



High-Dimensional Spatial Autoregression with Latent Factors by Diversified Projections

Jiaxin Shi, Xuening Zhu, Jing Zhou, Baichen Yu & Hansheng Wang

To cite this article: Jiaxin Shi, Xuening Zhu, Jing Zhou, Baichen Yu & Hansheng Wang (08 Dec 2025): High-Dimensional Spatial Autoregression with Latent Factors by Diversified Projections, Journal of the American Statistical Association, DOI: [10.1080/01621459.2025.2562644](https://doi.org/10.1080/01621459.2025.2562644)

To link to this article: <https://doi.org/10.1080/01621459.2025.2562644>



View supplementary material [↗](#)



Published online: 08 Dec 2025.



Submit your article to this journal [↗](#)



Article views: 430



View related articles [↗](#)



View Crossmark data [↗](#)



High-Dimensional Spatial Autoregression with Latent Factors by Diversified Projections

Jiaxin Shi^a, Xuening Zhu^{b,c,d}, Jing Zhou^e, Baichen Yu^a, and Hansheng Wang^a

^aGuanghua School of Management, Peking University, Beijing, China; ^bSchool of Management, Fudan University, Shanghai, China; ^cSchool of Data Science, Fudan University, Shanghai, China; ^dMOE Laboratory for National Development and Intelligent Governance, Fudan University, China; ^eCenter for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

ABSTRACT

We study one particular type of multivariate spatial autoregression (MSAR) model with diverging dimensions in both responses and covariates. This makes the usual MSAR models no longer applicable due to the high computational cost. To address this issue, we propose a factor-augmented spatial autoregression (FSAR) model. FSAR is a special case of MSAR but with a novel factor structure imposed on the high-dimensional random error vector. The latent factors of FSAR are assumed to be of a fixed dimension. Therefore, they can be estimated consistently by the diversified projections method, as long as the dimension of the multivariate response is diverging. Once the fixed-dimensional latent factors are consistently estimated, they are then fed back into the original SAR model and serve as exogenous covariates. This leads to a novel FSAR model. Thereafter, different components of the high-dimensional response can be modeled separately. To handle the high-dimensional feature, a smoothly clipped absolute deviation (SCAD) type penalized estimator is developed for each response component. We show theoretically that the resulting SCAD estimator is uniformly selection consistent, as long as the tuning parameter is selected appropriately. For practical selection of the tuning parameter, a novel BIC method is developed. Extensive numerical studies are conducted to demonstrate the finite sample performance of the proposed method. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY

Received August 2024
Accepted August 2025

KEYWORDS

Diversified projections;
Factor-augmented maximum
likelihood estimator;
High-dimensional spatial
data; Latent factor model

1. Introduction

Spatial data are frequently encountered in various statistical and econometric applications (Fujita, Krugman, and Venables 2001; Yin et al. 2022; Zhou et al. 2023). These applications include, but are not limited to, environmental analysis (Zhou et al. 2023), geographical science (Yin et al. 2022), political economics (Yu, Zhou, and Zhu 2016), and many others. To model the spatial dependence among multiple subjects/nodes, a variety of spatial autoregressive (SAR) models have been developed and extensively studied (Kelejian and Prucha 1998; Lee and Yu 2010; Huang et al. 2019). A number of estimation methods have been proposed, and their corresponding statistical properties have been carefully studied. Those estimation methods include, the quasi-maximum likelihood estimation (QMLE) method (Lee 2004), the generalized method of moments (GMM) (Lee 2007), the least squares estimation (LSE) methods (Huang et al. 2019), and many others (Su 2012).

It is remarkable that those classical SAR models are applicable only to datasets with univariate responses. In other words, only a univariate response is collected for every subject/node in a spatial/network dataset. However, in real-world applications, multivariate responses are frequently encountered. This leads to various multivariate spatial autoregressive (MSAR) models (Yang and Lee 2017; Zhu et al. 2020). Moreover, real network

datasets with high-dimensional multivariate responses are becoming increasingly available (Zhang, Xu, and Zhu 2022; Chen et al. 2025). Consider, for example, a regional economic dataset from China. The full dataset contains a total of 287 cities and 112 macroeconomic indicators. The objective here is to study the spatial spillover effects of regional economics, which is a problem of great importance for understanding spatial economic dynamics and regional economic growth in macroeconomic research (Anselin 1988; Zhou et al. 2023). For this dataset, each region can be treated as a node, which is spatially connected with others. This makes the SAR model a natural choice in empirical economics literature (Blasques et al. 2016; De Paula, Rasul, and Souza 2025). In addition to the spatial structure, we also observe a large number of economic indicators for each region (i.e., node). This leads to a high-dimensional response vector for each region. As a consequence, the MSAR models of Yang and Lee (2017) and Zhu et al. (2020) are difficult to apply directly. This is mainly because the associated computational cost becomes extremely expensive if the response dimension is relatively high. This interesting dataset motivates us to develop a novel method, which is able to model spatial dependence for datasets with high-dimensional responses.

To this end, we propose a factor-augmented SAR approach. This method assumes a standard SAR model for each compo-

nent of the high-dimensional response. By doing so, the spatial dependence structure can be flexibly modeled for different response components in a parallel way. This leads to a high-dimensional error vector for each node. To analyze the high-dimensional data, various factor modeling techniques have been developed (Fan, Fan, and Lv 2008; Bai 2012; Lam and Yao 2012; Fan and Liao 2022). We are then inspired to impose a factor model on this high-dimensional error vector. This leads to a new type of SAR model with a factor-augmented structure. For convenience, we refer to this as a factor-augmented spatial autoregressive (FSAR) model. It is worth noting that the FSAR model is related to the dynamic SAR model in the existing literature (Bai and Li 2021). However, there are two critical differences. First, FSAR is a static MSAR model without a dynamic panel structure over time. Second, FSAR allows different spatial effects for different responses. By assuming that the factor dimension is fixed as d , we obtain a highly simplified model structure with a total of only $(d + q + 2)p$ parameters, if the exogenous covariates are of dimension q . In contrast, a traditional MSAR model in this case (Yang and Lee 2017; Zhu et al. 2020) should consume a total of $(3p^2/2 + pq)$ parameters. Moreover, it is remarkable that the type of dependence captured by MSAR and FSAR models are different. The MSAR model is good at capturing weak dependence, which refers to the type of influence due to local network neighbors. In contrast, our FSAR model is good at capturing strong dependence, which reflects the type of the influence due to the global cross-sectional dependence. Therefore, the applications of MSAR and FSAR models are not the same.

To practically estimate the FSAR model, a three-step estimation procedure is developed. In the first step, the standard QMLE method (Lee 2004) is applied to each component of the high-dimensional response. This yields to a consistent initial estimator for each componentwise SAR model. Accordingly, the high-dimensional error vector can be consistently differentiated for each node. In the second step, the diversified projections method of Fan and Liao (2022) is applied to the estimated error vectors for factor estimation. By doing so, the latent factors can be consistently estimated up to an affine transformation. In the last step, we treat the estimated factors as exogenous covariates. Then, the FSAR model can be estimated for each response component in a fully parallel way (Lee 2004; Lee and Yu 2010). This leads to the final estimators for the spatial correlation parameters. Under appropriate regularity conditions, we show theoretically that the resulting estimator is \sqrt{n} -consistent and asymptotically normal. To handle high-dimensional exogenous covariates, a smoothly clipped absolute deviation (SCAD) penalized estimator (Fan and Li 2001) is developed for the FSAR model, and a novel BIC method is developed for tuning parameter selection (Wang, Li, and Tsai 2007; Chen and Chen 2008; Wang, Li, and Leng 2009). We show theoretically that the resulting estimator is uniformly selection consistent for every response component.

The rest of the article is organized as follows. Section 2 develops the FSAR model. The estimation methods and the associated asymptotic theory are also included. The numerical studies are presented in Section 3, which includes both extensive simulation experiments and a real data example. Finally, Section 4 concludes the article with a brief discussion. All technical proofs are left to the Appendix.

2. Methodology

2.1. The Model Setup

Consider a large-scale network with n nodes indexed by $1 \leq i \leq n$. Define an adjacency matrix of the network as $A = (a_{i_1 i_2}) \in \mathbb{R}^{n \times n}$, where $a_{i_1 i_2} = 1$ if the node i_1 is connected to the node i_2 and $a_{i_1 i_2} = 0$ otherwise. Following the existing literature (Lee 2004; Zhu et al. 2020), we set $a_{ii} = 0$ for every $1 \leq i \leq n$. Next, define a spatial weight matrix as $W = (w_{i_1 i_2}) \in \mathbb{R}^{n \times n}$ with $w_{i_1 i_2} = a_{i_1 i_2}/n_{i_1}$ and $n_{i_1} = \sum_{i_2=1}^n a_{i_1 i_2}$ so that each row of the weight matrix W sums up to one. For those zero-degree nodes (i.e., $\sum_{i_2=1}^n a_{i_1 i_2} = 0$), we set $w_{i_1 i_2} = 0$ for $1 \leq i_2 \leq n$ so that the useful covariate information contained in those nodes can be maintained. Next, for each node i , we observe a p -dimensional response vector $Y_i = (Y_{ij}) \in \mathbb{R}^p$ with $p \rightarrow \infty$ as $n \rightarrow \infty$ and a q -dimensional exogenous covariate vector as $X_i = (X_{im}) \in \mathbb{R}^q$ (Lee 2004; Lee and Yu 2010; Huang et al. 2021). Write $\mathbb{Y}_j = (Y_{ij}) \in \mathbb{R}^n$ as the response vector for the j th component. Accordingly, write $\mathbb{X} = (X_1, \dots, X_n)^\top \in \mathbb{R}^{n \times q}$ as the covariate matrix. Each component of \mathbb{Y}_j is expected to be spatially correlated with others through A . Therefore, we assume for \mathbb{Y}_j a standard spatial autoregressive (SAR) model as

$$\mathbb{Y}_j = \rho_j W \mathbb{Y}_j + \mathbb{X} \beta_j + \mathcal{E}_j, \quad (2.1)$$

where $\rho_j \in \mathbb{R}$ is the spatial correlation, $\beta_j = (\beta_{jm}) \in \mathbb{R}^q$ is the coefficient vector, and $\mathcal{E}_j = (\varepsilon_{ij}) \in \mathbb{R}^n$ is the error vector.

Let ρ_j^* and $\beta_j^* = (\beta_{jm}^*)$ be the true values of ρ_j and β_j , respectively. In this work, we allow q to diverge in the sense that $q \rightarrow \infty$ as $n \rightarrow \infty$. In this case, we should expect that a large number of covariates are redundant for any given response \mathbb{Y}_j (Tibshirani 1996; Fan and Li 2001; Huang et al. 2021). To reflect this phenomenon, define for every response \mathbb{Y}_j a true model set $\mathcal{S}_{(j),T} = \{1 \leq k \leq q : \beta_{jk}^* \neq 0\}$ with size $s_j = |\mathcal{S}_{(j),T}|$. Define $\mathcal{S}_{\text{True}} = \bigcup_{1 \leq j \leq p} \mathcal{S}_{(j),T}$. In this work, we assume the maximum size of the true model for every response component is upper bounded by a fixed number $m > 0$ (i.e., $\max_{1 \leq j \leq p} s_j \leq m$). However, we allow the total number of relevant covariates (i.e., $|\mathcal{S}_{\text{True}}|$) to diverge as $n \rightarrow \infty$. This allows a diverging amount of information to be used. Next, let $\mathbb{X}_k = (X_{ik}) \in \mathbb{R}^n$ be the k th column of \mathbb{X} . Write $\mathbb{X}_{(j)} = (\mathbb{X}_k : k \in \mathcal{S}_{(j),T}) \in \mathbb{R}^{n \times s_j}$ as the submatrix of \mathbb{X} corresponding to $\mathcal{S}_{(j),T}$. Similarly, define $\beta_{(j)} = (\beta_{jk} : k \in \mathcal{S}_{(j),T}) \in \mathbb{R}^{s_j}$. Then, model (2.1) becomes

$$\mathbb{Y}_j = \rho_j W \mathbb{Y}_j + \mathbb{X}_{(j)} \beta_{(j)} + \mathcal{E}_j. \quad (2.2)$$

Assume that $\mathcal{S}_{(j),T}$ s are already given at this moment. In practice, $\mathcal{S}_{(j),T}$ s are typically unknown. Therefore, they have to be consistently estimated based on the observed data. This is an important issue to be studied in Section 2.5.

Write $\varepsilon_i = (\varepsilon_{ij}) \in \mathbb{R}^p$ as the error vector associated with the i th node. Then, how to model the stochastic behavior of ε_i with a high dimension p becomes a problem of great interest. To address this issue, we follow the ideas of Fan, Fan, and Lv (2008) and Wang (2012) and assume a factor model as

$$\varepsilon_i = B Z_i + \omega_i, \quad (2.3)$$

where $Z_i = (Z_{ik}) \in \mathbb{R}^d$ is a d -dimensional latent factor for the i th node, $B = (b_{jk}) \in \mathbb{R}^{p \times d}$ is the loading matrix, and $\omega_i = (\omega_{ij}) \in \mathbb{R}^p$ represents the information contained in

ε_i but missed by Z_i . We assume that the factor dimension d is a fixed number, consistent with the existing literature (Bai 2012; Lam and Yao 2012; Fan and Liao 2022), and also with our empirical example, which is to be analyzed in Section 3.3. We assume that Z_i and ω_{ij} s are mutually independent with mean 0. Write $\text{cov}(\varepsilon_i) = \Sigma_\varepsilon = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p \times p}$, $\Sigma_\omega = \text{cov}(\omega_i) = (\tau_{j_1 j_2}) \in \mathbb{R}^{p \times p}$, and $\Sigma_Z = \text{cov}(Z_i) \in \mathbb{R}^{d \times d}$. Accordingly, the true parameters are denoted as $B^* = (b_{jk}^*)$, $\Sigma_\varepsilon^* = (\sigma_{j_1 j_2}^*)$, $\Sigma_\omega^* = (\tau_{j_1 j_2}^*)$, and Σ_Z^* . It then follows that $\Sigma_\varepsilon^* = B^* \Sigma_Z^* B^{*\top} + \Sigma_\omega^*$. For model identification, we assume that $\Sigma_Z^* = I_d$, which stands for a d -dimensional identity matrix. Otherwise, we can always re-define $Z_i := \Sigma_Z^{*-1/2} Z_i$ and $B := B \Sigma_Z^{*1/2}$ so that model (2.3) remains valid but with $\text{cov}(Z_i) = I_d$. Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be the largest and smallest eigenvalues of an arbitrary symmetric matrix A , respectively. Moreover, we assume that Σ_ω^* is a positive definite matrix. Notably, we do not require Σ_ω^* to be of a diagonal structure. The only constraint imposed on Σ_ω^* is that its eigenvalues are well bounded away from 0 and infinity as $p \rightarrow \infty$ (Wang, Li, and Leng 2009; Wang 2012; Fan and Liao 2022).

2.2. Componentwise Maximum Likelihood Estimators

We next consider how to estimate the model parameters. Here we temporarily assume that the true model sets $S_{j,T}$ s are given. Thus, the following estimators obtained in Sections 2.2–2.4 are the very ideal estimators, which are often referred as the oracle estimators in the literature (Donoho and Johnstone 1994; Fan and Li 2001). Unfortunately, those true model sets $S_{j,T}$ s are typically unknown in practice. Therefore, these oracle estimators cannot be practically computed and the true model sets have to be empirically estimated. That is the reason why we have developed in Section 2.5 a SCAD-penalized estimation method for a uniformly consistent selection of the true model sets $S_{j,T}$ s. This should be done in prior to applying the three-step estimation procedure as shown in Sections 2.2–2.4.

In this study, we focus on the QMLE method due to its theoretical importance. However, the method to be developed can readily be applied to other estimation methods without additional difficulty. We then apply the QMLE method to each response component j to obtain consistent initial estimators for the interested parameters (i.e., ρ , $\beta_{(j)}$, and σ_{jj}). For convenience, we refer to these as componentwise maximum likelihood estimators (CMLE). Specifically, under the following technical conditions C1–C2, we have $E(\mathcal{E}_j) = 0$ and $\text{cov}(\mathcal{E}_j) = \sigma_{jj}^* I_n$, where $\sigma_{jj}^* = \text{var}(\varepsilon_{ij}) = \|b_j^*\|^2 + \tau_{jj}^*$ and b_j^* is the j th column of B^* . Write $S_j = I_n - \rho_j W$. This leads to a reduced form as $\mathbb{Y}_j = S_j^{*-1} (\mathbb{X}_{(j)} \beta_{(j)}^* + \mathcal{E}_j)$ with $S_j^* = I_n - \rho_j^* W$. To ensure that S_j is invertible, we follow Lee (2004) and assume $|\rho_j| < 1$ for every $1 \leq j \leq p$. Define $\theta_j = (\rho_j, \beta_{(j)}^\top, \sigma_{jj})^\top \in \mathbb{R}^{j+2}$. Then, the log-likelihood function for CMLE is given by

$$\begin{aligned} \mathcal{L}_{\text{cmle}}^{(j)}(\theta_j) = & -\frac{n}{2} \log \sigma_{jj} + \log |S_j| \\ & - \frac{1}{2\sigma_{jj}} \left(S_j \mathbb{Y}_j - \mathbb{X}_{(j)} \beta_{(j)} \right)^\top \left(S_j \mathbb{Y}_j - \mathbb{X}_{(j)} \beta_{(j)} \right), \end{aligned} \quad (2.4)$$

where some irrelevant constants are ignored. The CMLE for θ_j can then be obtained by maximizing (2.4) as $\hat{\theta}_{j,\text{cmle}} =$

$(\hat{\rho}_{j,\text{cmle}}, \hat{\beta}_{(j),\text{cmle}}^\top, \hat{\sigma}_{jj,\text{cmle}})^\top = \arg\max_{\theta} \mathcal{L}_{\text{cmle}}^{(j)}(\theta)$. The asymptotic properties of $\hat{\theta}_{j,\text{cmle}}$ have been well studied in the existing literature (Lee 2004; Lee and Yu 2010; Yang and Lee 2017).

By Theorem 3.1 of Lee (2004), the componentwise estimator $\hat{\theta}_{j,\text{cmle}}$ is \sqrt{n} -consistent for every $1 \leq j \leq p$. However, to the best of our knowledge, it seems that no uniform convergence result has been established when $p \rightarrow \infty$. Nevertheless, a uniform convergence result about $\hat{\theta}_{j,\text{cmle}}$ s is critically important for our subsequent theory development. To this end, define the ℓ_q -norm of an arbitrary vector $v = (v_j) \in \mathbb{R}^p$ as $\|v\|_q = (\sum_{j=1}^p |v_j|^q)^{1/q}$ for any $q > 0$. For convenience, we omit the subscript q when $q = 2$. Additionally, denote a sub-Weibull distribution of order α as sub-Weibull(α). Let $U \in \mathbb{R}$ be an arbitrary random variable. Define its sub-Weibull(α) norm as $\|U\|_{\psi_\alpha} = \inf \{t > 0 : E \exp(|U|^\alpha / t^\alpha) \leq 2\}$. Then, the following technical conditions are necessarily needed.

- (C1) (SUB-WEIBULL DISTRIBUTION) Assume that both Z_{ik} and ω_{ij} independently follow sub-Weibull(α) distributions with $\alpha \in (0, 2]$ for every $1 \leq k \leq d$ and $1 \leq j \leq p$, and are independent of X_i . Furthermore, assume that there exists a positive and fixed constant C_{sw} such that $\max_k \|Z_{ik}\|_{\psi_\alpha} \leq C_{\text{sw}}$ and $\max_j \|\omega_{ij}\|_{\psi_\alpha} \leq C_{\text{sw}}$.
- (C2) (LOADING MATRIX) Assume that the loading matrix $B^* = (b_{jk}^*) \in \mathbb{R}^{p \times d}$ is fixed and there exists a fixed constant $C_B > 0$ such that $\max_{j,k} |b_{jk}^*| \leq C_B$.
- (C3) (BOUNDED PARAMETERS) Assume that there exist some positive constants $0 < \beta_{\min} < C_{\beta \max} < \infty$ and $0 < \tau_{\min} \leq \tau_{\max} < 1$ such that (1) $\max_j \|\beta_j^*\| \leq C_{\beta \max}$; (2) $\beta_{\min} \leq \min_{j,k \in S_{(j),T}} |\beta_{jk}^*|$; and (3) $\tau_{\min}^2 \leq \min_j \tau_{jj}^* \leq \max_j \tau_{jj}^* \leq \tau_{\max}^2$.

The sub-Weibull distribution assumption imposed by Condition C1 allows for heavier tails and is thus weaker than the popularly used sub-Gaussian assumption in high-dimensional literature (Wainwright 2019). However, it is stronger than the moment conditions widely used in the classical SAR literature (Lee 2004; Zhu et al. 2020). Condition C1 is necessary in our setting since we are dealing with a problem with a diverging dimension. Consequently, appropriate uniform convergence results are inevitably needed; see for example the uniform consistent result of Theorem 1. This also explains why similar tail conditions like C1 have been seldom used in the classical SAR literature of a fixed dimension (Lee 2004; Zhu et al. 2020). However, they are extensively used in the high-dimensional literature (Wainwright 2019). Moreover, Condition C1 assumes that both Z_{ik} and ω_{ij} are independent of the exogenous X_i . As a consequence, the regression effect term $\mathbb{X}_{(j)} \beta_{(j)}$ for every $1 \leq j \leq p$ can be interpreted in the same way as the usual SAR model (Yang and Lee 2017).

Condition C2 requires that the true value of the factor loading matrix B^* to be elementwise uniformly bounded (Fan, Fan, and Lv 2008; Bai 2012). Condition C3 assumes that: (a) $\|\beta_j^*\|$ s are uniformly upper bounded, (b) the minimum of nonzero $|\beta_{jk}^*|$ s are uniformly lower bounded away from 0, and (c) the error variance τ_{jj}^* s are uniformly bounded away from both 0 and 1 as $p \rightarrow \infty$. Similar conditions have been used by Fan and Lv (2011) and Wang (2012).

- (C4) (DIVERGING RESPONSE DIMENSION) Assume that (1) $\sqrt{n}/p \rightarrow 0$ as $n \rightarrow \infty$ and (2) $\log p = O(n^{\alpha\gamma})$, where $\alpha \in (0, 2]$ is the sub-Weibull parameter specified in Condition C1 and $\gamma \in (0, 1/4)$ is some fixed constant.
- (C5) (DIVERGING FEATURE DIMENSION) Assume that (1) $(q \log q)^{1/\alpha} / n^{1/2-2\gamma} \rightarrow 0$ as $n \rightarrow \infty$, and (2) $q / \{(\log q)^{2/\alpha} \log n\} \rightarrow 0$ as $n \rightarrow \infty$, where $\alpha \in (0, 2]$ is the sub-Weibull parameter specified in Condition C1 and $\gamma \in (0, 1/4)$ is defined in Condition C4.

The first part of Condition C4 requires the response dimension p to be sufficiently large so that the latent factors can be estimated consistently (Fan and Liao 2022). The second part of Condition C4 allows the response dimension p to diverge at an exponentially fast rate (Fan, Fan, and Lv 2008; Fan and Liao 2022). By Condition C5, we require that the diverging rate of the feature dimension q cannot be too fast (Fan, Fan, and Lv 2008; Wang, Li, and Leng 2009; Cho and Qu 2013).

Next, let $\|A\|_1 = \max_{1 \leq j \leq n} |\sum_{i=1}^m a_{ij}|$ and $\|A\|_\infty = \max_{1 \leq i \leq m} |\sum_{j=1}^n a_{ij}|$ stand for the ℓ_1 -norm (i.e., the maximum absolute column sum) and ∞ -norm (i.e., the maximum absolute row sum) of an arbitrary matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, respectively. In this article, we use “ \rightarrow_d ” to denote “convergence in distribution” and “ \rightarrow_p ” to denote “convergence in probability”. Write $G_j = WS_j^{-1}$ with $S_j = I_n - \rho_j W$. Define $\tilde{\mathbb{X}}_\beta^* = (G_1 \mathbb{X} \beta_1^*, \dots, G_p \mathbb{X} \beta_p^*) \in \mathbb{R}^{n \times p}$ and $\tilde{\mathbb{Z}}_b^* = (G_1 \mathbb{Z} b_1^*, \dots, G_p \mathbb{Z} b_p^*) \in \mathbb{R}^{n \times p}$.

- (C6) (NETWORK MATRIX) Assume that there exists a sufficiently large but fixed constant $C_W > 0$ such that $\|W\|_1 \leq C_W$ and $\max_j (\|S_j^{-1}\|_1, \|S_j^{-1}\|_\infty) \leq C_W$ uniformly in $\rho_j \in [-\rho_{\max}, \rho_{\max}]$ for some fixed constant $\rho_{\max} \in (0, 1)$.
- (C7) (LAW OF LARGE NUMBERS) Assume that some positive and fixed constants $\kappa_{Gj1}, \kappa_{Gj2}, \kappa_{Gj3}, \kappa_{GGj1}, \kappa_{GGj2}$, and κ_{Gjd} , such that (1) $\max_j |\text{tr}(G_j^v)/n - \kappa_{Gjv}| = o(1)$ for $v = 1, 2, 3$; (2) $\max_j |\text{tr}\{(G_j^\top G_j)^v\}/n - \kappa_{GGjv}| = o(1)$ for $v = 1, 2$; and (3) $\max_j |\text{tr}\{\text{diag}^2(G_j)\}/n - \kappa_{Gjd}| = o(1)$ as $n \rightarrow \infty$ uniformly in $\rho_j \in [-\rho_{\max}, \rho_{\max}]$ for the same ρ_{\max} given in Condition C6.
- (C8) (IDENTIFICATION) Assume a fixed and nonsingular matrix $\Sigma_{\mathbb{X}\mathbb{Z}}^* \in \mathbb{R}^{(q+d+2p) \times (q+d+2p)}$ such that (1) $\|(\mathbb{X}, \tilde{\mathbb{X}}_\beta^*, \mathbb{Z}, \tilde{\mathbb{Z}}_b^*)^\top (\mathbb{X}, \tilde{\mathbb{X}}_\beta^*, \mathbb{Z}, \tilde{\mathbb{Z}}_b^*)/n - \Sigma_{\mathbb{X}\mathbb{Z}}^*\| = o_p(1)$, and (2) $\nu_{\min} \leq \lambda_{\min}(\Sigma_{\mathbb{X}\mathbb{Z}}^*) \leq \lambda_{\max}(\Sigma_{\mathbb{X}\mathbb{Z}}^*) \leq \nu_{\max}$ for some fixed constants $\nu_{\min} > 0$ and $\nu_{\max} > 0$.

Condition C6 assumes that both W and S_j^{-1} are uniformly bounded in both the column and row sums as $n \rightarrow \infty$. Condition C7 is a set of Law of Large Numbers type conditions. Condition C8 is a sufficient identification condition for θ_j^* and b_j^* with $1 \leq j \leq p$. All those conditions are fairly standard in the literature of spatial autoregression (Lee 2004; Lee, Liu, and Lin 2010; Yang and Lee 2017).

Then, the uniform convergence of $\hat{\theta}_{j,\text{cmle}}$ is given in Theorem 1, which is proved in Appendix A.1. By Theorem 1, we know that $\hat{\theta}_{j,\text{cmle}}$ is uniformly consistent for θ_j^* over $1 \leq j \leq p$. The uniform convergence rate is slightly slower than the standard rate of $1/\sqrt{n}$ by a factor $(\log p)^{1/\alpha}$. This is the price paid for uniform convergence (Fan, Liao, and Mincheva 2013; Fan, Guo,

and Zheng 2022). In the case of $\alpha = 2$ (i.e., sub-Gaussian), this uniform convergence rate becomes $\sqrt{\log p/n}$, which is consistent with the classical results in the existing literature (Fan, Guo, and Hao 2012; Wang 2012; Fan, Liao, and Mincheva 2013). However, this uniform convergence rate becomes slower if $0 < \alpha < 2$ with heavier distribution tails.

Theorem 1. Assume the conditions C1–C8 hold, we then have

$$\max_{1 \leq j \leq p} \|\hat{\theta}_{j,\text{cmle}} - \theta_j^*\| = O_p((\log p)^{1/\alpha} / \sqrt{n}).$$

2.3. Latent Factor Estimation by Diversified Projections

Next, consider how to estimate the latent factors in model (2.3). Following the idea of Fan and Liao (2022), we develop a method of diversified projections for factor estimation. Specifically, let $M = (m_{jk}) \in \mathbb{R}^{p \times d_{\max}}$ be a pre-specified projection matrix such that $M^\top M/p \rightarrow \Sigma_M \in \mathbb{R}^{d_{\max} \times d_{\max}}$ for some positive definite matrix Σ_M as $p \rightarrow \infty$. Here $d_{\max} \geq d$ is a pre-specified working number of factors. Then, the latent factor Z_i can be estimated by $M^\top \varepsilon_i/p = HZ_i + M^\top \omega_i/p$ up to an $d_{\max} \times d$ affine transformation $H = M^\top B^*/p \in \mathbb{R}^{d_{\max} \times d}$ and an estimation error $M^\top \omega_i/p$. Recall that $\mathbb{E} = (\varepsilon_{ij}) = (\varepsilon_1, \dots, \varepsilon_n)^\top = (\mathcal{E}_1, \dots, \mathcal{E}_p) \in \mathbb{R}^{n \times p}$, where $\varepsilon_i = (\varepsilon_{ij}) \in \mathbb{R}^p$ is the i th row vector of \mathbb{E} and $\mathcal{E}_j = (\varepsilon_{ij}) \in \mathbb{R}^n$ is the j th column vector of \mathbb{E} . Note that $\mathcal{E}_j = (I_n - \rho_j^* W) \mathbb{Y}_j - \mathbb{X}_{(j)} \beta_{(j)}^*$ by model (2.2). Then, a natural estimator for \mathcal{E}_j can be formed as $\hat{\mathcal{E}}_j = (I_n - \hat{\rho}_{j,\text{cmle}} W) \mathbb{Y}_j - \mathbb{X}_{(j)} \hat{\beta}_{(j),\text{cmle}}$. This leads to an estimated residual matrix $\hat{\mathbb{E}} = (\hat{\varepsilon}_{ij}) \in \mathbb{R}^{n \times p}$, which serves an estimator for $\mathbb{E} = (\varepsilon_{ij}) \in \mathbb{R}^{n \times p}$. In practice, Z_i can be estimated by $\hat{Z}_i = M^\top \hat{\varepsilon}_i/p$. However, whether the estimation error between \hat{Z}_i and Z_i is asymptotically negligible is not clear. Therefore, we are motivated to study the asymptotic behaviors of \hat{Z}_i rigorously.

To this end, one more technical condition is needed. For an arbitrary matrix, define $\|A\| = \lambda_{\max}^{1/2}(A^\top A)$. Then following Fan and Liao (2022), we further impose the following technical condition.

- (C9) Assume that (1) $\max_{1 \leq j \leq p} |m_{jk}| \leq C > 0$ for every $1 \leq k \leq d_{\max}$ with some positive constant C , and (2) $\text{rank}(H) = d$, $\lambda_{\min}(H^\top H) \gg 1/p$ and $\lambda_{\max}(H^\top H) \leq C \lambda_{\min}(H^\top H)$ with $H = M^\top B^*/p \in \mathbb{R}^{d_{\max} \times d}$.

Condition C9 is a combination of Assumptions 2.1 and 2.2 in Fan and Liao (2022), focusing on the projection matrix (i.e., M) and transformation matrix (i.e., H). Specifically, the first part of Condition C9 requires that the projection matrix M should be uniformly bounded elementwise. The second part of Condition C9 prevents M from being orthogonal to B^* . Otherwise, the projected random variable $M^\top \varepsilon_i/p$ becomes uncorrelated with the latent factor of interest Z_i . In that case, we lose the opportunity to estimate Z_i consistently. Write $\hat{\mathbb{Z}} = (\hat{Z}_1, \dots, \hat{Z}_n)^\top \in \mathbb{R}^{n \times d_{\max}}$. Then, we have the following theorem.

Theorem 2. Assume conditions C1–C8 hold, we then have

$$\|\hat{\mathbb{Z}} - \mathbb{Z}H^\top\|/\sqrt{n} = O_p(1/\sqrt{n} + 1/\sqrt{p}).$$

The detailed proof of Theorem 2 is given in Appendix A.2. By Theorem 2, we know that $\|\hat{\mathbb{Z}} - \mathbb{Z}H^\top\|/\sqrt{n}$ converges to 0 at

a rate $O_p(1/\sqrt{n} + 1/\sqrt{p})$ with $\mathbb{Z} = (Z_1, \dots, Z_n)^\top \in \mathbb{R}^{n \times d}$. This convergence rate contains two parts. The first part $1/\sqrt{n}$ is due to the estimation error of the initial estimator $\hat{\theta}_{j,\text{cmle}}$ and the second part $1/\sqrt{p}$ is due to the projection error of the diversified projections.

It is remarkable that Condition C9 plays an important role in Theorem 2. Therefore, an appropriate specification of M is practically important. In this regard, Fan and Liao (2022) propose four effective solutions. The first solution is called loading characteristics. In this approach, one can construct M by the observed characteristics related to factor loadings. The second solution is called moving window estimation. In this approach, one needs to divide the data into two parts. One can then construct M by the principal component loadings on the first part and then estimate factors by the second part. The third solution is called initial transformation. One can construct M by an appropriate transformation of the initial observation. The fourth solution is called Hadamard projection, which is based on Walsh–Hadamard matrix from a carefully designed statistical experiment. In this work, we implement a random partition method, which is similar to the second solution of Fan and Liao (2022).

2.4. Factor-Augmented Maximum Likelihood Estimators

By the SAR model (2.2) and the factor model (2.3), we obtain a factor-augmented spatial autoregressive (FSAR) model as

$$\mathbb{Y}_j = \rho_j W \mathbb{Y}_j + \mathbb{X}_{(j)} \beta_{(j)} + \tilde{\mathbb{Z}} \tilde{b}_j + \Omega_j, \quad (2.5)$$

where $\tilde{\mathbb{Z}} = \mathbb{Z} H^\top \in \mathbb{R}^{n \times d_{\max}}$ is the common factor after H -transformation, $\tilde{b}_j = H(H^\top H)^{-1} b_j \in \mathbb{R}^{d_{\max}}$, $b_j = (b_{jk}) \in \mathbb{R}^d$ is the j th row vector of B , and $\Omega_j = (\omega_{ij}) \in \mathbb{R}^n$ is the independent random noise with mean 0 and covariance $\tau_{jj} I_n$. For a given j , model (2.5) is similar to the spatial autoregressive model with additional X -covariates (i.e., $\tilde{\mathbb{Z}}$). However, there is a critical difference. That is an “additional X -covariates” here (i.e., $\tilde{\mathbb{Z}}$) is a latent random matrix and cannot be directly observed. A natural solution is to replace $\tilde{\mathbb{Z}}$ by its estimator $\hat{\mathbb{Z}}$. Then a factor-augmented log-likelihood function can be specified out as

$$\begin{aligned} \mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \hat{\mathbb{Z}}) = & -\frac{n}{2} \log \tau_{jj} + \log |S_j| \\ & - \frac{1}{2\tau_{jj}} \left(S_j \mathbb{Y}_j - \mathbb{X}_{(j)} \beta_{(j)} - \hat{\mathbb{Z}} \tilde{b}_j \right)^\top \\ & \left(S_j \mathbb{Y}_j - \mathbb{X}_{(j)} \beta_{(j)} - \hat{\mathbb{Z}} \tilde{b}_j \right), \end{aligned}$$

where $\Theta_j = (\rho_j, \beta_{(j)}^\top, \tilde{b}_j^\top, \tau_{jj})^\top \in \mathbb{R}^{g_j}$ with $g_j = s_j + d_{\max} + 2$. By maximizing $\mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \hat{\mathbb{Z}})$ with respect to Θ_j , we obtain $\hat{\Theta}_{j,\text{fmle}} = (\hat{\rho}_{j,\text{fmle}}, \hat{\beta}_{(j),\text{fmle}}, \hat{b}_{j,\text{fmle}}, \hat{\tau}_{jj,\text{fmle}})^\top = \arg\max_{\Theta_j} \mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \hat{\mathbb{Z}})$. Here we refer to $\hat{\Theta}_{j,\text{fmle}}$ as a factor-augmented maximum likelihood estimator (FMLE). Numerically, the FMLE $\hat{\Theta}_{j,\text{fmle}}$ with different j can be computed in a fully parallel or distributed way.

We next consider how to establish the asymptotic properties of $\hat{\Theta}_{j,\text{fmle}}$. Note that the resulting estimator $\hat{\Theta}_{j,\text{fmle}}$ is defined as the maximizer of $\mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \hat{\mathbb{Z}})$ with the estimated latent factors $\hat{\mathbb{Z}}$. It is unclear whether the estimation error of $\hat{\mathbb{Z}}$ affects the

asymptotic behaviors of $\hat{\Theta}_{j,\text{fmle}}$. Recall that $\hat{\mathbb{Z}}$ is a consistent estimator for $\tilde{\mathbb{Z}}$ by Theorem 2. Let $\Theta_j^* = (\rho_j^*, \beta_{(j)}^{*\top}, \tilde{b}_j^{*\top}, \tau_{jj}^*)^\top$ denotes the true value of Θ_j with $\tilde{b}_j^* = H(H^\top H)^{-1} b_j^*$. Then we can apply the Taylor's expansion and obtain the following asymptotic approximation for $\hat{\Theta}_{j,\text{fmle}}$ as

$$\begin{aligned} \sqrt{n}(\hat{\Theta}_{j,\text{fmle}} - \Theta_j^*) = & \left\{ -\ddot{\mathcal{L}}_{\Theta_j^* \Theta_j^*}(\Theta_j^*, \tilde{\mathbb{Z}})/n \right\}^{-1} \\ & \left\{ \dot{\mathcal{L}}_{\Theta_j^*}(\Theta_j^*, \tilde{\mathbb{Z}})/\sqrt{n} + \sum_{i=1}^n \ddot{\mathcal{L}}_{\Theta_j^* \tilde{Z}_i}(\Theta_j^*, \tilde{\mathbb{Z}})(\tilde{Z}_i - \tilde{Z}_i)/\sqrt{n} + o_p(1) \right\}, \end{aligned} \quad (2.6)$$

where $\dot{\mathcal{L}}_{\Theta_j}(\Theta_j, \tilde{\mathbb{Z}}) = \partial \mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \tilde{\mathbb{Z}})/\partial \Theta_j \in \mathbb{R}^{g_j}$ and $\ddot{\mathcal{L}}_{\Theta_j \Theta_j}(\Theta_j, \tilde{\mathbb{Z}}) = \partial^2 \mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \tilde{\mathbb{Z}})/\partial \Theta_j \partial \Theta_j^\top \in \mathbb{R}^{g_j \times g_j}$ are the 1st and 2nd order partial derivatives of $\mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \tilde{\mathbb{Z}})$ with respect to Θ_j , respectively. Here $\ddot{\mathcal{L}}_{\Theta_j \tilde{Z}_i}(\Theta_j, \tilde{\mathbb{Z}}) = \partial^2 \mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \tilde{\mathbb{Z}})/\partial \Theta_j \partial \tilde{Z}_i^\top \in \mathbb{R}^{g_j \times d_{\max}}$ is the 2nd order partial derivative of $\mathcal{L}_{\text{fmle}}^{(j)}(\Theta_j, \tilde{\mathbb{Z}})$ with respect to Θ_j and \tilde{Z}_i . Compared with the classical approximation theory (Lee 2004; Lee and Yu 2010), there involves an extra term $\sum_{i=1}^n \ddot{\mathcal{L}}_{\Theta_j^* \tilde{Z}_i}(\Theta_j^*, \tilde{\mathbb{Z}})(\tilde{Z}_i - \tilde{Z}_i)/\sqrt{n}$ in (2.6) due to the estimation of $\tilde{\mathbb{Z}}$. This motivates us to study this extra term rigorously. It can be theoretically verified that

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \ddot{\mathcal{L}}_{\Theta_j^* \tilde{Z}_i}(\Theta_j^*, \tilde{\mathbb{Z}}) \sqrt{n}(\tilde{Z}_i - \tilde{Z}_i) \\ & = -\frac{1}{p\tau_{jj}} \sum_{k=1}^p \left\{ c_{jk}^{\Theta_j^*} \sqrt{n}(\rho_k^* - \hat{\rho}_{k,\text{cmle}}) \right. \\ & \quad \left. + Q_{jk}^{\Theta_j^*} \sqrt{n}(\beta_{(k)}^* - \hat{\beta}_{(k),\text{cmle}}) \right\} + o_p(1), \end{aligned} \quad (2.7)$$

where $c_{jk}^{\Theta_j^*} \in \mathbb{R}^{g_j}$ and $Q_{jk}^{\Theta_j^*} \in \mathbb{R}^{g_j \times s_k}$ are some unknown parameters defined in (A.23) and (A.24) of Appendix A.3, respectively.

Note that $\hat{\rho}_{k,\text{cmle}}$ and $\hat{\beta}_{(k),\text{cmle}}$ for every $1 \leq k \leq p$ are the initial estimators defined in Section 2.2. By (2.7), we find that the estimation error of $\hat{\mathbb{Z}}$ should play an important role in determining the asymptotic distribution of $\hat{\Theta}_{j,\text{fmle}}$ through $\hat{\rho}_{k,\text{cmle}}$ and $\hat{\beta}_{(k),\text{cmle}}$. Then, the asymptotic properties of $\hat{\Theta}_{j,\text{fmle}}$ can be well studied by combining (2.6) and (2.7). To this end, one more technical condition is needed.

(C10) Assume that $\sum_{k=1}^p |\tau_{jk}^*| = O(1)$, $\lambda_{\max}(\mathcal{A}_j/p^2) = o(1)$, and $\lambda_{\min}(\mathcal{D}_j/p^2) \geq C_d$ for every $1 \leq j \leq p$, where $C_d > 0$ is a fixed constant, \mathcal{A}_j and \mathcal{D}_j are some matrices defined in (A.30) and (A.33) of Appendix A.3, respectively.

Condition C10 puts one particular type of sparsity constraint on the covariance matrix $\Sigma_\omega^* = (\tau_{jk}^*) \in \mathbb{R}^{p \times p}$ (Fan, Fan, and Lv 2008; Bai 2012; Wang 2012). It can be well satisfied for many important special cases, such as $\Sigma_\omega = \text{diag}(\tau_{11}, \dots, \tau_{pp}) \in \mathbb{R}^{p \times p}$. We then have the following theorem about the asymptotic behavior of $\hat{\Theta}_{j,\text{fmle}}$, which is proved in Appendix A.3.

Theorem 3. Assume conditions C1–C10 hold, we then have $\sqrt{n}(\hat{\Theta}_{j,\text{fmle}} - \Theta_j^*) \rightarrow_d N(0, \Sigma_{2\Theta_j^*}^{-1} \Sigma_{1\Theta_j^*}^{-1} \Sigma_{2\Theta_j^*}^{-1})$ as $n \rightarrow \infty$, where

$$\Sigma_{1\Theta_j^*}^* = \Sigma_{2\Theta_j^*}^* + \Delta_{\Theta_j^*}^* + \Sigma_{Q_j} \in \mathbb{R}^{g_j \times g_j},$$

$$\Sigma_{2\Theta_j^*}^* = \begin{pmatrix} \Sigma_{2\rho_j^* \rho_j^*}^* & \Sigma_{2\rho_j^* \rho_j^*}^{\top} & \Sigma_{2b_j^* \rho_j^*}^{\top} & \Sigma_{2\tau_{jj}^* \rho_j^*}^* \\ \Sigma_{2\rho_j^* \rho_j^*}^* & \Sigma_{2\rho_j^* \rho_j^*}^{\top} & 0_{q, d_{\max}} & 0_q \\ \Sigma_{2b_j^* \rho_j^*}^* & 0_{d_{\max}, q} & \Sigma_{2b_j^* b_j^*}^* & 0_{d_{\max}} \\ \Sigma_{2\tau_{jj}^* \rho_j^*}^* & 0_q^{\top} & 0_{d_{\max}}^{\top} & \Sigma_{2\tau_{jj}^* \tau_{jj}^*}^* \end{pmatrix},$$

$$\Delta_{\Theta_j^*} = \begin{pmatrix} \Delta_{\rho_j^* \rho_j^*}^* & \Delta_{\rho_j^* \rho_j^*}^{\top} & \Delta_{b_j^* \rho_j^*}^{\top} & \Delta_{\tau_{jj}^* \rho_j^*}^* \\ \Delta_{\rho_j^* \rho_j^*}^* & 0_{q, q} & 0_{q, d_{\max}} & \Delta_{\tau_{jj}^* \rho_j^*}^* \\ \Delta_{b_j^* \rho_j^*}^* & 0_{d_{\max}, q} & 0_{d_{\max}, d_{\max}} & \Delta_{\tau_{jj}^* \rho_j^*}^* \\ \Delta_{\tau_{jj}^* \rho_j^*}^* & \Delta_{\tau_{jj}^* \rho_j^*}^{\top} & \Delta_{\tau_{jj}^* b_j^*}^{\top} & \Delta_{\tau_{jj}^* \tau_{jj}^*}^* \end{pmatrix},$$

and $\Sigma_{Q_j} \in \mathbb{R}^{g_j \times g_j}$. The analytical expressions of the matrices $\Sigma_{2\Theta_j^*}^*$, $\Delta_{\Theta_j^*}^*$, and Σ_{Q_j} are given in Appendix A.3 and Appendix C.3, respectively.

By Theorem 3, we know that $\hat{\Theta}_{j, \text{fmle}}$ is \sqrt{n} -consistent and asymptotically normal. Note that asymptotic covariance of $\hat{\Theta}_{j, \text{fmle}}$ consists of three parts. The first part $\Sigma_{2\Theta_j^*}^*$ represents a typical information matrix under normality. The second part $\Delta_{\Theta_j^*}$ contains high order moments of the disturbances (Lee 2004; Lee and Yu 2010; Yang and Lee 2017). This part becomes zero if Ω_j in (2.5) follows a normal distribution strictly. Moreover, the last term Σ_{Q_j} is due to the estimation error of $\tilde{\mathbb{Z}}$. We should have $\Sigma_{Q_j} = 0$ if the true $\tilde{\mathbb{Z}}$ were actually observed. In this case, $\hat{\Theta}_{j, \text{fmle}}$ becomes statistically as efficient as the oracle estimator $\tilde{\Theta}_{j, \text{fmle}} = \arg\max_{\Theta} \mathcal{L}_{\text{fmle}}^{(j)}(\Theta, \tilde{\mathbb{Z}})$.

2.5. Shrinkage Estimation and Uniform Selection Consistency

Next, we consider how to consistently estimate the unknown $S_{(j), T}$ for every response \mathbb{Y}_j in practice. To this end, various shrinkage estimation techniques can be considered (Tibshirani 1996; Fan and Li 2001; Fan, Wang, and Zhu 2021). In this work, we focus on the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001) and Fan and Lv (2011) due to its excellent theoretical properties. The methodology developed below can be readily applied to other popular shrinkage methods without additional difficulty (Tibshirani 1996; Efron et al. 2004; Fan, Wang, and Zhu 2021). Specifically, we define a penalized likelihood function for the model (2.1) as $Q_{\lambda}^{(j)}(\theta_j) = \mathcal{L}^{(j)}(\theta_j) - n \sum_{k=1}^q p_{\lambda}(|\beta_{jk}|)$, where $\mathcal{L}^{(j)}(\theta_j) = -n(\log \sigma_{jj})/2 + \log |S_j| - (S_j \mathbb{Y}_j - \mathbb{X} \beta_j)^{\top} (S_j \mathbb{Y}_j - \mathbb{X} \beta_j) / (2\sigma_{jj})$ and $p_{\lambda}(\cdot)$ is the SCAD penalty function with its first order derivative given by $\dot{p}_{\lambda}(t) = \lambda[I(t \leq \lambda) + (a\lambda - t)_+ I(t > \lambda)] / \{(a-1)\lambda\}$. Here a is some constant that is often taken to be 3.7 (Fan and Li 2001; Fan and Lv 2011), λ is a tuning parameter, $(t)_+ = tI(t > 0)$, and $I(\cdot)$ is the indicator function. Then, a SCAD estimator can be obtained as $\hat{\theta}_{j, \lambda} = (\hat{\rho}_{j, \lambda}, \hat{\beta}_{j, \lambda}^{\top}, \hat{\sigma}_{j, \lambda})^{\top} = \arg\max_{\Theta} Q_{\lambda}^{(j)}(\Theta)$.

As demonstrated by Fan and Li (2001) and many subsequent works (Fan and Lv 2011; Fan et al. 2020), the SCAD estimator has excellent model selection capabilities for various statistical models. It is then of interest to study whether similar properties can be reproduced in our case. Following the literature (Wang,

Li, and Tsai 2007; Wang, Li, and Leng 2009), write λ_n as a tuning parameter sequence indexed by n . Accordingly, define $\hat{S}_{(j), \lambda_n} = \{1 \leq k \leq q : \hat{\beta}_{jk, \lambda_n} \neq 0\}$ as the model set selected by $\hat{\theta}_{j, \lambda_n}$. Recall that $\beta_{(j)} = \{\beta_{jk} : k \in S_{(j), T}\} \in \mathbb{R}^{g_j}$ is the sub-vector associated with the nonzero coefficients. Define $\beta_{(-j)} = \{\beta_{jk} : k \notin S_{(j), T}\} \in \mathbb{R}^{q-g_j}$ to be the sub-vector associated with the zero coefficients for every $1 \leq j \leq p$. Write $\hat{\beta}_{j, \lambda_n} = (\hat{\beta}_{(j), \lambda_n}^{\top}, \hat{\beta}_{(-j), \lambda_n}^{\top})^{\top} \in \mathbb{R}^q$ and $\beta_j^* = (\beta_{(j)}^{\top}, \beta_{(-j)}^{\top})^{\top} \in \mathbb{R}^q$ with $\beta_{(j)}^* \neq 0$ and $\beta_{(-j)}^* = 0$. Then, the following theorem establishes the uniform selection consistency of $\hat{S}_{(j), \lambda_n}$ over $1 \leq j \leq p$.

Theorem 4. Assume the conditions C1–C10 hold. Further assume that $\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n / \{\log(pq)^{1/\alpha}\} \rightarrow \infty$ as $n \rightarrow \infty$, we then have

$$P(\hat{S}_{(j), \lambda_n} = S_{(j), T}, \text{ for every } 1 \leq j \leq p) \rightarrow 1.$$

The detailed proof of Theorem 4 is provided in Appendix A.4. By Theorem 4, we know that, with probability tending to one, the selected set $\hat{S}_{(j), \lambda_n}$ consistently recovers the true set $S_{(j), T}$ exactly in a way uniformly over $1 \leq j \leq p$. It is remarkable that this uniform selection consistency result is stronger than the conventional selection consistency discussed in the existing literature (Shao 1993; Wang, Li, and Tsai 2007; Wang, Li, and Leng 2009), which focuses on a single $j \in \{1, \dots, p\}$.

We next consider how to specify λ_n practically. To this end, a number of Bayesian information criterion (BIC) methods have been developed under various model setups (Wang, Li, and Tsai 2007; Chen and Chen 2008; Wang, Li, and Leng 2009; Wang 2012). We are then inspired to develop for our model (2.1) a similar BIC-type criterion as

$$\text{BIC}^{(j)}(\lambda) = -\frac{1}{n} \mathcal{L}^{(j)}(\hat{\theta}_{j, \lambda}) + \frac{1}{n} |\hat{S}_{(j), \lambda}| (\log n) \{\log(pq)\}^{2/\alpha}.$$

Note that this BIC criterion contains two components. The first component $\mathcal{L}^{(j)}(\hat{\theta}_{j, \lambda})/n$ reflects the goodness-of-fit. The second component penalizes the model complexity $|\hat{S}_{(j), \lambda}|$ by a factor $(\log n) \{\log(pq)\}^{2/\alpha}/n$. The first factor $\log n$ is due to the diverging sample size n , and the second factor $\{\log(pq)\}^{2/\alpha}/n$ is due to the diverging feature q and the diverging response p . Penalizing factors of a similar form have been popularly used in the literature (Chen and Chen 2008; Wang, Li, and Leng 2009; Zhang, Wang, and Zhang 2024).

Then, an optimal tuning parameter can be selected as $\hat{\lambda}_{(j), \text{BIC}} = \arg\min_{\lambda} \text{BIC}^{(j)}(\lambda)$. Note that $\hat{\lambda}_{(j), \text{BIC}}$ is an estimator depending on n . This leads to a selected model as $\hat{S}_{(j), \hat{\lambda}_{(j), \text{BIC}}}$. To study its uniform selection consistency property, define $S_F = \{1, \dots, q\}$ as the full model. Write $S_{(j)} \subset S_F$ with size $q_j = |S_{(j)}|$ as an arbitrary working model for the j th response. Define $\mathcal{R}_{(j)}(\theta) = E\{-\mathcal{L}^{(j)}(\theta)/n\}$ as the risk function, and let $\mathcal{R}_{(j), \min}^* = \mathcal{R}_{(j)}(\theta_j^*)$ be its minimum value evaluated at the true parameter. Then, the following technical condition is necessarily needed.

(C11) Assume that there exists some positive and fixed constant $\delta_{\min} > 0$ such that $\min_{1 \leq j \leq p} \min_{S_{(j)} \not\supset S_{(j), T}} \inf_{\theta_j \in S_{(j)}} \{\mathcal{R}_{(j)}(\theta_j, S_{(j)}) - \mathcal{R}_{(j), \min}^*\} \geq \delta_{\min}$.

Condition **C11** imposes a strict separation condition on the risk function $\mathcal{R}_{(j)}(\theta)$. It ensures that the minimal risk of any underfitted working model (i.e., $\mathcal{S}_{(j)} \not\supset \mathcal{S}_{(j),T}$) must be strictly larger than that of the true model by a fixed margin δ_{\min} . Similar conditions have been widely used in the literature; see for example Condition (2.5) in Shao (1993), Condition 2 in Wang, Li, and Tsai (2007), and Assumption 2 in Fan, Guo, and Hao (2012). Then, the uniform selection consistency of $\mathcal{S}_{(j),\hat{\lambda}_{(j),\text{BIC}}}$ can be rigorously established by Theorem 5, whose detailed proof is given in Appendix A.5. By Theorem 5, we know that, with probability tending to one, the selected model $\hat{\mathcal{S}}_{(j),\hat{\lambda}_{(j),\text{BIC}}}$ recovers the true model $\mathcal{S}_{(j),T}$ uniformly over $1 \leq j \leq p$.

Theorem 5. Assume the conditions **C1–C11** hold, we then have as $n \rightarrow \infty$,

$$P(\hat{\mathcal{S}}_{(j),\hat{\lambda}_{(j),\text{BIC}}} = \mathcal{S}_{(j),T}, \text{ for every } 1 \leq j \leq p) \rightarrow 1.$$

As we mentioned before, these true model sets $\mathcal{S}_{j,T}$ s are practically unknown and therefore have to be empirically estimated by $\hat{\mathcal{S}}_{(j),\hat{\lambda}_{(j),\text{BIC}}}$ s. Once they are empirically estimated, they are then treated as if they were the truth. Thereafter, the three-step estimators as developed in Sections 2.2–2.4 can be readily computed. This leads to the final empirical estimators. Strictly speaking, the empirical estimators finally computed are different from the oracle estimators studied in Sections 2.2–2.4, since there exists a positive probability for $\hat{\mathcal{S}}_{(j),\hat{\lambda}_{(j),\text{BIC}}} \neq \mathcal{S}_{j,T}$. Nevertheless, this probability shrinks to zero as $n \rightarrow \infty$ due to the uniform selection consistency results as established in Theorems 4 and 5. Therefore, the two estimators (i.e., the oracle estimators and the empirical estimators) share the same asymptotic distribution. Therefore, we are able to claim that both estimators are equivalent asymptotically.

3. Numerical Studies

3.1. Simulation Models

To demonstrate the finite sample performance of the FSAR model, we conduct a number of simulation studies. For each simulation replication, we first generate the adjacency matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{n \times n}$, and then set the diagonal element $a_{ii} = 0$ for every $1 \leq i \leq n$. Note that A is not necessarily a symmetric matrix. Thereafter, the adjacency matrix A is row-normalized as $w_{i_1 i_2} = a_{i_1 i_2}/n_{i_1}$ for each row $1 \leq i_1 \leq n$. This leads to the spatial weight matrix $W = (w_{i_1 i_2}) \in \mathbb{R}^{n \times n}$. Regarding the adjacency matrix A , three widely standard network structures are considered.

Example 1 (Dyad Independence Model, DIM). Following Holland and Leinhardt (1981), define a dyad as $\mathcal{A}_{i_1 i_2} = (a_{i_1 i_2}, a_{i_2 i_1})$ for any $1 \leq i_1 < i_2 \leq n$. Different $\mathcal{A}_{i_1 i_2}$ s are assumed to be mutually independent. Next, following Zhu et al. (2020), define $P\{\mathcal{A}_{i_1 i_2} = (1, 1)\} = 2n^{-1}$ and $P\{\mathcal{A}_{i_1 i_2} = (1, 0)\} = P\{\mathcal{A}_{i_1 i_2} = (0, 1)\} = 0.5n^{-0.8}$. As a result, the expected number of the mutually connected dyads with $\mathcal{A}_{i_1 i_2} = (1, 1)$ is $O(n)$. In the meanwhile, the expected degree of each node to be slowly diverging in the order of $O(n^{0.2})$. Then, we have $P\{\mathcal{A}_{i_1 i_2} =$

$(0, 0)\} = 1 - 2n^{-1} - n^{-0.8}$, which is close to 1 as the network size $n \rightarrow \infty$.

Example 2 (Stochastic Block Model, SBM). Consider a network structure generated from the stochastic block model. Specifically, set $K = 5$ be the total number of blocks. Next, following Nowicki and Snijders (2001), we randomly assign each node a block label ($k = 1, \dots, K$) with equal probability $1/K$. Let $P(a_{i_1 i_2} = 1) = 9/n$ if i_1 and i_2 belongs to the same block, and $P(a_{i_1 i_2} = 1) = 3/n$ otherwise. Therefore, nodes within the same block are more likely to be connected with each other.

Example 3 (Latent Space Model, LSM). Following Hoff, Raftery, and Handcock (2002), assume that the node i has a low-dimensional position d_i in the latent space for every $1 \leq i \leq n$. The probability of two nodes being connected (i.e., $P(a_{i_1 i_2} = 1)$) is determined by the distance between their respective latent positions (i.e., $d_{i_1 i_2} = \|d_{i_1} - d_{i_2}\|$). Here, we set $P(a_{i_1 i_2} = 1) = \exp(-0.25nd_{i_1 i_2}) / \{1 + \exp(-0.25nd_{i_1 i_2})\}$, where d_i is independently and identically uniformly distributed on $(0, 1)$ for every $1 \leq i \leq n$.

For each network structure, consider multiple network sizes (i.e., $n = 500, 1000$, and 1500), response dimensions (i.e., $p = 50, 100$, and 200), covariate dimensions (i.e., $q = 5, 10$, and 20), and latent factor dimensions (i.e., $d = 1, 2$, and 3). Next, generate $\omega_i = (\omega_{i1}, \dots, \omega_{ip})^\top \in \mathbb{R}^p$ with ω_{ij} independently drawn from $N(0, \tau_{jj})$, where τ_{jj} is independently generated from a uniform distribution $U(0.1, 0.2)$. Then, both $X_i \in \mathbb{R}^q$ and $Z_i \in \mathbb{R}^d$ are sampled from a standard multivariate normal distribution. The true model size of each response is set to be $s_j = 2$ for $q = 5$, $s_j = 5$ for $q = 10$, and $s_j = 10$ for $q = 20$. Then, for $\mathcal{B} = (\beta_{jh}) \in \mathbb{R}^{p \times q}$, we independently sample β_{jh} s from $U(0.5, 1)$ if $1 \leq h \leq s_j$, and 0 otherwise. Then, for the true parameters $B = (b_{jk}) \in \mathbb{R}^{p \times d}$ and $\rho = (\rho_1, \dots, \rho_p)^\top \in \mathbb{R}^p$, we independently sample b_{jk} s from $N(0, 1)$, and ρ_j s from $U(0.2, 0.9)$, respectively. Accordingly, the high-dimensional response matrix $Y \in \mathbb{R}^{n \times p}$ can be obtained according to model (2.5).

3.2. Simulation Results

We start with the uniform convergence of CMLE $\hat{\rho}_{j,\text{cmle}}$ s. Given a specification (W, n, p, q, d) , we randomly replicate the experiment for a total of $R = 500$ times. Since the simulation results are qualitatively similar, we only report here the case with $(q, d) = (20, 3)$. We use $\hat{\rho}_{j,\text{cmle}}^{(r)}$ to represent one particular estimator obtained in the r th replication ($1 \leq r \leq R$). The true parameter is denoted by ρ_j . Define the estimation error (Err) as $\text{Err}_{j,c}^{(r)} = |\hat{\rho}_{j,\text{cmle}}^{(r)} - \rho_j|$ for every $\hat{\rho}_{j,\text{cmle}}^{(r)}$, and the maximum error (MaxErr) over j as $\text{MaxErr}_c^{(r)} = \max_{1 \leq j \leq p} \text{Err}_{j,c}^{(r)}$. This leads to a total of R MaxErr_c values, which are then log-transformed and box-plotted in Figure 1. By Figure 1, we obtain the following two findings. First, for a fixed W and p , the maximum error (MaxErr) decreases as the sample size n increases. This provides empirical evidence for the uniform consistency of $\hat{\rho}_{j,\text{cmle}}$ over $1 \leq j \leq p$. Moreover, with a fixed W and n but diverging p , the maximum error (MaxErr) increases slowly. This suggests that the uniform convergence rate of $\hat{\rho}_{j,\text{cmle}}$ diverges with respect to p but at a slow

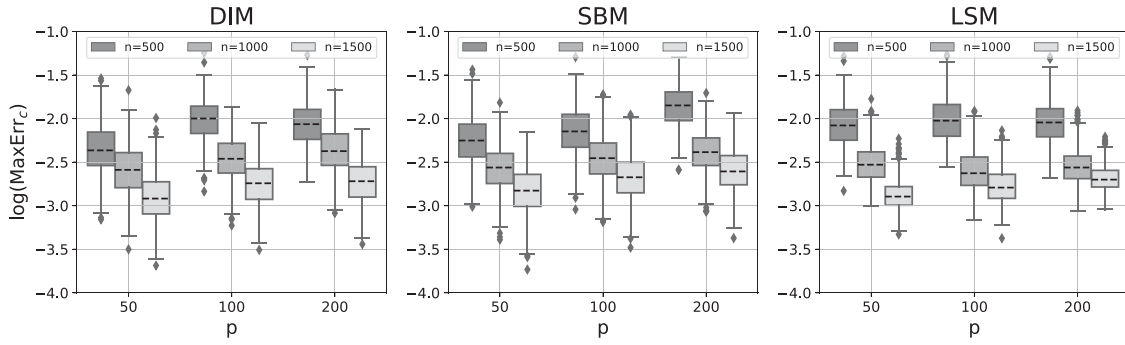


Figure 1. The $\log(\text{MaxErr}_c)$ values for CMLE $\hat{\rho}_{j,\text{cmle}}$ with $d = 3$. Different panels correspond to different network structures: DIM (the left), SBM (the middle), and LSM (the right). For a given panel, different groups correspond to different feature dimensions with $p = 50, 100$, and 200 , respectively. For a given group, the lighter the color of the box is, the larger the sample size is.

Table 1. The simulation results of FMLE for three networks.

n	$p = 50$			$p = 100$			$p = 200$		
	MErr_f	$\text{RIM}(\%)$	$\text{CP}(\%)$	MErr_f	$\text{RIM}(\%)$	$\text{CP}(\%)$	MErr_f	$\text{RIM}(\%)$	$\text{CP}(\%)$
<i>DIM</i>									
500	0.018	41.58	95.52	0.019	51.98	95.40	0.020	42.75	95.20
1000	0.016	32.47	95.00	0.015	42.44	95.20	0.014	46.58	94.40
1500	0.011	36.79	96.00	0.010	46.00	95.60	0.010	44.31	95.60
<i>SBM</i>									
500	0.025	29.48	95.60	0.021	40.29	95.80	0.020	52.91	95.00
1000	0.016	38.31	96.00	0.014	42.85	95.60	0.014	44.51	95.00
1500	0.012	37.74	95.00	0.012	42.07	95.00	0.012	43.40	94.00
<i>LSM</i>									
500	0.035	26.06	93.40	0.034	24.16	95.20	0.031	24.14	93.75
1000	0.026	16.96	94.20	0.023	17.42	94.00	0.022	17.86	95.03
1500	0.020	12.42	93.72	0.020	15.17	94.13	0.021	14.73	93.00

rate. All of these results are in line with our theoretical finding in [Theorem 1](#).

We next study the finite sample performance of the FMLE $\hat{\rho}_{j,\text{fmle}}$ s. To this end, we need to specify the projection matrix M . Similar to the moving window estimation method of Fan and Liao (2022), we implement here a random partition method, which uses 10% of the randomly generated sample to estimate the projection matrix M . Once M is specified, the rest 90% samples are then used to conduct the subsequent analysis. First, we compute for each $\hat{\rho}_{j,\text{fmle}}$ an Err value decoded by $\text{Err}_{j,f}^{(r)}$ at the r th replication. Then, we obtain the mean error as $\text{MErr}_f = (Rp)^{-1} \sum_{r=1}^R \sum_{j=1}^p \text{Err}_{j,f}^{(r)}$. For comparison purposes, the same value is also computed for the CMLE and is denoted as MErr_c . Next, we compare their relative efficiency by the relative improvement margin $\text{RIM} = (1 - \text{MErr}_f / \text{MErr}_c) \times 100\%$. Moreover, for each $1 \leq j \leq p$, a 95% confidence interval is constructed for ρ_j as $\text{CI}_j^{(r)} = (\hat{\rho}_{j,\text{fmle}} - z_{0.975} \widehat{\text{SE}}_j^{(r)}, \hat{\rho}_{j,\text{fmle}} + z_{0.975} \widehat{\text{SE}}_j^{(r)})$, where $\widehat{\text{SE}}_j^{(r)}$ is square root of the first (1, 1) component of $\widehat{\Sigma}_{2\Theta_j}^{-1} \widehat{\Sigma}_{1\Theta_j}^* \widehat{\Sigma}_{2\Theta_j}^{-1}$, and $\widehat{\Sigma}_{2\Theta_j}^{-1} \widehat{\Sigma}_{1\Theta_j}^* \widehat{\Sigma}_{2\Theta_j}^{-1}$ is a plug-in estimator of the asymptotic covariance matrix $\Sigma_{2\Theta_j}^{-1} \Sigma_{1\Theta_j}^* \Sigma_{2\Theta_j}^{-1}$ given in [Theorem 3](#). Here z_α is the α th quantile of a standard normal distribution. Then the coverage probability (CP) is computed as $\text{CP}_j = R^{-1} \sum_{r=1}^R I(\rho_j \in \text{CI}_j^{(r)})$. Different RIM values for different combinations (n, p, W) are computed and reported in [Table 1](#). For ease of presentation, a CP_j is randomly selected over $1 \leq j \leq p$ and also reported in [Table 1](#). By [Table 1](#), we find

that given p and W , larger sample sizes always lead to smaller MErr_f values. This confirms the consistency of FMLE, which in line with the result in [Theorem 3](#). Compared with CMLE, the estimation efficiency of FMLE is improved by about 35% on average. Moreover, the reported coverage probability values are all fairly close to the nominal level of 95%. This implies that the estimated standard error approximates the true standard error very well. Those results provide numerical evidence of the asymptotic theory obtained in [Theorem 3](#).

Lastly, we study the model selection results. Following [Section 2.5](#), we compute the SCAD estimators and use the BIC method to select the optimal tuning parameter λ . In this case, we randomly replicate the experiment for a total of $R = 100$ times for each (n, p, q, d, W) specification. Let $\widehat{\mathcal{S}}_{(j), \widehat{\lambda}_{(j), \text{BIC}}}^{(r)}$ represent one particular model set obtained in the r th replication ($1 \leq r \leq R$). Define the percentage of experiments with correctly identified true models (CM) as

$$\text{CM} = \frac{1}{R} \sum_{r=1}^R I(\widehat{\mathcal{S}}_{(j), \widehat{\lambda}_{(j), \text{BIC}}}^{(r)} = \mathcal{S}_{(j), T}, \text{ for every } 1 \leq j \leq p) \times 100\%.$$

This provides a uniform criterion for assessing model selection accuracy. Since the simulation results are qualitatively similar, we only report here case with $(q, d) = (20, 3)$ in [Figure 2](#). By [Figure 2](#), we find that CM values converge to 100% rapidly as the sample size n increases. This suggests that $\widehat{\mathcal{S}}_{(j), \widehat{\lambda}_{(j), \text{BIC}}}^{(r)}$ is uniformly consistent for recovering $\mathcal{S}_{(j), T}$, which is in line with our theoretical findings in [Theorem 5](#).

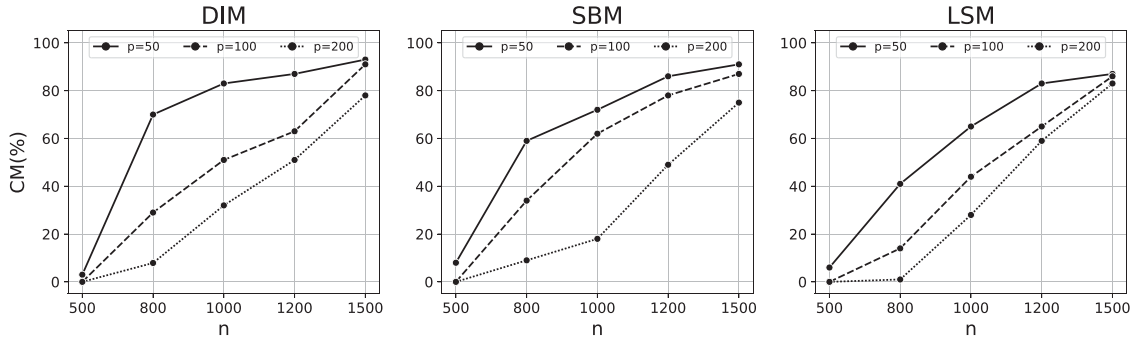


Figure 2. The CM values with $q = 20$ and $d = 3$. Different panels correspond to different network structures: DIM (the left), SBM (the middle), and LSM (the right). For a given panel, different lines correspond to different feature dimensions with $p = 50$ (solid), 100 (dashed), and 200 (dotted), respectively.

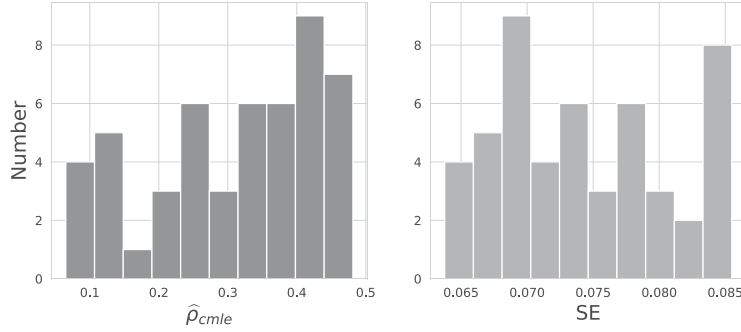


Figure 3. The histogram of the CMLE $\hat{\rho}_{j,cmle}$ s (the left panel) and the histogram of the estimated SE (the right panel).

3.3. A Real Data Example

To demonstrate the practical applications of the proposed FSAR model, we present here a case study. Specifically, we consider an urban statistics dataset collected from Urban Statistical Yearbook 2019 of China, which is published by National Bureau of Statistics (<http://www.stats.gov.cn/sj/ndsj/>). The full dataset contains a total of 287 nodes, with each node representing a city. For each node (i.e., city), we collect a total of 112 macroeconomic indicators from 2019. Those indicators provide detailed information on city-level statistics. However, some indicators suffer from a large proportion (more than 15%) of missing values and are then omitted for the subsequent analysis. For the remaining 50 indicators, the proportion of missing values does not exceed 5%. Details of these 50 indicators are provided in Table 2 of Appendix D. For these 50 indicators, the neighbor year interpolation method of Lunardi (2018) is employed to impute the missing values. Subsequently, these completed indicators are then log-transformed and standardized to have mean 0 and variance 1. This leads to a final high-dimensional dataset $Y = (Y_{ij}) \in \mathbb{R}^{n \times p}$ with $n = 287$ and $p = 50$. A spatial weight matrix $W = (w_{i1i2}) \in \mathbb{R}^{n \times n}$ is then constructed based on geographical locations (Lee and Yu 2010). We next fit an FSAR model to this dataset. Note that each node (i.e., city) is spatially connected with others through W . The associated spatial correlations ρ_{js} reflect the spatial spillover effects among cities (Zhou et al. 2023).

We start with computing the CMLE as an important initial estimator. The histograms of those estimators and their standard errors (SE) are then plotted in the left and right panel of Figure 3, respectively. We find that the resulting estimates $\hat{\rho}_{j,cmle}$ s varies greatly, ranging from 0.05 to 0.50 with estimated SEs ranging from 0.06 to 0.09. It is then of interest to understand

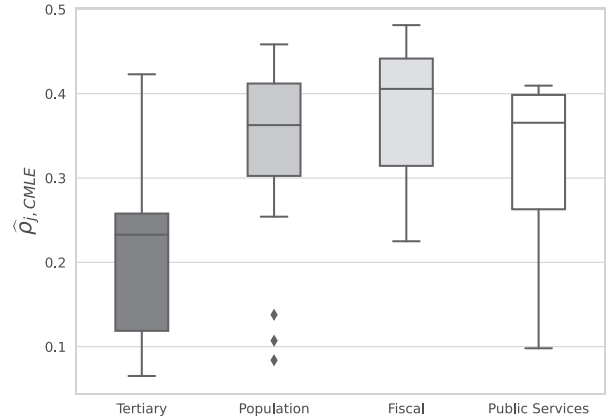


Figure 4. The boxplots of the CMLE $\hat{\rho}_{j,cmle}$ s for four categories. From left to right: Tertiary (tertiary industry), Population (labor and population development), Fiscal (fiscal and financial resources), and Public Services (infrastructure and public services).

the reason behind such considerable variation. To this end, we classify the 50 indicators into a total of four groups. They are, respectively, (a) tertiary industry related indicators, capturing the development of services and knowledge-based sectors such as accommodation and food services, and wholesale and retail trade (Kenessey 1987); (b) labor and population development related indicators, reflecting the dynamics of urban labor supply, employment structure, and demographic shifts (Fujita, Krugman, and Venables 2001); (c) fiscal and financial resources related indicators, indicating the capacity of local governments to mobilize and allocate financial resources, the strength of local fiscal institutions, and the accessibility of financial services (Gyourko and Tracy 1991); and (d) infrastructure and public

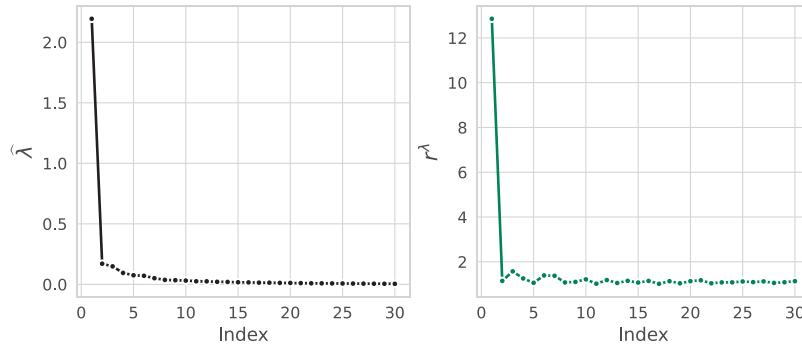


Figure 5. The top 30 estimated eigenvalues of $\hat{\mathbb{E}}$ (the left panel) and the top 30 eigenvalue ratios (the right panel).

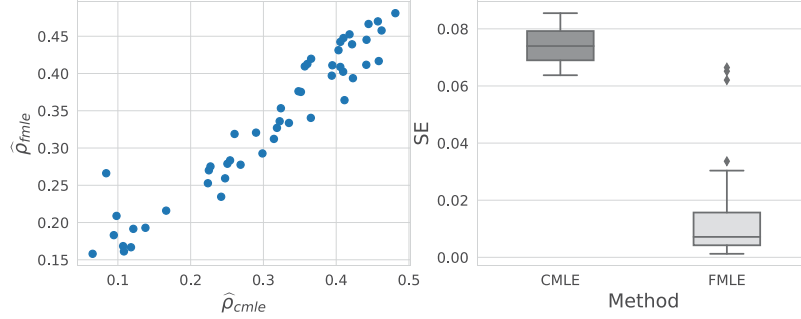


Figure 6. The scatterplot of the FMLE $\hat{\rho}_{fmle}$ and the associated CMLE $\hat{\rho}_{cmle}$ (the left panel). The boxplots of their standard errors (the right panel).

services related indicators, measuring the regional capacity to provide essential physical and social infrastructure (Démurger 2001). The group sizes are 14, 16, 13, and 7, respectively. The CMLEs of each group are then boxplotted in Figure 4. We find that the spatial spillover effects of fiscal and financial resources are the strongest, with the largest $\hat{\rho}_{j,cmle}$ s on average. This result aligns well with empirical findings in the urban and regional economics literature (Gyourko and Tracy 1991; Auerbach, Gorodnichenko, and Murphy 2020). It is also noteworthy that the spatial spillover effects of tertiary industry are the weakest on average. Practically, this pattern may be explained by the relatively localized nature of the service-oriented economic activities (Kenessey 1987; Yin et al. 2022).

Next, we need to decide the factor dimension. To this end, we compute $\hat{\varepsilon}_i = (\hat{\varepsilon}_{ij}) \in \mathbb{R}^p$ for each city i . Then, we compute the eigenvalues $(\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p)$ of the covariance matrix of $\hat{\mathbb{E}} = (\hat{\varepsilon}_{ij}) \in \mathbb{R}^{n \times p}$. The top 30 eigenvalues are then plotted in the left panel of Figure 5. It seems that the first eigenvalue is notably larger than others. Following Luo, Wang, and Tsai (2009) and Lam and Yao (2012), we calculate the eigenvalue ratio statistic as $r_j^\lambda = \hat{\lambda}_j / \hat{\lambda}_{j+1}$ with $1 \leq j \leq p - 1$. These values are then plotted in the right panel of Figure 5, which provides strong evidence for the existence of a one-dimensional factor structure. This finding is not totally surprising, as these 50 indicators are all macroeconomics related. Therefore, they are heavily correlated with the overall macroeconomics status (i.e., one common factor) of the target city. This makes the underlying factor structure of $\hat{\varepsilon}_i$ relatively simple. Similar low-dimensional factor structures are also often observed in empirical macroeconomics literature (Bai and Ng 2002; Bernanke, Boivin, and Elias 2005).

Lastly, we apply the proposed factor estimation method in Section 2.3 and obtain the estimated latent factor \hat{Z}_i for each

city i . The choice of the projection matrix is similar to that in simulation studies. Next, we compute the FMLE for every ρ_j ($1 \leq j \leq p$). The resulting FMLE estimates, along with the initial CMLE estimates, are then plotted in the left panel of Figure 6. We find that $\hat{\rho}_{fmle}$ is in line with $\hat{\rho}_{cmle}$. Moreover, their standard errors are boxplotted in the right panel of Figure 6. We find that the SEs of FMLE are considerably smaller than those of CMLE. Our estimation results reveal that there exists the significant spatial correlation in various macroeconomic indicators among these cities. Specifically, the largest spatial spillover effect is detected for revenue in the gross regional product (GRP) growth rate with $\hat{\rho}_{1,fmle} = 0.48$. In contrast, the smallest spatial spillover effect is detected for persons employed in culture, sports and entertainment with $\hat{\rho}_{37,fmle} = 0.16$.

4. Concluding Remarks

In this work, we study the problem of spatial autoregressive modeling for network data with high-dimensional responses and covariates. The key contribution lies in the development of a flexible factor-augmented spatial autoregressive (FSAR) model that accommodates both high-dimensionality and complex cross-sectional dependence across response variables. To conclude this article, we discuss here several interesting topics for future research. First, it is worth noting that the FSAR model requires the high-dimensional responses to be continuous. Then how to relax this continuity assumption is an interesting direction for future exploration. Second, it is assumed that the dimension of the latent factors is fixed. How to allow for a diverging number of latent factors should be another intriguing topic for the future study (Fan, Fan, and Lv 2008). Third, for the real urban statistics dataset analysis, the cross-response spillover effects

(e.g., the cross effect of retail sales and household wealth) are not explicitly characterized. Developing novel tools for better interpretation is also worth pursuing.

Supplementary Materials

Appendices A–C provide the proofs of all theoretical results and some useful lemmas, and Appendix D contains the supplementary table. The code is publicly available on GitHub at <https://github.com/Shi12056/FactorSAR.git>.

Acknowledgments

The authors are very grateful to the editor, the associate editor, and referees for their constructive comments and suggestions, which greatly improved the quality of this article.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Xuening Zhu's research is supported by the National Natural Science Foundation of China (Nos. 72222009, 72573038, 12331009), MOE Laboratory for National Development and Intelligent Governance, Fudan University. The research of Jing Zhou is partially supported by the National Natural Science Foundation of China (Nos. 72171226, 11971504) and the National Statistical Science Research Project (No. 2023LD008). Hansheng Wang's research is partially supported by the National Natural Science Foundation of China (Nos. 72495123, 12271012).

References

- Anselin, L. (1988), *Spatial Econometrics: Methods and Models* (Vol. 4), Dordrecht: Springer. [1]
- Auerbach, A., Gorodnichenko, Y., and Murphy, D. (2020), "Local Fiscal Multipliers and Fiscal Spillovers in the USA," *IMF Economic Review*, 68, 195–229. [10]
- Bai, J. (2012), "Statistical Analysis of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436–465. [2,3,5]
- Bai, J., and Li, K. (2021), "Dynamic Spatial Panel Data Models with Common Shocks," *Journal of Econometrics*, 224, 134–160. [2]
- Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [10]
- Bernanke, B. S., Boivin, J., and Elias, P. (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *The Quarterly Journal of Economics*, 120, 387–422. [10]
- Blasques, F., Koopman, S. J., Lucas, A., and Schaumburg, J. (2016), "Spillover Dynamics for Systemic Risk Measurement Using Spatial Financial Time Series Models," *Journal of Econometrics*, 195, 211–223. [1]
- Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [2,6]
- Chen, J., Li, D., Li, Y.-N., and Linton, O. (2025), "Estimating Time-Varying Networks for High-Dimensional Time Series," *Journal of Econometrics*, 249, 105941. [1]
- Cho, H., and Qu, A. (2013), "Model Selection for Correlated Data with Diverging Number of Parameters," *Statistica Sinica*, 23, 901–927. [4]
- De Paula, A., Rasul, I., and Souza, P. C. (2025), "Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition," *Review of Economic Studies*, 92, 2691–2729. [1]
- Démurger, S. (2001), "Infrastructure Development and Economic Growth: An Explanation for Regional Disparities in China?" *Journal of Comparative Economics*, 29, 95–117. [10]
- Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [3]
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–451. [6]
- Fan, J., Fan, Y., and Lv, J. (2008), "High Dimensional Covariance Matrix Estimation Using a Factor Model," *Journal of Econometrics*, 147, 186–197. [2,3,4,5,10]
- Fan, J., Guo, J., and Zheng, S. (2022), "Estimating Number of Factors by Adjusted Eigenvalues Thresholding," *Journal of the American Statistical Association*, 117, 852–861. [4]
- Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society, Series B*, 74, 37–65. [4,7]
- Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [2,3,6]
- Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical Foundations of Data Science*, New York: Chapman and Hall/CRC. [6]
- Fan, J., and Liao, Y. (2022), "Learning Latent Factors From Diversified Projections and its Applications to Over-Estimated and Weak Factors," *Journal of the American Statistical Association*, 117, 909–924. [2,3,4,5,8]
- Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society, Series B*, 75, 603–680. [4]
- Fan, J., and Lv, J. (2011), "Nonconcave Penalized Likelihood with NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [3,6]
- Fan, J., Wang, W., and Zhu, Z. (2021), "A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery," *Annals of Statistics*, 49, 1239–1266. [6]
- Fujita, M., Krugman, P. R., and Venables, A. (2001), *The Spatial Economy: Cities, Regions, and International Trade*, Cambridge, MA: MIT Press. [1,9]
- Gyourko, J., and Tracy, J. (1991), "The Structure of Local Public Finance and the Quality of Life," *Journal of Political Economy*, 99, 774–806. [9,10]
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent Space Approaches to Social Network Analysis," *Journal of the American Statistical Association*, 97, 1090–1098. [7]
- Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76, 33–50. [7]
- Huang, D., Lan, W., Zhang, H. H., and Wang, H. (2019), "Least Squares Estimation of Spatial Autoregressive models for Large-Scale Social Networks," *Electronic Journal of Statistics*, 13, 1135–1165. [1]
- Huang, D., Zhu, X., Li, R., and Wang, H. (2021), "Feature Screening for Network Autoregression Model," *Statistica Sinica*, 31, 1239–1259. [2]
- Kelejian, H. H., and Prucha, I. R. (1998), "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances," *The Journal of Real Estate Finance and Economics*, 17, 99–121. [1]
- Kenessey, Z. (1987), "The Primary, Secondary, Tertiary and Quaternary Sectors of the Economy," *Review of Income and Wealth*, 33, 359–385. [9,10]
- Lam, C., and Yao, Q. (2012), "Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors," *The Annals of Statistics*, 40, 694–726. [2,3,10]
- Lee, L.-F. (2004), "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models," *Econometrica*, 72, 1899–1925. [1,2,3,4,5,6]
- (2007), "GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models," *Journal of Econometrics*, 137, 489–514. [1]
- Lee, L.-F., Liu, X., and Lin, X. (2010), "Specification and Estimation of Social Interaction Models with Network Structures," *The Econometrics Journal*, 13, 145–176. [4]
- Lee, L.-f., and Yu, J. (2010), "Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects," *Journal of Econometrics*, 154, 165–185. [1,2,3,5,6,9]
- Lunardi, A. (2018), *Interpolation Theory* (Vol. 16), Pisa: Springer. [9]
- Luo, R., Wang, H., and Tsai, C.-L. (2009), "Contour Projected Dimension Reduction," *The Annals of Statistics*, 37, 3743–3778. [10]

- Nowicki, K., and Snijders, T. A. B. (2001), “Estimation and Prediction for Stochastic Blockstructures,” *Journal of the American Statistical Association*, 96, 1077–1087. [7]
- Shao, J. (1993), “Linear Model Selection by Cross-Validation,” *Journal of the American statistical Association*, 88, 486–494. [6,7]
- Su, L. (2012), “Semiparametric GMM Estimation of Spatial Autoregressive Models,” *Journal of Econometrics*, 167, 543–560. [1]
- Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [2,6]
- Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Vol. 48), Cambridge: Cambridge University Press. [3]
- Wang, H. (2012), “Factor Profiled Sure Independence Screening,” *Biometrika*, 99, 15–28. [2,3,4,5,6]
- Wang, H., Li, B., and Leng, C. (2009), “Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters,” *Journal of the Royal Statistical Society, Series B*, 71, 671–683. [2,3,4,6]
- Wang, H., Li, R., and Tsai, C.-L. (2007), “Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method,” *Biometrika*, 94, 553–568. [2,6,7]
- Yang, K., and Lee, L.-f. (2017), “Identification and QML Estimation of Multivariate and Simultaneous Equations Spatial Autoregressive Models,” *Journal of Econometrics*, 196, 196–214. [1,2,3,4,6]
- Yin, J., Li, S., Zhou, L., Jiang, L., and Ma, W. (2022), “Spatial Heterogeneity of the Economic Growth Pattern and Influencing Factors in Formerly Destitute Areas of China,” *Journal of Geographical Sciences*, 32, 829–852. [1,10]
- Yu, J., Zhou, L.-A., and Zhu, G. (2016), “Strategic Interaction in Political Competition: Evidence from Spatial Effects Across Chinese Cities,” *Regional Science and Urban Economics*, 57, 23–37. [1]
- Zhang, X., Xu, G., and Zhu, J. (2022), “Joint Latent Space Models for Network Data with High-Dimensional Node Variables,” *Biometrika*, 109, 707–720. [1]
- Zhang, Y., Wang, J., and Zhang, W. (2024), “Variable Selection and Subgroup Analysis for High-Dimensional Censored Data,” *Statistical Theory and Related Fields*, 8, 211–231. [6]
- Zhou, K., Yang, J., Yang, T., and Ding, T. (2023), “Spatial and Temporal Evolution Characteristics and Spillover Effects of China’s Regional Carbon Emissions,” *Journal of Environmental Management*, 325, 116423. [1,9]
- Zhu, X., Huang, D., Pan, R., and Wang, H. (2020), “Multivariate Spatial Autoregressive Model for Large Scale Social Networks,” *Journal of Econometrics*, 215, 591–606. [1,2,3,7]