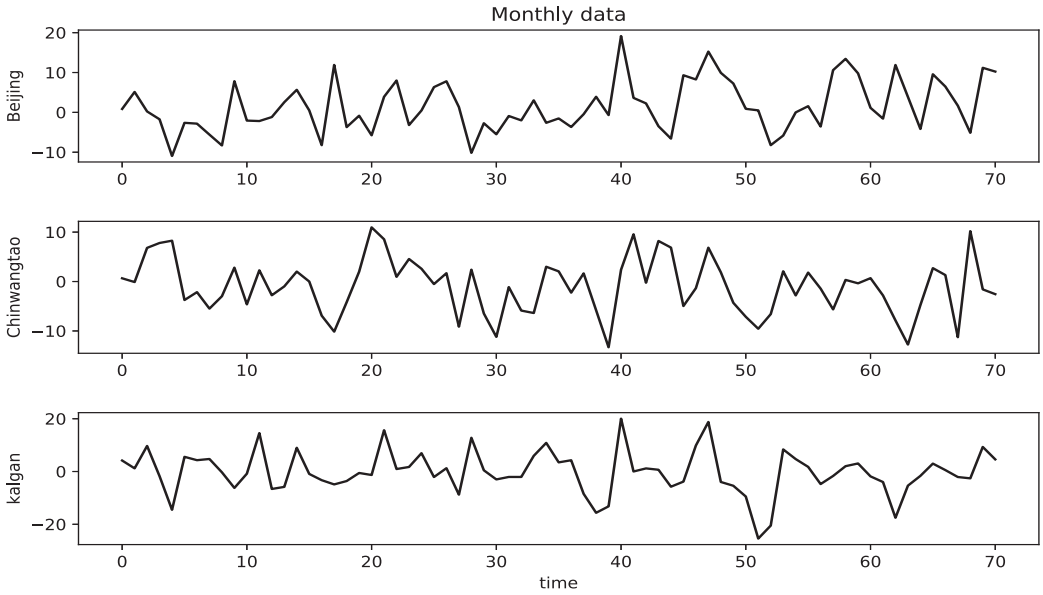

Figure 5: Time series plots of monthly estimated error terms at, from top to bottom, Beijing, Chinwangtao and Kalgan.

Case 2. Now we consider the yearly effect of public capital, private capital, employment and the rate of unemployment on gross state product of 48 states (except Hawaii and Alaska) in the United States over the years 1970-1984. Data can be availabled from https://www.ceicdata.com/zh-hans. Note that now location size $\bar{n} = 48$ and time size $n_0 = 15$. With the logarithmic transformation to the data of variables (except the rate of unemployment), Figure 6 shows the time series plots of the estimated error terms from those transformed data at six randomly selected states. In order to fit the banded model with $p = 1, K = 15, r = 3$, and detect the importance of location order to prediction, we just arrange them in alphabetical order. Moreover, we choose the spatial contiguity matrix in this case to predict based on the STEM. The mean squared predictive errors of one-step and two-step ahead for two different methods are both reported in three different directions in Table 2. According that, STBEM has the better one-step ahead predictions clearly in three different directions, and the two-step ahead predictions are not worse than that of STEM.

6 Concluding remark

We propose in this paper a new class of semiparametric banded spatio-temporal models with autoregressive errors. No matter what form of the implied auto-covariance matrices, the

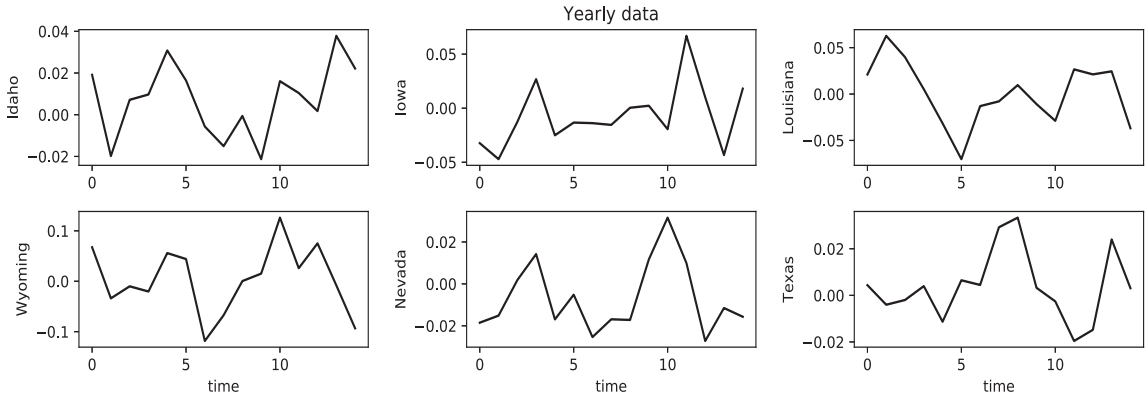

Figure 6: Time series plots of yearly estimated error terms at six randomly selected states in the United States.

Table 2: The estimated bandwidth and mean squared predictive errors for case 1 and case 2.

		Case 1		Case 2	
		STBEM	STEM	STBEM	STEM
No particular order	One-step ahead	2.9715	3.1290	0.0026	0.0034
	Two-step ahead	5.4154	6.4205	0.0054	0.0054
	$\hat{k}$	7		13	
Northeast to southwest	One-step ahead	1.4954	3.8940	0.0025	0.0035
	Two-step ahead	2.4314	5.3306	0.0053	0.0056
	$\hat{k}$	7		13	
Southeast to northwest	One-step ahead	1.1075	1.5415	0.0023	0.0035
	Two-step ahead	4.2776	4.5084	0.0048	0.0056
	$\hat{k}$	7		15	

setting can include as many panels as possible. The coefficient matrices are estimated based on generalized Yule – Walker equations, and the optimal bandwidth of the coefficient matrices is determined by the marginal Bayesian information criterion. Both the asymptotic properties and numerical results show that the proposed models perform well.

Acknowledgements

We thank the editors and the referees for their insightful comments leading to a substantial improvement of the paper. This research was supported by the National Social Science Foundation of China under Grant [17CTJ016].

Appendix. Proofs

Lemma 1. Under conditions A1, A2, and A4, we have

$$\frac{1}{n^*}\boldsymbol{\mathcal{X}}^{\mathsf{T}}\mathbf{W}_0\boldsymbol{\mathcal{X}} \to_p \mathbf{U}.$$

Lemma 1 is similar to Lemma 2.1 in Hallin et al. (2004), so we omit the proof. For our purposes, we make a new notation before Lemma 2.

$$\mathbf{H}_{\boldsymbol{n}} = \begin{bmatrix} \frac{1}{n^*}\sum\limits_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum\limits_{t\in T_{\boldsymbol{n}}} \boldsymbol{K}_{\boldsymbol{i},t}(\rho_{\boldsymbol{i},t}+R_{\boldsymbol{i},t}) \\ \frac{1}{n^*}\sum\limits_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum\limits_{t\in T_{\boldsymbol{n}}} (\frac{X_{\boldsymbol{i},t}-x_0}{\boldsymbol{h}})\boldsymbol{K}_{\boldsymbol{i},t}(\rho_{\boldsymbol{i},t}+R_{\boldsymbol{i},t}) \end{bmatrix}$$

where $\rho_{\boldsymbol{i},t} = m(\boldsymbol{X}_{\boldsymbol{i},t}) - m(\boldsymbol{x}_0) - \boldsymbol{h}^{\mathsf{T}}m^{'}(\boldsymbol{x}_0)$.

Lemma 2. If conditions A1-A4 hold, then

$$(n^*h_1\cdots h_l)^{\frac{1}{2}}E(\mathbf{H}_{\boldsymbol{n}}) \to 0.$$

Proof. It suffices to prove

$$(n^*h_1\cdots h_l)^{\frac{1}{2}}\frac{1}{n^*}\sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}} E\left(\boldsymbol{K}_{\boldsymbol{i},t}(\rho_{\boldsymbol{i},t}+R_{\boldsymbol{i},t})\right) \to 0.$$

We first rewritten $\rho_{\boldsymbol{i},t}$ by the Taylor expansion,

$$\rho_{\boldsymbol{i},t} = (\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0)^{\mathsf{T}}m^{''}(x_0+\rho(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0))(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0),$$

where $|\rho|<1$. Since the explanatory variable is independent of the error term, $E(\boldsymbol{K}_{\boldsymbol{i},t}R_{\boldsymbol{i},t}) = E(\boldsymbol{K}_{\boldsymbol{i},t})E(R_{\boldsymbol{i},t}) = 0$, and their expectations are both exist. By condition A4,

$$\begin{aligned}
& (n^*h_1\cdots h_l)^{\frac{1}{2}}\frac{1}{n^*}\sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}} E\left(\boldsymbol{K}_{\boldsymbol{i},t}(\rho_{\boldsymbol{i},t}+R_{\boldsymbol{i},t})\right) \\
= \;& (n^*h_1\cdots h_l)^{-\frac{1}{2}}\sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}} E\left\{\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})(\rho_{\boldsymbol{i},t}+R_{\boldsymbol{i},t})\right\} \\
\leq \;& \boldsymbol{h}_{\max}^2(n^*h_1\cdots h_l)^{-\frac{1}{2}}\sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}} E\left\{\boldsymbol{K}(\boldsymbol{u})\rho_{\boldsymbol{i},t}(\boldsymbol{u})\right\} \\
= \;& \boldsymbol{h}_{\max}^2(n^*h_1\cdots h_l)^{\frac{1}{2}}f(x_0)\int_{\mathbb{R}^l}\boldsymbol{K}(\boldsymbol{u})\rho_{\boldsymbol{i},t}(\boldsymbol{u})d\boldsymbol{u} \\
\leq \;& C\boldsymbol{h}_{\max}^2(n^*h_1\cdots h_l)^{\frac{1}{2}} \to 0.
\end{aligned}$$

This completes the proof of Lemma 2.

Proof of Theorem 1. By equation (3), we have

$$\begin{aligned} R_{\boldsymbol{i},t} - \widehat{R}_{\boldsymbol{i},t} &= \hat{m}(x_0) - m(x_0) \\ &= \frac{1}{n^*}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\boldsymbol{\mathcal{X}}\mathbf{W}_0 \begin{bmatrix} \rho_{1,1} + R_{1,1} \\ \vdots \\ \rho_{\bar{n},n_0} + R_{\bar{n},n_0} \end{bmatrix} \\ &= \frac{1}{n^*}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\boldsymbol{\mathcal{X}}\mathbf{W}_0 \begin{bmatrix} \rho_{1,1} \\ \vdots \\ \rho_{\bar{n},n_0} \end{bmatrix} + \frac{1}{n^*}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\boldsymbol{\mathcal{X}}\mathbf{W}_0 \begin{bmatrix} R_{1,1} \\ \vdots \\ R_{\bar{n},n_0} \end{bmatrix} \\ &\equiv \mathbf{A}_1 + \mathbf{A}_2 \end{aligned}$$

for the first part. In addition, we can easily derive that the equation above equals $\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\mathbf{H}_{\boldsymbol{n}}$. Then, to prove the zero mean, it suffices to prove $E(\sqrt{n^*h_1\cdots h_l}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\mathbf{H}_{\boldsymbol{n}}) = 0$. The Lemma 1 means that we just need to prove that for any vector $\boldsymbol{c} = (c_0, \cdots, c_l)^{\mathsf{T}}$, it holds

$$E\left((n^*h_1\cdots h_l)^{\frac{1}{2}}\boldsymbol{c}^{\mathsf{T}}\mathbf{H}_{\boldsymbol{n}}\right) = 0. \tag{18}$$

It is easy to prove that holds by Lemma 2. Next, we consider the asymptotic behaviour of the first part by observing $\mathbf{A}_1$ and $\mathbf{A}_2$. After that, we know that it suffices to prove

$$(n^*h_1\cdots h_l)^{\frac{1}{2}}\mathbf{A}_1 \to 0 \tag{19}$$

and

$$(n^*h_1\cdots h_l)^{\frac{1}{2}}\mathbf{A}_2 \to_L N(\mathbf{0}, \sigma^2\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}^{-1}\boldsymbol{\Sigma}(\mathbf{U}^{-1})^{\mathsf{T}}\boldsymbol{e}_1).$$

Without loss of generality, we take $d = 2$ in the proof process, then $\boldsymbol{i}$ can be rewritten (i, j). Similar to the proof of Lemma 2.1 in Hallin et al. (2004), by Taylor expansion, conditions A2 and A4,

$$\begin{aligned} & (n^*h_1\cdots h_l)^{\frac{1}{2}}\mathbf{A}_1 \\ = \; & \left(\frac{h_1\cdots h_l}{n^*}\right)^{\frac{1}{2}}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}\boldsymbol{\mathcal{X}}\mathbf{W}_0 \begin{bmatrix} \rho_{1,1} \\ \vdots \\ \rho_{\bar{n},n_0} \end{bmatrix} \\ = \; & (n^*h_1\cdots h_l)^{-\frac{1}{2}}\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1} \begin{bmatrix} \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}}\left\{\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})\rho_{\boldsymbol{i},t}\right\} \\ \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}}\left\{\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}}\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})\rho_{\boldsymbol{i},t}\right\} \end{bmatrix} \\ = \; & \frac{\boldsymbol{e}_1^{\mathsf{T}}\mathbf{U}_{\boldsymbol{n}}^{-1}}{n^*h_1\cdots h_l} \begin{bmatrix} \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}}\left\{\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0)^{\mathsf{T}}m''(\boldsymbol{x}_0+\rho(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0))(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0)\right\} \\ \sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}}\left\{\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}}\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0)^{\mathsf{T}}m''(\boldsymbol{x}_0+\rho(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0))(\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0)\right\} \end{bmatrix} \\ \to_p \; & (n^*h_1\cdots h_l)^{\frac{1}{2}}\boldsymbol{h}_{\max}^2 \begin{bmatrix} \frac{1}{2}f(\boldsymbol{x}_0)\sum_{i=1}^{l}\sum_{j=1}^{l}\frac{\partial^2 m(\boldsymbol{x}_0)}{\partial\boldsymbol{x}_{0_i}\partial\boldsymbol{x}_{0_j}}\int_{\mathbb{R}^l} z_iz_j\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} \\ \frac{1}{2}f(\boldsymbol{x}_0)\sum_{i=1}^{l}\sum_{j=1}^{l}\frac{\partial^2 m(\boldsymbol{x}_0)}{\partial\boldsymbol{x}_{0_i}\partial\boldsymbol{x}_{0_j}}\int_{\mathbb{R}^l} z_iz_j\boldsymbol{z}\boldsymbol{K}(\boldsymbol{z})d\boldsymbol{z} \end{bmatrix} \\ \to_p \; & 0. \end{aligned}$$

To prove (20), similar to (18), it suffices to prove that for any vector $\boldsymbol{c} = (c_0, \cdots, c_l)^{\mathsf{T}} = (c_0, (\boldsymbol{c}')^{\mathsf{T}})^{\mathsf{T}}$, it holds

$$\Delta = (n^*h_1\cdots h_l)^{-\frac{1}{2}}\sum_{\boldsymbol{i}\in\Lambda_{\boldsymbol{n}}}\sum_{t\in T_{\boldsymbol{n}}}\left\{\left[c_0 + (\boldsymbol{c}')^{\mathsf{T}}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})\right]\boldsymbol{K}(\frac{\boldsymbol{X}_{\boldsymbol{i},t}-\boldsymbol{x}_0}{\boldsymbol{h}})R_{\boldsymbol{i},t}\right\} \to_d N(\mathbf{0}, \sigma^2\boldsymbol{c}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{c}). \tag{20}$$

We know the expectation of Δ is zero already from the front, and similar to the proof of Lemma 1, it is easy to prove the variance of Δ is equal to $\sigma^2 \boldsymbol{c}^{\mathsf{T}} \boldsymbol{\Sigma} \boldsymbol{c}$. By conditions A1-A5, (20) is hold. Thus, the proof of Theorem 1 is completed.

Proof of Theorem 2.

Our goal is to prove that $P(\hat{k}_0 = k_0) \to 1$. It is sufficient to show that $P(\hat{k}_0 < k_0) \to 0$ and $P(\hat{k}_0 > k_0) \to 0$. Our proof follows the arguments in Wang et al. (2009).

Without loss of generality, we take $p = 1$ in the proof process. Consider the first case. Since $\hat{k}_0 = \max_{1 \le i_0 \le \bar{n}} \hat{k}_{i_0}$, then it holds that $P(\hat{k}_0 < k_0) \le P(\hat{k}_{i_0} < k_0)$ for some $i_0 \in \{1, \cdots, \bar{n}\}$, and the event $(\hat{k}_{i_0} < k_0)$ implies

$$\boldsymbol{\Psi}_{i_0, n_0} \equiv \min_{k < k_0} BIC_{i_0}(k) < BIC_{i_0}(k_0).$$

That means we just need to prove $P(\boldsymbol{\Psi}_{i_0, n_0}) \to 0$ for some x. We observe the situation under $k = k_0$ first. By (4), we have

$$\widehat{\mathbf{V}}_{i_0} = (\widehat{\boldsymbol{\Sigma}}_p^{\mathsf{T}}, \cdots, \widehat{\boldsymbol{\Sigma}}_0), \quad \boldsymbol{\beta}_{i_0} = ((\boldsymbol{d}_{i_0}^{0'})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0}^{p'})^{\mathsf{T}})^{\mathsf{T}},$$

which all correspond to the non-zero elements of $((\boldsymbol{d}_{i_0}^{0})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0}^{p})^{\mathsf{T}})^{\mathsf{T}}$. It follows from the compatibility of matrix L_2 norm and some properties of projection matrix that

$$\begin{aligned}
\mathrm{RSS}_{i_0}(k_0) &= \frac{1}{\bar{n}} \left\| \widehat{\mathbf{z}}_{i_0} - \widehat{\mathbf{V}}_{i_0} \widehat{\boldsymbol{\beta}}_{i_0} \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| \widehat{\mathbf{V}}_{i_0} \boldsymbol{\beta}_{i_0} + \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0, t} - \widehat{\mathbf{V}}_{i_0} \widehat{\boldsymbol{\beta}}_{i_0} \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0}) \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0, t} \right\|_2^2 \\
&\le \frac{1}{\bar{n}} \left\| \mathbf{I} - \mathbf{H}_{i_0} \right\|_2^2 \left\| \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0, t} \right\|_2^2 \\
&\le \frac{1}{\bar{n}} \left\| \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0, t} \right\|_2^2,
\end{aligned}$$

where $\mathbf{H}_{i_0} = \widehat{\mathbf{V}}_{i_0}(\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0})^{-1}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}$. By proposition 2.5 of Fan and Yao (2003), we have

$$\begin{aligned}
& E\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right)^2 \\
= \;& \mathrm{Var}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{x,t}\right) + E^2\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \\
= \;& \mathrm{Var}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) + 0 \\
= \;& \frac{1}{n_0^2}\sum_{t=p+1}^{n_0}\mathrm{Var}\left(\sum_{t=p+1}^{n_0}\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) + \frac{1}{n_0^2}\sum_{t\neq s}\mathrm{Cov}(\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}, \boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{s-p}\varepsilon_{i_0,s}) \\
= \;& \frac{C}{n_0} + \frac{1}{n_0^2}\sum_{t\neq s}8\alpha(|t-s|)^{\frac{\gamma}{4+\gamma}}\left(E|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}|^{2+\frac{\gamma}{2}}\right)^{\frac{2}{4+\gamma}} \times \left(E|\boldsymbol{e}_j^{\mathsf{T}}\boldsymbol{R}_{s-p}\varepsilon_{i_0,s}|^{2+\frac{\gamma}{2}}\right)^{\frac{2}{4+\gamma}} \\
\leq \;& \frac{C}{n_0} + \frac{C}{n_0^2}\sum_{t\neq s}\alpha(|t-s|)^{\frac{\gamma}{4+\gamma}} \\
\leq \;& \frac{C}{n_0} + \frac{C}{n_0}\sum_{j=1}^{n_0}\alpha(j)^{\frac{\gamma}{4+\gamma}} \\
= \;& O(\frac{1}{n_0})
\end{aligned}$$

for $j = 1, \cdots, \bar{n}$, where $\alpha(j)$ is the mixing coefficient, and it is easy to prove that $\sum_{j=1}^{\infty}\alpha(j)^{\frac{\gamma}{4+\gamma}} < \infty$ for some positive constant γ by Theorem 2.1 of Pham and Tran (1985). Meanwhile, the constant C is independent of $\bar{n}$, then we obtain

$$\left\|\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_2^2 = \sum_{j=1}^{\bar{n}}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right)^2 = O_p(\frac{\bar{n}}{n_0}).$$

Hence

$$\mathrm{RSS}_{i_0}(k_0) = O_p(\frac{1}{n_0}). \tag{21}$$

For $k < k_0$, we set

$$\begin{aligned}
\widehat{\mathbf{V}}_{i_0} &= (\mathbf{A}_{i_0,k}^{(1)}, \widehat{\boldsymbol{\Sigma}}_{p,k}^{\mathsf{T}}, \mathbf{A}_{i_0,k}^{(2)}, \mathbf{A}_{i_0,k}^{(3)}, \widehat{\boldsymbol{\Sigma}}_{p-1,k}^{\mathsf{T}}, \mathbf{A}_{i_0,k}^{(4)}, \cdots, \mathbf{A}_{i_0,k}^{(2p+1)}, \widehat{\boldsymbol{\Sigma}}_{0,k}, \mathbf{A}_{i_0,k}^{(2p+2)}), \\
\boldsymbol{\beta}_{i_0} &= \left((\boldsymbol{d}_{i_0,k}^{0(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{0})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{0(2)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{1(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{1})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{1(2)})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0,k}^{p(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{p})^{\mathsf{T}}, (\boldsymbol{d}_{i_0,k}^{p(2)})^{\mathsf{T}}\right)^{\mathsf{T}},
\end{aligned}$$

where $\widehat{\mathbf{V}}_{i_0,k} = (\widehat{\boldsymbol{\Sigma}}_{p,k}^{\mathsf{T}}, \cdots, \widehat{\boldsymbol{\Sigma}}_{0,k})$ and $\boldsymbol{\beta}_{i_0} = ((\boldsymbol{d}_{i_0,k}^{0})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0,k}^{p})^{\mathsf{T}})^{\mathsf{T}}$, which correspond to $\tau_{i_0}(k)$ non-zero elements of $((\boldsymbol{d}_{i_0}^{0})^{\mathsf{T}}, \cdots, (\boldsymbol{d}_{i_0}^{p})^{\mathsf{T}})^{\mathsf{T}}$. Meanwhile, note $\mathbf{A}_{i_0,k} = (\mathbf{A}_{i_0,k}^{(1)}, \mathbf{A}_{i_0,k}^{(2)}, \cdots, \mathbf{A}_{i_0,k}^{(2p+2)})$ and $\boldsymbol{d}_{i_0,k} =$

$((\boldsymbol{d}_{i_0,k}^{0(1)})^\mathsf{T}, (\boldsymbol{d}_{i_0,k}^{0(2)})^\mathsf{T}, \cdots, (\boldsymbol{d}_{i_0,k}^{p(1)})^\mathsf{T}, (\boldsymbol{d}_{i_0,k}^{p(2)})^\mathsf{T})^\mathsf{T}$. By (10) and (7),

$$\begin{aligned}
\mathrm{RSS}_{i_0}(k) &= \frac{1}{\bar{n}} \left\| \widehat{\mathbf{z}}_{i_0} - \widehat{\mathbf{V}}_{i_0,k} (\widehat{\mathbf{V}}_{i_0,k}^\mathsf{T} \widehat{\mathbf{V}}_{i_0,k})^{-1} \widehat{\mathbf{V}}_{i_0,k}^\mathsf{T} \widehat{\mathbf{z}}_{i_0} \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \widehat{\mathbf{z}}_{i_0} \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) (\widehat{\mathbf{V}}_{i_0,k} \boldsymbol{\beta}_{i_0,k} + \mathbf{A}_{i_0,k} \boldsymbol{d}_{i_0,k} + \frac{1}{n_0} \sum_{t=p+1}^{n_0} \mathbf{R}_{t-p} \varepsilon_{i_0,t}) \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \mathbf{A}_{i_0,k} \boldsymbol{d}_{i,k} + (\mathbf{I} - \mathbf{H}_{i_0,k}) \frac{1}{n_0} \sum_{t=p+1}^{n_0} \mathbf{R}_{t-p} \varepsilon_{i_0,t} \right\|_2^2 \\
&= \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \mathbf{A}_{i_0,k} \boldsymbol{d}_{i_0,k} \right\|_2^2 + \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \frac{1}{n_0} \sum_{t=p+1}^{n_0} \mathbf{R}_{t-p} \varepsilon_{i_0,t} \right\|_2^2 \\
&\quad + \frac{2}{\bar{n}} \boldsymbol{d}_{i_0,k}^\mathsf{T} \mathbf{A}_{i_0,k}^\mathsf{T} (\mathbf{I} - \mathbf{H}_{i_0,k}) \frac{1}{n_0} \sum_{t=p+1}^{n_0} \mathbf{R}_{t-p} \varepsilon_{i_0,t}, \qquad (22)
\end{aligned}$$

where $\mathbf{H}_{i_0,k} = \widehat{\mathbf{V}}_{i_0,k} (\widehat{\mathbf{V}}_{i_0,k}^\mathsf{T} \widehat{\mathbf{V}}_{i_0,k})^{-1} \widehat{\mathbf{V}}_{i_0,k}^\mathsf{T}$. For the orthogonal matrix $\mathbf{P}$ and diagonal matrix $\boldsymbol{\Lambda}$, it follows from condition A7 and properties of spatial weight matrix that

$$\begin{aligned}
\frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \mathbf{A}_{i_0,k} \boldsymbol{d}_{i_0,k} \right\|_2^2 &= \frac{1}{\bar{n}} \boldsymbol{d}_{i_0,k}^\mathsf{T} \mathbf{A}_{i_0,k}^\mathsf{T} (\mathbf{I} - \mathbf{H}_{i_0,k}) \mathbf{A}_{i_0,k} \boldsymbol{d}_{i_0,k} \\
&= \frac{1}{\bar{n}} \boldsymbol{d}_{i_0,k}^\mathsf{T} \mathbf{P}^\mathsf{T} \boldsymbol{\Lambda} \mathbf{P} \boldsymbol{d}_{i_0,k} \\
&\geq \frac{1}{\bar{n}} \boldsymbol{d}_{i_0,k}^\mathsf{T} \mathbf{P}^\mathsf{T} \lambda_1 \mathbf{I} \mathbf{P} \boldsymbol{d}_{i_0,k} \\
&\geq \frac{1}{\bar{n}} \lambda_1 \sum_{j=0}^{p} ((d_{i_0,i_0-k_0}^j)^2 + (d_{i_0,i_0+k_0}^j)^2).
\end{aligned}$$

And it is easy to get the upper boundary $\lambda_2 \|\boldsymbol{\beta}_{i_0}\|_2^2 / \bar{n}$. We can further relax the boundaries of the first term of (22) to

$$\frac{C_{n_0} k_0 \lambda_1 \log(\bar{n} \vee n_0)}{\bar{n} n_0} \leq \frac{1}{\bar{n}} \left\| (\mathbf{I} - \mathbf{H}_{i_0,k}) \mathbf{A}_{i_0,k} \boldsymbol{d}_{i_0,k} \right\|_2^2 \leq \frac{\lambda_2 k_0 O(1)}{\bar{n}} \qquad (23)$$

by conditions A6. The order of second term is same as that of $\mathrm{RSS}_{i_0}(k_0)$, and the third term can be bounded by the sum of the first and the second terms by Cauchy-Schwarz inequality. Therefore, under the condition of k_0 is fixed,

$$\frac{C_{n_0} k_0 \lambda_1 \log(\bar{n} \vee n_0)}{\bar{n} n_0} + O_p(\frac{1}{n_0}) \leq \mathrm{RSS}_{i_0}(k) \leq \frac{O(1) \lambda_2}{\bar{n}}.$$

For the deviance of number of elements between different bandwidth k, we have $\tau_{i_0}(k_0) - \tau_{i_0}(k) = \mu(\bar{n}+1)$ with a positive constant μ by (5). Then we get that

$$\begin{aligned}
&\min_{k<k_0} BIC_{i_0}(k) - BIC_{i_0}(k_0) \\
= \ & \log \frac{\mathrm{RSS}_{i_0}(k)}{\mathrm{RSS}_{i_0}(k_0)} + \frac{1}{n_0} C_{n_0} \log(\bar{n} \vee n_0)(\tau_{i_0}(k) - \tau_{i_0}(k_0)) \\
\geq \ & \log(1 + \frac{C_{n_0} k_0 \lambda_1 \log(\bar{n} \vee n_0)}{\bar{n}}) - \frac{1}{n_0} \mu C_{n_0} (\bar{n}+1) \log(\bar{n} \vee n_0) \to \infty
\end{aligned}$$

with the lower bound of $\mathrm{RSS}_{i_0}(k)$. Therefore, we obtain $P(\boldsymbol{\Psi}_{i_0,n_0}) \to 0$. The lower bound of $\mathrm{RSS}_{i_0}(k)$ stays same for $k_0 = o\left(C_{n_0}^{-1} n_0 / \log(\bar{n} \vee n_0)\right)$, it means the result of (24) still holds.

Similarly, we also know $P(\hat{k}_0 > k_0) \leq P(\hat{k}_{i_0} > k_0)$ for some $i_0 \in \{1, \cdots, \bar{n}\}$, and the event $(\hat{k}_{i_0} > k_0)$ implies

$$\boldsymbol{\Phi}_{i_0,n_0} \equiv \frac{\min_{k>k_0} BIC_{i_0}(k)}{BIC_{i_0}(k_0)} < 1,$$

it suffices to prove $P(\mathbf{\Phi}_{i_0,n_0}) \to 0$ for some i_0. Let

$$\widehat{\mathbf{V}}_{i_0,k} = (\mathbf{B}_{i_0,k}^{(1)}, \widehat{\mathbf{\Sigma}}_p^{\mathsf{T}}, \mathbf{B}_{i_0,k}^{(2)}, \mathbf{B}_{i_0,k}^{(3)}, \widehat{\mathbf{\Sigma}}_{p-1}^{\mathsf{T}}, \mathbf{B}_{i_0,k}^{(4)}, \cdots, \mathbf{B}_{i_0,k}^{(2p+1)}, \widehat{\mathbf{\Sigma}}_0, \mathbf{B}_{i_0,k}^{(2p+2)}),$$
$$\boldsymbol{\beta}_{i_0,k} = \left((\boldsymbol{b}_{i_0,k}^{0(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0}^{0'})^{\mathsf{T}}, (\boldsymbol{b}_{i_0,k}^{0(2)})^{\mathsf{T}}, (\boldsymbol{b}_{i_0,k}^{1(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0}^{1'})^{\mathsf{T}}, (\boldsymbol{b}_{i_0,k}^{1(2)})^{\mathsf{T}}, \cdots, (\boldsymbol{b}_{i_0,k}^{p(1)})^{\mathsf{T}}, (\boldsymbol{d}_{i_0}^{p'})^{\mathsf{T}}, (\boldsymbol{b}_{i_0,k}^{p(2)})^{\mathsf{T}}\right)^{\mathsf{T}},$$

then the residual sum of squares can be rewritten as

$$\mathrm{RSS}_{i_0}(k) = \frac{1}{\bar{n}} \min_{\boldsymbol{v}_1 \boldsymbol{v}_2} \left\| \widehat{\mathbf{z}}_{i_0} - \widehat{\mathbf{V}}_{i_0} \boldsymbol{v}_1 - \mathbf{B}_{i_0,k} \boldsymbol{v}_2 \right\|_2^2, \tag{24}$$

where $\mathbf{B}_{i_0,k} = (\mathbf{B}_{i_0,k}^{(1)}, \mathbf{B}_{i_0,k}^{(2)}, \cdots, \mathbf{B}_{i_0,k}^{(2p+1)}, \mathbf{B}_{i_0,k}^{(2p+2)})$. By the least squares method, we obtain

$$\widehat{\boldsymbol{v}}_1 = (\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0})^{-1} \widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \left[\mathbf{I} - \mathbf{B}_{i_0,k} (\mathbf{G}_{i_0,k}^{\mathsf{T}} \mathbf{G}_{i_0,k})^{-1} \mathbf{G}_{i_0,k}^{\mathsf{T}} \right] \widehat{\mathbf{z}}_{i_0},$$
$$\widehat{\boldsymbol{v}}_2 = (\mathbf{G}_{i_0,k}^{\mathsf{T}} \mathbf{G}_{i_0,k})^{-1} \mathbf{S}_{i_0,k}^{\mathsf{T}} (\mathbf{I} - \mathbf{H}_{i_0}) \widehat{\mathbf{z}}_{i_0} = (\mathbf{G}_{i_0,k}^{\mathsf{T}} \mathbf{G}_{i_0,k})^{-1} \mathbf{G}_{i_0,k}^{\mathsf{T}} \widehat{\mathbf{z}}_{i_0},$$

where $\mathbf{G}_{i_0,k} = (\mathbf{I} - \mathbf{H}_{i_0}) \mathbf{S}_{i_0,k}$. Obviously, we can see that $\widehat{\mathbf{z}}_{i_0}^{\mathsf{T}} \mathbf{G}_{i_0,k} \widehat{\boldsymbol{v}}_2 = \widehat{\boldsymbol{v}}_2^{\mathsf{T}} \mathbf{G}_{i_0,k}^{\mathsf{T}} \widehat{\mathbf{z}}_x = \widehat{\boldsymbol{v}}_2^{\mathsf{T}} \mathbf{G}_{i_0,k}^{\mathsf{T}} \mathbf{G}_{i_0,k} \widehat{\boldsymbol{v}}_2$. It follows from (21) and (24) that

$$\begin{aligned}
\mathrm{RSS}_{i_0}(k) &= \frac{1}{\bar{n}} \left\| \widehat{\mathbf{z}}_{i_0} - \widehat{\mathbf{V}}_{i_0} \widehat{\boldsymbol{v}}_1 - \mathbf{B}_{i_0,k} \widehat{\boldsymbol{v}}_2 \right\|_2^2 \\
&= \mathrm{RSS}_{i_0}(k_0) - \frac{1}{\bar{n}} \| \mathbf{G}_{i_0,k} \widehat{\boldsymbol{v}}_2 \|_2^2 \\
&\leq O_p(\frac{1}{n_0}) + \frac{1}{\bar{n}} \| \mathbf{G}_{i_0,k} (\mathbf{G}_{i_0,k}^{\mathsf{T}} \mathbf{G}_{i_0,k})^{-1} \mathbf{G}_{i_0,k}^{\mathsf{T}} \|_2^2 \| \frac{1}{n_0} \sum_{t=p+1}^{n_0} \mathbf{R}_{t-p} \varepsilon_{i_0,t} \|_2^2 \\
&= O_p(\frac{1}{n_0}).
\end{aligned} \tag{25}$$

It is briefly that $\tau_{i_0}(k) - \tau_{i_0}(k_0) = \mu(\bar{n}+1)$, and the constant $\mu > 0$. Hence

$$\begin{aligned}
\frac{\min_{k>k_0} BIC_{i_0}(k)}{BIC_{i_0}(k_0)} &= \frac{-\log n_0 + \frac{1}{n_0} C_{n_0} (\tau_{i_0}(k_0) + \mu(\bar{n}+1)) \log(\bar{n} \vee n_0)}{-logn_0 + \frac{1}{n_0} C_{n_0} \tau_{i_0}(k_0) \log(\bar{n} \vee n_0)} \\
&= \frac{O_p(\frac{1}{n_0} \mu C_{n_0} (\bar{n}+1)) - 1}{O_p(\frac{1}{n_0} C_{n_0} \tau_{i_0}(k_0)) - 1} \\
&= 1.
\end{aligned}$$

Therefore, we obtain $P(\min_{k>k_0} BIC_{i_0}(k) = BIC_{i_0}(k_0)) \to 1$, it is equal to $P(\mathbf{\Phi}_{i_0,n_0}) \to 0$ for all types of k_0. The proof of Theorem 2 is completed.

Proof of Theorem 3.

Under the condition $\bar{n} = o(n_0)$ of Theorem 3, the result of Theorem 2 holds. To prove Part (i) of Theorem 3 for a fixed bandwidth k_0 over the conclusion of Theorem 2, it is equivalent to prove $\sqrt{n_0} \mathbf{U}_{i_0}^{-\frac{1}{2}} \mathbf{K}_{i_0} (\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0})^{-1} \widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}$ is asymptotically normal, then we just need to verify the assertion (1) and (2) below by (8).

(1)

$$\sqrt{n_0} \mathbf{U}_{i_0}^{-\frac{1}{2}} \begin{pmatrix}
\frac{1}{n_0} \sum_{t=p+1}^{n_0} (R_{j,t})_{j \in S_{i_0}} \boldsymbol{R}_{t-p}^{\mathsf{T}} \left(\frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \right) \\
\frac{1}{n_0} \sum_{t=p+1}^{n_0} (R_{j,t-1})_{j \in S_{i_0}^+} \boldsymbol{R}_{t-p}^{\mathsf{T}} \left(\frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \right) \\
\vdots \\
\frac{1}{n_0} \sum_{t=p+1}^{n_0} (R_{j,t-p})_{j \in S_{i_0}^+} \boldsymbol{R}_{t-p}^{\mathsf{T}} \left(\frac{1}{n_0} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \right)
\end{pmatrix} \to_d N(0, \mathbf{I}_{\tau_{i_0}}),$$

(2)

$$\mathbf{K}_{i_0} (\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}} \widehat{\mathbf{V}}_{i_0})^{-1} \to_p \mathbf{I}_{\tau_{i_0}},$$

To prove assertion (1), we just need to prove that for any nonzero vector $\boldsymbol{a} = (\boldsymbol{a}_0^\mathsf{T}, \cdots, \boldsymbol{a}_p^\mathsf{T})$, where $\boldsymbol{a}_0 \in \mathbb{R}^{S_{i_0}}$ and $\boldsymbol{a}_j \in \mathbb{R}^{S_{i_0}^+}$ for other values of j, the linear combination

$$\sqrt{n_0}\boldsymbol{a}^\mathsf{T}\begin{pmatrix} \frac{1}{n_0}\sum_{t=p+1}^{n_0}(R_{j,t})_{j\in S_x}\boldsymbol{R}_{t-p}^\mathsf{T}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0}(R_{j,t-1})_{j\in S_{i_0}^+}\boldsymbol{R}_{t-p}^\mathsf{T}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \\ \vdots \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0}(R_{j,t-p})_{j\in S_{i_0}^+}\boldsymbol{R}_{t-p}^\mathsf{T}\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \end{pmatrix}$$

is asymptotic normal. Let us consider one term for each $j \in S_{i_0}$ in the first block of (26) first, thus we have

$$\begin{aligned} & \frac{1}{n_0}\sum_{t=p+1}^{n_0} R_{j,t}\boldsymbol{R}_{t-p}^\mathsf{T}(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}) \\ = & \frac{1}{n_0}\sum_{t=p+1}^{n_0}(R_{j,t}\boldsymbol{R}_{t-p}^\mathsf{T} - E(R_{j,t}\boldsymbol{R}_{t-p}^\mathsf{T}))\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} \\ & + \frac{n_0-1}{n_0}E(R_{j,t}\boldsymbol{R}_{t-p}^\mathsf{T})\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} \\ = & \left[\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}^\mathsf{T}(\boldsymbol{e}_j^\mathsf{T}\boldsymbol{R}_t) - \boldsymbol{e}_j^\mathsf{T}\boldsymbol{\Sigma}_p\right]\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} \\ & + \frac{n_0-1}{n_0}\boldsymbol{e}_j^\mathsf{T}\boldsymbol{\Sigma}_1\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} \\ = & E_1 + E_2. \end{aligned}$$

By similar method of (21) for term E_1 and $k = 1, \cdots, \bar{n}$, we have

$$E\left[\frac{1}{n_0}\sum_{t=p+1}^{n_0}\left((\boldsymbol{e}_k^\mathsf{T}\boldsymbol{R}_{t-p})(\boldsymbol{e}_j^\mathsf{T}\boldsymbol{R}_t) - \boldsymbol{e}_k^\mathsf{T}\boldsymbol{\Sigma}_p^\mathsf{T}\boldsymbol{e}_j\right)\right]^2 \le O(\frac{1}{n_0}).$$

Then,

$$\left\|\frac{1}{n_0}\sum_{t=p+1}^{n_0}\left(\boldsymbol{R}_{t-p}(\boldsymbol{e}_j^\mathsf{T}\boldsymbol{R}_t) - \boldsymbol{e}_k^\mathsf{T}\boldsymbol{\Sigma}_p^\mathsf{T}\boldsymbol{e}_j\right)\right\|_2^2 = O_p(\frac{\bar{n}}{n_0}).$$

Thus,

$$E_1 \le \left\|\frac{1}{n_0}\sum_{t=p+1}^{n_0}\left(\boldsymbol{R}_{t-p}(\boldsymbol{e}_j^\mathsf{T}\boldsymbol{R}_t) - \boldsymbol{e}_k^\mathsf{T}\boldsymbol{\Sigma}_p^\mathsf{T}\boldsymbol{e}_j\right)\right\|_2 \left\|\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_2 = O_p(\frac{\bar{n}}{n_0}). \tag{26}$$

Similarly, we obtain $E_2 = O(\frac{1}{\sqrt{n_0}})$. Under the condition $\bar{n} = o(\sqrt{n_0})$, it holds that $\sqrt{n_0}E_1 = o_p(1)$ and $\sqrt{n_0}E_2 = O_p(1)$. Hence,

$$\frac{1}{\sqrt{n_0}}\sum_{t=p+1}^{n_0} R_{j,t}\boldsymbol{R}_{t-p}^\mathsf{T}(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}) = \boldsymbol{e}_j^\mathsf{T}\boldsymbol{\Sigma}_p\frac{1}{\sqrt{n_0}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} + o_p(1), \qquad j \in S_{i_0}.$$

Similarly, we obtain

$$\frac{1}{\sqrt{n_0}}\sum_{t=p+1}^{n_0} R_{j,t-q}\boldsymbol{R}_{t-p}^\mathsf{T}(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}) = \boldsymbol{e}_j^\mathsf{T}\boldsymbol{\Sigma}_{p-q}\frac{1}{\sqrt{n_0}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t} + o_p(1), \qquad j \in S_{i_0}^+$$

for $q = 1, \cdots, p$. Now it suffices to prove

$$S_{\bar{n},n_0} \equiv \boldsymbol{a}_0^{\mathsf{T}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \frac{1}{\sqrt{n_0}} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} + \sum_{j=0}^{p-1} \boldsymbol{a}_{p-j}^{\mathsf{T}} \mathbf{I}_{S_{i_0}^+}^{\mathsf{T}} \boldsymbol{\Sigma}_j \frac{1}{\sqrt{n_0}} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}$$

is asymptotic normal.

Obviously, the mean of $S_{\bar{n},n_0}$ is zero. Next, we calculate the variance of one term of $S_{\bar{n},n_0}$ first,

$$\begin{aligned} & \mathrm{Var}(\boldsymbol{a}_0^{\mathsf{T}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \frac{1}{\sqrt{n_0}} \sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}) \\ = \; & \boldsymbol{a}_0^{\mathsf{T}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \frac{n_0 - p}{n_0} \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(0) \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} \boldsymbol{a}_0 \\ & + \boldsymbol{a}_0^{\mathsf{T}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \sum_{j=1}^{n_0-p-1} (1 - \frac{p+j}{n_0}) \left[\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(j) + \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}^{\mathsf{T}}(j) \right] \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} \boldsymbol{a}_0. \end{aligned} \tag{27}$$

We note that it holds that

$$E|\boldsymbol{e}_j^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}|^{\frac{4+\gamma}{2}} \leq [E|\boldsymbol{e}_j^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{R}_{t-p}|^{4+\gamma}]^{\frac{1}{2}} [E|\varepsilon_{i_0,t}|^{4+\gamma}]^{\frac{1}{2}} \leq \infty.$$

by the probability form of Cauchy inequality. It follows from the discrete form of Hölder inequality that

$$\begin{aligned} & \sup_{\bar{n}} \sum_{j=1}^{\infty} \left| \boldsymbol{a}_0^{\mathsf{T}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \boldsymbol{\Sigma}_p \left[\boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(j) + \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}^{\mathsf{T}}(j) \right] \boldsymbol{\Sigma}_p^{\mathsf{T}} \mathbf{I}_{S_{i_0}} \boldsymbol{a}_0 \right| \\ \leq \; & C \sup_{j_1,j_2 \leq \bar{n}} \sum_{j=1}^{\infty} \left| \boldsymbol{e}_{j_1}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_{\boldsymbol{R},\boldsymbol{\varepsilon}_{i_0}}(j) \boldsymbol{\Sigma}_p^{\mathsf{T}} \boldsymbol{e}_{j_2} \right| \\ \leq \; & C \sup_{j_1,j_2 \leq \bar{n}} \sum_{t-s=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} \left[\left(E \left| \boldsymbol{e}_{j_1}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \right|^{4+\gamma} \right)^{\frac{1}{4+\gamma}} \left(E \left| \boldsymbol{e}_{j_2}^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{R}_{s-p} \varepsilon_{i_0,s} \right|^{4+\gamma} \right)^{\frac{1}{4+\gamma}} \right]^2 \\ \leq \; & C \sup_{l \leq \bar{n}} \left[E \left| \boldsymbol{e}_l^{\mathsf{T}} \boldsymbol{\Sigma}_p \boldsymbol{R}_{t-p} \right|^{4+\gamma} E \left| \varepsilon_{i_0,t} \right|^{4+\gamma} \right]^{\frac{1}{4+\gamma}} \sum_{j=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} \\ < \; & \infty. \end{aligned} \tag{28}$$

Similarly, we obtain the same boundary results of all covariance and variance of other terms. Back to the variance of $S_{\bar{n},n_0}$, there exit control functions for each term by the proof of (28). it follows from dominate convergence theorem that $\mathrm{Var}(S_{\bar{n},n_0}) = \boldsymbol{a}^{\mathsf{T}} \mathbf{U}_{i_0} \boldsymbol{a}$.

To prove the asymptotic normality of $S_{\bar{n},n_0}$ by Lévy continuity theorem, we employ the small-block and large-block arguments as follows. We partition the set $\{1, \cdots, n_0\}$ into $2k' + 1$ subsets with large blocks of size l', small blocks of size s' and the last remaining set of size $n_0 - k'l' - k's'$. Then we have

$$\begin{aligned} S_{\bar{n},n_0} \; = \; & \boldsymbol{a}_0^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(0)} + \boldsymbol{a}_0^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \eta_j^{(0)} + \boldsymbol{a}_0^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \zeta^{(0)} \\ + \; & \boldsymbol{a}_1^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(1)} + \boldsymbol{a}_1^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \eta_j^{(1)} + \boldsymbol{a}_1^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \zeta^{(1)} \\ + \; & \cdots \\ + \; & \boldsymbol{a}_p^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(p)} + \boldsymbol{a}_p^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \eta_j^{(p)} + \boldsymbol{a}_p^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \zeta^{(p)}, \end{aligned}$$

where

$$\xi_j^{(0)} = \sum_{t=(j-1)(l^{'}+s^{'})+1}^{jl^{'}+(j-1)s^{'}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}, \qquad \eta_j^{(0)} = \sum_{t=jl^{'}+(j-1)s^{'}+1}^{j(l^{'}+s^{'})} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t},$$

$$\zeta_j^{(0)} = \sum_{t=k^{'}(l^{'}+s^{'})+1}^{n_0} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}, \quad \cdots, \quad \xi_j^{(p)} = \sum_{t=(j-1)(l^{'}+s^{'})+1}^{jl^{'}+(j-1)s^{'}} \mathbf{I}_{S_{i_0}^{+}}^{\mathsf{T}} \mathbf{\Sigma}_0 \boldsymbol{R}_{t-p} \varepsilon_{i_0,t},$$

$$\eta_j^{(p)} = \sum_{t=jl^{'}+(j-1)s^{'}+1}^{j(l^{'}+s^{'})} \mathbf{I}_{S_{i_0}^{+}}^{\mathsf{T}} \mathbf{\Sigma}_0 \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}, \qquad \zeta_j^{(p)} = \sum_{t=k^{'}(l^{'}+s^{'})+1}^{n_0} \mathbf{I}_{S_{i_0}^{+}}^{\mathsf{T}} \mathbf{\Sigma}_0 \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}.$$

Put

$$l^{'} = \frac{\sqrt{n_0}}{\ln n_0}, \quad s^{'} = (\sqrt{n_0} \ln n_0)^{\chi}, \quad k^{'} = \frac{n_0}{l^{'}+s^{'}},$$

where $\frac{\gamma}{4+\gamma} \leq \chi < 1$. Also $\sum_{j=1}^{\infty} \alpha(j)^{\frac{\gamma}{4+\gamma}} < \infty$, we can easily obtain $\alpha(n_0) = o(n_0^{\frac{4+\gamma}{\gamma}})$ by using harmonic series, then it holds that $k^{'}\alpha(s^{'}) = o(1)$. By using the Hölder inequality successively for each $\boldsymbol{\theta}_t$, we have that

$$|\mathrm{Cov}(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+r})| \leq \alpha(r)^{1-\frac{2}{\delta}} [E|\boldsymbol{\theta_1}|^{\delta}]^{\frac{2}{\delta}} < O(r^{-1})[E|\theta|^{\delta}]^{\frac{2}{\delta}} = O(\frac{1}{r}),$$

where $\boldsymbol{\theta}_t = \boldsymbol{e}_j^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{x,t}$ and $\delta = \frac{\gamma+4}{2}$. It follows from Theorem 2.17 of Fan and Yao (2003) that

$$\frac{1}{n_0} \left| E \left(\sum_{j=1}^{k^{'}} \eta_j^{(i)} \right)^2 \right| = \frac{1}{n_0} E \left(\sum_{j=1}^{k^{'}} \eta_j^{(i)} \right)^2 \leq \frac{C k^{'} s^{'}}{n_0} \to 0$$

for $i = 0, \cdots, p$. Similarly, it holds that

$$\frac{1}{n_0} \left| E \left(\zeta^{(i)} \right)^2 \right| = \frac{1}{n_0} E \left(\zeta^{(i)} \right)^2 \leq \frac{C}{n_0} [n - k^{'}(s^{'} + l^{'})] \to 0$$

for $i = 0, \cdots, p$. Thus, $S_{\bar{n},n_0}$ can be rewritten as

$$S_{\bar{n},n_0} = \boldsymbol{a}_0^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k^{'}} \xi_j^{(0)} + \cdots + \boldsymbol{a}_p^{\mathsf{T}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k^{'}} \xi_j^{(p)} + o_p(1) \equiv T_{\bar{n},n_0} + o_p(1).$$

Similar to (27) and (28), w can calculate the variance of $T_{\bar{n},n_0}$ and it holds that

$$\mathrm{Var} \left(\frac{T_{\bar{n},n_0}}{\sqrt{\mathbf{a}^{\mathsf{T}} \mathbf{U}_{i_0} \mathbf{a}}} \right) \to 1.$$

Next, we just need to prove the asymptotic normality of $T_{\bar{n},n_0}$. We partition $T_{\bar{n},n_0}$ into two parts via truncation. Specifically, we define

$$\xi_j^{(0)L} = \sum_{t=(j-1)(l^{'}+s^{'})+1}^{jl^{'}+(j-1)s^{'}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \mathbf{I}_{\{\|\mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}\|_2 \leq L\}},$$

$$\xi_j^{(0)R} = \sum_{t=(j-1)(l^{'}+s^{'})+1}^{jl^{'}+(j-1)s^{'}} \mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t} \mathbf{I}_{\{\|\mathbf{I}_{S_{i_0}}^{\mathsf{T}} \mathbf{\Sigma}_p \boldsymbol{R}_{t-p} \varepsilon_{i_0,t}\|_2 > L\}}.$$

Similarly, we can define $\xi_j^{(i)L}$ and $\xi_j^{(i)R}$ for $i = 0, \cdots, p$. Then

$$\begin{aligned} T_{\bar{n},n_0} &= \left(\boldsymbol{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(0)L} + \cdots + \boldsymbol{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(p)L} + o_p(1) \right) \\ &+ \left(\boldsymbol{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(0)R} + \cdots + \boldsymbol{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(p)R} + o_p(1) \right) \\ &\equiv T_{\bar{n},n_0}^L + T_{\bar{n},n_0}^R. \end{aligned}$$

Similar to computing the variance of $S_{\bar{n},n_0}$ and $T_{\bar{n},n_0}$, note that

$$\mathrm{Var}\left(\frac{T_{\bar{n},n_0}^L}{\sigma_L} \right) \to 1, \qquad \mathrm{Var}\left(\frac{T_{\bar{n},n_0}^R}{\sigma_R} \right) \to 1,$$

where we denote σ_L as the asymptotic variance of $T_{\bar{n},n_0}^L$, and σ_R as the asymptotic variance of $T_{\bar{n},n_0}^R$. Define

$$M_{\bar{n},n_0} = \left| E \exp\left(\frac{itT_{\bar{n},n_0}}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - \exp\left(-\frac{t^2}{2} \right) \right|,$$

where $i = \sqrt{-1}$ now. We bound $M_{\bar{n},n_0}$ as follows

$$\begin{aligned} M_{\bar{n},n_0} &\le E \left| \exp\left(\frac{itT_{\bar{n},n_0}^L}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) \left[\exp\left(\frac{itT_{\bar{n},n_0}^R}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - 1 \right] \right| \\ &+ \left| E \exp\left(\frac{itT_{\bar{n},n_0}^L}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - \prod_{j=1}^{k'} E \exp\left[\frac{it\left(\mathbf{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(0)L} + \cdots + \mathbf{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(p)L} \right)}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right] \right| \\ &+ \left| \prod_{j=1}^{k'} E \exp\left[\frac{it\left(\mathbf{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(0)L} + \cdots + \mathbf{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(p)L} \right)}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right] - \exp\left(-\frac{t^2}{2} \frac{\sigma_L^2}{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}} \right) \right| \\ &+ \left| \exp\left(-\frac{t^2}{2} \frac{\sigma_L^2}{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}} \right) - \exp\left(-\frac{t^2}{2} \right) \right|. \end{aligned} \tag{29}$$

For the first term of (29), according to the property of characteristic function that $|E(e^{itX})| \le 1$, we have

$$E \left| \exp\left(\frac{itT_{\bar{n},n_0}^L}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) \left[\exp\left(\frac{itT_{\bar{n},n_0}^R}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - 1 \right] \right| \le E \left| \exp\left(\frac{itT_{\bar{n},n_0}^R}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - 1 \right| \to 0$$

when L is enough large. The last term of (29) may also have the same result by choosing large L as well. By proposition 2.6 of (Fan and Yao2003), the second term of (29) is bounded by $16(k^{'} - 1)\alpha(s^{'})$, which converges to 0. In addition, we can rewritten the third term of (29) as follows,

$$\begin{aligned} &\left| \prod_{j=1}^{k'} E \exp\left[\frac{it\left(\mathbf{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(0)L} + \cdots + \mathbf{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(p)L} \right)}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right] - \exp\left(-\frac{t^2}{2} \frac{\sigma_L^2}{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}} \right) \right| \\ = \; & E \exp\left[\frac{it \sum_{j=1}^{k'} \left(\mathbf{a}_0^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(0)L} + \cdots + \mathbf{a}_p^\mathsf{T} \frac{1}{\sqrt{n_0}} \xi_j^{(p)L} \right)}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right] - \exp\left(-\frac{t^2}{2} \frac{\sigma_L^2}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) \\ = \; & E \exp\left(\frac{itT_{\bar{n},n_0}^L}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right) - \exp\left(-\frac{t^2}{2} \frac{\sigma_L^2}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}}} \right). \end{aligned}$$

Now it suffices to prove $\frac{T^L_{\bar{n},n_0}}{\sigma_L} \to_d N(0,1)$. Under the condition that all ξ_j are mutually independent, we just need to prove

$$\mathbf{a}_i^\mathsf{T} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{k'} \xi_j^{(i)L} \to_d N(0, \nu_i^2)$$

for all $i = 0, \cdots, p$ by the additive property of normal distribution , where ν_i^2 is the asymptotic variance of the term above, and $\sigma_L^2 = \sum_{i=0}^{p} \nu_i^2$. It follows from Theorem 2.20 of Fan and Yao (2003) and (27) that

$$\frac{1}{l'}\mathrm{Var}(\xi_1^{(0)L}) \to \mathrm{Var}\left(\mathbf{I}_{S_{i_0}}^\mathsf{T} \boldsymbol{\Sigma}_p \boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) + 2\sum_{j=1}^{\infty} \mathrm{Cov}\left(\mathbf{I}_{S_{i_0}}^\mathsf{T} \boldsymbol{\Sigma}_p \boldsymbol{R}_0\varepsilon_{i_0,1}, \mathbf{I}_{S_{i_0}}^\mathsf{T} \boldsymbol{\Sigma}_p \boldsymbol{R}_j\varepsilon_{i_0,j+1}\right) < \infty,$$

then we obtain $\mathrm{Var}(\xi_1^{(0)L}) = O(l')$. Meanwhile,

$$\begin{aligned} E\left((\xi_1^{(0)L})^2 \cdot \mathbf{I}(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi)\right) &\leq \left|E\left((\xi_1^{(0)L})^2 \cdot \mathbf{I}(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi)\right)\right| \\ &\leq \sqrt{E[(\xi_1^{(0)L})^4]}\sqrt{E\left[\mathbf{I}^2(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi)\right]} \\ &= \sqrt{E[(\xi_1^{(0)L})^4]}P(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi), \end{aligned}$$

where $\phi = \frac{\nu_1}{\sqrt{\mathbf{a}^\mathsf{T}\mathbf{a}}}$ and $\epsilon > 0$ is arbitrarily positive number. For the probability term above, we have

$$P(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi) = P((\xi_1^{(0)L})^2 > n_0\epsilon^2\phi^2) \leq \frac{E[(\xi_1^{(0)L})^2]}{n_0\epsilon^2\phi^2} = \frac{O(l')}{n_0\epsilon^2\phi^2}.$$

Thus,

$$E\left((\xi_1^{(0)L})^2 \cdot \mathbf{I}(|\xi_1^{(0)L}| > \epsilon\sqrt{n_0}\phi)\right) \leq \sqrt{O(l'^2)}\frac{O(l')}{n_0\epsilon^2\phi^2} = O(\frac{l'^2}{n_0}).$$

Therefore, the Lindberg condition can be bounded as follows,

$$\lim_{k'\to\infty} \frac{1}{n_0\phi^2} \sum_{j=1}^{k'} E\left\{(\xi_j^{(0)L})^2 \cdot \mathbf{I}(|\xi_j^{(0)L}| > \epsilon\sqrt{n_0}\phi)\right\} \leq \lim_{k'\to\infty} \frac{1}{n_0\phi^2} k' O(\frac{l'^2}{n_0}) \to 0.$$

Similar proof for the rest items of the third term of $M_{\bar{n},n_0}$, we can obtain the similar conclusion, and summing up them. Hence the third term of (29) converges to 0. By the Lévy continuity theorem, it holds that

$$\sqrt{n_0}\mathbf{a}^\mathsf{T} \begin{pmatrix} \frac{1}{n_0}\sum_{t=p+1}^{n_0} (R_{j,t})_{j\in S_{i_0}} \boldsymbol{R}_{t-p}^\mathsf{T} \left(\frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} (R_{j,t-1})_{j\in S_{i_0}^+} \boldsymbol{R}_{t-p}^\mathsf{T} \left(\frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \\ \vdots \\ \frac{1}{n_0}\sum_{t=p+1}^{n_0} (R_{j,t-p})_{j\in S_{i_0}^+} \boldsymbol{R}_{t-p}^\mathsf{T} \left(\frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right) \end{pmatrix} / \sqrt{\mathbf{a}^\mathsf{T}\mathbf{U}_{i_0}\mathbf{a}} \to_d N(0,1).$$

Substituting $\mathbf{a}$ by $(\mathbf{U}_{i_0}^{-\frac{1}{2}})^\mathsf{T}\mathbf{a}$, then assertion (1) holds.

To prove assertion (2), let us look at one element of $\widehat{\mathbf{V}}_x^{\mathsf{T}}\widehat{\mathbf{V}}_x$ first. For some $j_1, j_2 \in S_{i_0}$,

$$\begin{aligned}
&\frac{1}{n_0}\sum_{t=p+1}^{n_0} \boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{R}_t\boldsymbol{R}_{t-p}^{\mathsf{T}}\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\boldsymbol{R}_t^{\mathsf{T}}\boldsymbol{e}_{j_2} \\
= &\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{R}_t\boldsymbol{R}_{t-p}^{\mathsf{T}} - \boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{\Sigma}_p\right)\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\boldsymbol{R}_t^{\mathsf{T}}\boldsymbol{e}_{j_2} - \boldsymbol{\Sigma}_p^{\mathsf{T}}\boldsymbol{e}_{j_2}\right) \\
+ &\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{\Sigma}_p\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\boldsymbol{R}_t^{\mathsf{T}}\boldsymbol{e}_{j_2} - \boldsymbol{\Sigma}_p^{\mathsf{T}}\boldsymbol{e}_{j_2}\right) \\
+ &\left(\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{R}_t\boldsymbol{R}_{t-p}^{\mathsf{T}} - \boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{\Sigma}_p\right)\boldsymbol{\Sigma}_p^{\mathsf{T}}\boldsymbol{e}_{j_2} \\
+ &\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_p^{\mathsf{T}}\boldsymbol{e}_{j_2}. \qquad (30)
\end{aligned}$$

Using the same arguments as (26), the first term is $O_p(\frac{\bar{n}}{n_0})$, the second and third terms are of order $O_p(\frac{1}{\sqrt{n_0}})$. Hence given $\bar{n} = o(n_0)$, it holds that

$$\frac{\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{R}_t\boldsymbol{R}_{t-p}^{\mathsf{T}}\frac{1}{n_0}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\boldsymbol{R}_t^{\mathsf{T}}\boldsymbol{e}_{j_2}}{\boldsymbol{e}_{j_1}^{\mathsf{T}}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_p^{\mathsf{T}}\boldsymbol{e}_{j_2}} \to 1.$$

Applying the same arguments to the other elements of $\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0}$, it holds that

$$\mathbf{K}_{i_0}(\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0})^{-1} \to_p \mathbf{I}_{\tau_{i_0}}.$$

When $k_0 = o\left(C_{n_0}^{-1}n_0/\log(\bar{n}\vee n_0)\right)$, we have $\lambda_{\min}(\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0}) \geq c$ with probability tending to 1. By (5) and (26),

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 &= \left\|\frac{1}{n_0}(\widehat{\mathbf{V}}_i^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0})^{-1}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_2 \\
&\leq \frac{1}{c}\left\|\frac{1}{n_0}\widehat{\mathbf{V}}_x^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_2 \\
&= \frac{1}{c}\sqrt{\sum_{j=1}^{\tau_{i_0}}\left[O_p(\frac{\bar{n}}{n_0}) + O_p(\frac{1}{\sqrt{n_0}})\right]^2} \\
&= O_p(\sqrt{\frac{k_0}{n_0}}), \qquad i_0 = 1, \cdots, \bar{n}.
\end{aligned}$$

If k_0 is not fixed in the part (ii) of Theorem 3, we obtain

$$\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_2 \leq \frac{1}{c}\left\|\frac{1}{n_0}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_2 = O_p(\frac{\sqrt{k_0}\bar{n}}{n_0}), \qquad i_0 = 1, \cdots, \bar{n}.$$

The required asymptotic result for a fixed k_0 follows from the above result directly. This completes the proof of Theorem 3.

Proof of Theorem 4.

By Theorem 3, it holds that

$$\begin{aligned}
\left\|\widehat{\boldsymbol{\beta}}_{i_0} - \boldsymbol{\beta}_{i_0}\right\|_1 &= \left\|\frac{1}{n_0}(\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\widehat{\mathbf{V}}_{i_0})^{-1}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_1 \\
&\leq C\left\|\frac{1}{n_0}\widehat{\mathbf{V}}_{i_0}^{\mathsf{T}}\sum_{t=p+1}^{n_0}\boldsymbol{R}_{t-p}\varepsilon_{i_0,t}\right\|_1 \\
&= C\sum_{j=1}^{\tau_{i_0}}\left|O_p(\frac{\bar{n}}{n_0}) + O_p(\frac{1}{\sqrt{n_0}})\right|
\end{aligned}$$

for $i_0 = 1, \cdots, \bar{n}$. The required asymptotic result follows from the above result easily.

References

[1] Biau, G., Cadre, B., 2004. Nonparametric Spatial Prediction. Statistical Inference for Stochastic Process. 7(3), 327 – 349.

[2] Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion Ltd., London.

[3] Cressie N.A.C., 1993. Statistics for Spatial Data[M]. New York: Wiley.

[4] Deutsch, S.J., Ramos, J.A., 1986. Space-time modeling of vector hydrologic sequences. JAWRA Journal of the American Water Resources Association. 22(6), 967-981.

[5] Dou, B.J., Parrella, M.L., Yao, Q.W., 2016. Generalized yule – walker estimation for spatio-temporal models with unknown diagonal coefficients. Journal of Econometrics. 194(2), 369 – 382.

[6] Fan, J.Q., Yao, Q.W., 2003. Nonlinear Time Series Analysis: Nonparametric and Parametric Methods. Springer, New York.

[7] Fu, H., Yuan, G., Özkan, K., Johansson, L. S., Sondergaard, M., Lauridsen, T. L., Jeppesen, E. 2020. Seasonal and long-term trends in the spatial heterogeneity of lake phytoplankton communities over two decades of restoration and climate change. Science of The Total Environment, 141106.

[8] Fu, Z.Y., Li, R., 2020. The contributions of socioeconomic indicators to global PM 2.5 based on the hybrid method of spatial econometric model and geographical and temporal weighted regression. Science of the Total Environment. 703.

[9] Gao, Z.X., Ma, Y.Y., Wang, H.S., Yao, Q.W., 2019. Banded spatio-temporal autoregressions. Journal of Econometrics. 208(1), 211-230.

[10] Guo, S.J., Wang, Y.Z., Yao, Q.W., 2016. High dimensional and banded vector autoregressions. Biometrika. 103(4), 889 – 903.

[11] Hallin, M., Lu, Z., and Tran, L.T., 2004. Local Linear Spatial Regression. The Annals of Statistics. 32, 2469 – 2500.

[12] Kelejian, H.H., Prucha, I.R., 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of Econometrics. 157(1), 53 – 67.

[13] Lütkepohl, H., 2007. New Introduction to Multiple Time Series Analysis. Springer, New York.

[14] Lam, C., Yao, Q.W., 2012. Factor modeling for high-dimensional time series: inference for the number of factors. The Annals of Statistics. 40(2), 694 – 726.

[15] Lin, X., Lee, L.F., 2010. GMM estimation of spatial autoregressive models with unknown heteroskedasticity. Journal of Econometrics. 157(1), 34-52.

[16] Linton, O., Xiao, Z.J., 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. Econometric Theory. 23(3), 371-413.

[17] Masry, E., 1986. Recursive probability density estimation for weakly dependent stationary processes. IEEE Transactions on Information Theory. 32(2):254-267.

[18] Moran, P.A.P., 1948. The Interpretation of Statistical Maps. Journal of the Royal Statistical Society: Series B (Methodological). 10(2), 243 – 251.

[19] Pham, T.D., Tran, L.T., 1985. Some mixing properties of time series models. Stochastic Processes and Their Applications. 19(2), 297 – 303.

[20] Robinson, P.M., 2010. Efficient estimation of the semiparametric spatial autoregressive model. Journal of Econometrics. 157(1), 6 – 17.

[21] Szummer, M., Picard, R.W., 1996. Temporal Texture Modeling. Cambridge, MA: MIT Media Laboratory, Manuscript.

[22] Tran, L.T., 1990. Kernel density estimation on random fields. Journal of Multivariate Analysis. 34(1), 37 – 53.

[23] Wang, H.S., Li, B., Leng, L.C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. Journal of the Royal Statal Society: Series B (Statal Methodology). 71(3), 671-683.

[24] Wang, B., Sun, Y.F., Wang, Z.H., 2018. Agglomeration effect of CO2 emissions and emissions reduction effect of technology: A spatial econometric perspective based on China's province-level data. Journal of Cleaner Production.

[25] Wang, H.X., Wang, J.D., Huang B., 2012. Prediction for spatio-temporal models with autoregression in errors. Journal of Nonparametric Statistics. 24(1), 217-244.

[26] Yu, J.H., Jong, R.D., Lee, L.f., 2012. Estimation for spatial dynamic panel data with fixed effects: the case of spatial cointegration. Journal of Econometrics. 167(1), 16 – 37.