AIMS Mathematics, 11(5): 13981–13998.
DOI: 10.3934/math.2026575
Received: 24 February 2026
Revised: 26 April 2026
Accepted: 12 May 2026
Published: 18 May 2026

https://www.aimspress.com/journal/Math

Research article

The profiled variable selection for high-dimensional spatial auto-regressive model

Zengchao Xu and Yu Liu*

Department of Mathematics, Shanghai Normal University, 100 Guilin Road, Shanghai, 200234, China

* **Correspondence:** Email: liuyu2021@shnu.edu.cn.

Abstract: The spatial auto-regressive (SAR) model serves as a foundational tool for investigating spatial dependencies across various scientific fields. In this paper, we proposed a sequential variable selection approach named the profiled variable selection (PVS) procedure for the SAR model. The PVS procedure was designed to efficiently handle high-dimensional cases while maintaining high scalability, even as the number of potential covariates p grows exponentially with the sample size n. However, existing penalization methods are inadequate in variable selection for the SAR model when $p > n$, due to the necessity for a consistent initial estimator. The selection consistency of the PVS procedure was established under mild conditions. Numerical simulations demonstrated the promising performance of the PVS procedure in variable selection for SAR models. Additionally, a real data analysis on housing prices across Chinese cities illustrated the practical utility of our method.

Keywords: spatial auto-regressive model; variable selection; sequential method; partial profile score; high-dimensional data
Mathematics Subject Classification: 62H11, 62J07

1. Introduction

The spatial auto-regressive (SAR) model is fundamental for analyzing data where observations exhibit spatial dependence, a common feature in fields such as economics, environmental science, and social research. These models explicitly account for the interdependence among neighboring units, formalizing the intuitive notion that proximate entities often influence one another. The SAR model is formulated as

$$\boldsymbol{y} = \rho W_0 \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the observed matrix of p-dimensional covariates associated with $\boldsymbol{y}$, and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$ denotes the error vector with unknown variance σ^2. The matrix $W_0 = (w_{ij}) \in \mathbb{R}^{n\times n}$ serves as a pre-specified spatial weight matrix satisfying $w_{ii} = 0$ for all i, and

$\sum_j w_{ij} = 1$ for all $1 \leq i \leq n$. In the SAR model (1.1), $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector of interest, whereas ρ is a nuisance parameter such that $|\rho| < 1$, reflecting the strength of spatial autocorrelation between neighboring observations.

When the dimensionality of covariates is large, the full model, which includes all covariates, is often suboptimal and may even fail to fit. Therefore, it is crucial to identify a reduced model formulated by an appropriate subset of the potential covariates, which helps stabilize coefficient estimation and improve model generalizability. Consequently, variable selection is an indispensable task for identifying the most important covariates.

The existing literature primarily focuses on penalization methods for variable selection in SAR models. These methods introduce a penalty term to the log-likelihood function, effectively shrinking the coefficients of less important covariates toward zero. Research in this area has progressed along several lines. For the standard parametric SAR model (1.1), key works include the adaptive LASSO estimator with established statistical properties [1], the SCAD-penalized quasi-maximum likelihood estimator possessing oracle properties [2], and an instrumental variable-based SCAD method for settings with a diverging number of parameters [3]. More recently, attention has shifted toward semi-parametric spatial frameworks. For semi-parametric varying coefficient models, Luo et al. [4] utilized B-spline approximations coupled with a two-stage least squares and SCAD penalty. Building on this foundation, Liu et al. [5, 6] developed adaptive LASSO penalized estimators via profile maximum likelihood for varying-coefficient spatial panel models with random and fixed effects, respectively, also establishing asymptotic normality and consistency.

Despite their theoretical appeal and successful applications, penalization methods face several notable challenges. First, their implementation often requires solving potentially complex non-convex optimization problems and carefully choosing tuning parameters, which can become computationally intensive, especially in high-dimensional settings. Second, the performance of non-convex penalties, such as the SCAD penalty, typically depends on a consistent initial estimator [2, 3], adding another layer of complexity to the method. Most critically, existing theoretical guarantees for these penalized SAR estimators are primarily confined to scenarios where the number of covariates p is fixed or grows slower than the sample size n. As a consequence, their performance and properties in genuinely high-dimensional contexts, $p > n$, remain largely unexplored, representing a significant gap in the current literature of spatial econometrics.

We focus on the sequential strategy for variable selection in SAR models, which has received limited systematic research in the literature. Sequential methods start with a null model, i.e., $\boldsymbol{\beta} = \mathbf{0}$, and progressively add covariates into the model one at a time. At each step, the covariate that has the most significant importance among the unselected covariates is identified and added into the current model, and this augmented model will be evaluated by a model selection criterion. Examples of sequential methods include forward regression [7], orthogonal matching pursuit [8], and boosting techniques [9]. These methods were developed for linear regression models, in which the importance of the unselected covariates is measured essentially by their Pearson correlation with the residuals of a model fitted on the currently selected covariates. The sequential strategy is particularly attractive for variable selection under high dimensions due to its computational scalability and stable performance. Specifically, it allows for the manageable exploration of large feature spaces without necessitating the exhaustive search of all submodels, significantly reducing computational burden while maintaining the robustness of the selected model, thus making it particularly suitable for complex spatial auto-regressive models.

In this paper, we propose a sequential variable selection procedure for high-dimensional SAR models, referred to as the profiled variable selection (PVS) procedure. In the PVS procedure, a novel measure of importance, the partial profile score, is introduced for unselected covariates. Specifically, the partial profile score for an unselected covariate is defined based on its partial gradient of the likelihood function profiled on the currently selected model. At each step, the covariate with the largest absolute partial profile score is identified as the candidate to be added to the model. The extended Bayesian information criterion serves as the stopping rule to evaluate the updated model during the PVS procedure. The contributions of the proposed PVS procedure for the SAR model are twofold: (1) It extends variable selection for the SAR model to an ultra high-dimensional setting, where the number of covariates p can grow exponentially as the sample size n increases. (2) The selection consistency for the PVS procedure is established rigorously under mild conditions. Both issues have remained unresolved in existing penalization methods. The numerical simulations and empirical data analysis demonstrate that the PVS procedure achieves superior variable selection accuracy compared to existing penalization methods in finite samples, while also maintaining high scalability in high dimensions.

The remainder of the paper is arranged as follows. In Section 2, the partial profile score and the PVS procedure for the SAR model are described. In Section 3, the property of selection consistency of the PVS procedure is presented. Section 4 reports the numerical simulations, and real data analysis is presented in Section 5. Technical details are provided in Appendix A.

2. The PVS procedure for the SAR model

Let $S = \{1, 2, \cdots, p\}$ be the universal index set of all potential covariates in the vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^\top$. Denote by s the index set of selected covariates, which is a subset of S, and denote by $\bar{s}$ the index set of unselected covariates, which is the complementary set of s in S. The notation s also refers to the reduced model with covariates in s, which will be clear from the context. Furthermore, let $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_{\bar{s}}$ be the sub-vectors of $\boldsymbol{\beta}$ with components indexed by s and $\bar{s}$, respectively.

2.1. The partial profile score for the SAR model

Given the observed data $\{X, \boldsymbol{y}\}$, the log-likelihood function (up to an irrelevant constant) of SAR model (1.1) is given by

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}}) = -\frac{n}{2}\log\sigma^2 + \log|A(\rho)| - \frac{||A(\rho)\boldsymbol{y} - X\boldsymbol{\beta}||^2}{2\sigma^2}, \tag{2.1}$$

wherein $\boldsymbol{\theta} = (\rho, \sigma^2, \boldsymbol{\beta})$, $\boldsymbol{\theta}_s = (\rho, \sigma^2, \boldsymbol{\beta}_s)$, and $A(\rho) = I_n - \rho W_0$. The log-likelihood with respect to $\boldsymbol{\beta}_{\bar{s}}$, profiled on the selected model s, is given by

$$\widetilde{\ell}(\boldsymbol{\beta}_{\bar{s}}) = \sup_{\boldsymbol{\theta}_s} \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}}) = \ell(\widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}}), \boldsymbol{\beta}_{\bar{s}}),$$

where $\widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}})$ denotes the maximizer of $\boldsymbol{\theta}_s$ for fixed $\boldsymbol{\beta}_s$. Given the selected model s, for an unselected covariate x_j, $j \in \bar{s}$, we define its partial profile score as

$$\psi(j \mid s) = \frac{\partial\widetilde{\ell}(\boldsymbol{\beta}_{\bar{s}})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}_{\bar{s}}=\mathbf{0}}. \tag{2.2}$$

Intuitively, a covariate x_j with larger absolute partial profile score will increase the level of log-likelihood more significantly when it is added into the current model s.

Note the fact that $\widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}})$ is the maximizer of $\ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}})$ with fixed $\boldsymbol{\beta}_{\bar{s}}$, implying that $\frac{\partial \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}})}{\partial \boldsymbol{\theta}_s}\Big|_{\boldsymbol{\theta}_s = \widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}})} = 0$. As a consequence, we have that, for each $j \in \bar{s}$,

$$\frac{\partial \widetilde{\ell}(\boldsymbol{\beta}_{\bar{s}})}{\partial \beta_j} = \Big(\frac{\partial \widetilde{\boldsymbol{\theta}}_s^{\top}}{\partial \beta_j}\frac{\partial \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}})}{\partial \boldsymbol{\theta}_s} + \frac{\partial \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}})}{\partial \beta_j}\Big)\Big|_{\boldsymbol{\theta}_s = \widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}})} = \frac{\partial \ell(\boldsymbol{\theta}_s, \boldsymbol{\beta}_{\bar{s}})}{\partial \beta_j}\Big|_{\boldsymbol{\theta}_s = \widetilde{\boldsymbol{\theta}}_s(\boldsymbol{\beta}_{\bar{s}})}.$$

In the definition of the partial profile score, the value of $\boldsymbol{\beta}_{\bar{s}}$ is set as $\mathbf{0}$, thus the profiled maximizer $\widetilde{\boldsymbol{\theta}}_s(\mathbf{0})$ is exactly the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}_s$ for the reduced model s. Denote $\widetilde{\boldsymbol{\theta}}_s(\mathbf{0})$ by $\widehat{\boldsymbol{\theta}}_s$ for notation simplicity. The partial profile score is equivalently given by

$$\psi(j \mid s) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j}\Big|_{\boldsymbol{\theta} = (\widehat{\boldsymbol{\theta}}_s, \mathbf{0})} = \frac{1}{\widehat{\sigma}^2}(A(\widehat{\rho})\mathbf{y} - X_s\widehat{\boldsymbol{\beta}}_s)^{\top}\boldsymbol{x}_j,$$

where $\widehat{\boldsymbol{\theta}}_s = (\widehat{\rho}, \widehat{\sigma}^2, \widehat{\boldsymbol{\beta}}_s)$, X_s is the sub-matrix of X with columns indexed in s, and $\boldsymbol{x}_j$ is the column vector of X corresponding to β_j.

The MLE $\widehat{\boldsymbol{\theta}}_s$ of the reduced model s can be obtained through a dynamic programming approach. Specifically, given ρ, maximizing the log-likelihood function (2.1) yields the closed-form of the MLE for $(\boldsymbol{\beta}_s, \sigma^2)$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_s(\rho) &= (X_s^{\top}X_s)^{-1}X_s^{\top}A(\rho)\mathbf{y}, \\ \widehat{\sigma}^2(\rho) &= \frac{1}{n}\|A(\rho)\mathbf{y} - X_s\widehat{\boldsymbol{\beta}}_s(\rho)\|^2. \end{aligned} \tag{2.3}$$

Plugging (2.3) into the log-likelihood function (2.1), the MLE of ρ can be efficiently obtained by some standard numerical optimization routine for the univariate function with bounded domain $\rho \in (-1, 1)$, and then $\widehat{\boldsymbol{\beta}}_s(\rho)$ and $\widehat{\sigma}^2(\rho)$ can be estimated recursively using Eq (2.3).

2.2. The PVS procedure

As a sequential approach, the PVS procedure starts with the null model, i.e., $s = \emptyset$, which corresponds to initialization $\boldsymbol{\beta} = \mathbf{0}$. At each step of the PVS procedure, the covariate, denoted by x_{j^*}, with the largest absolute partial profile score is identified to be added into the currently selected model s, yielding an augmented model $s^+ = s \cup \{j^*\}$. We employ the extended Bayesian information criterion (EBIC, [10]) to evaluate the saturability and complexity of the models during the PVS procedure. The EBIC for model s is defined as

$$\mathrm{EBIC}_{\gamma}(s) = -2\ell(\widehat{\boldsymbol{\theta}}_s) + |s|\log n + 2\gamma\log\binom{p}{|s|}, \tag{2.4}$$

where $|s|$ denotes the number of selected features in s, and $\gamma = \max\{1 - \log n/(2\log p), 0\}$. If the augmented model reduces the level of the EBIC compared with the current model, i.e., $\mathrm{EBIC}_{\gamma}(s^+) < \mathrm{EBIC}_{\gamma}(s^*)$, the current model is updated by the augmented model and the process is iterated; otherwise, the PVS procedure stops.

The sequential strategy has a well-known greedy nature, meaning that it prefers to select more spurious covariates beyond the true ones. The EBIC extends the classical BIC by adding an extra

penalty term $2\gamma \log \binom{p}{|s|}$ on the size of the model, thus encouraging a parsimonious model. Moreover, the EBIC has a clear interpretation from the Bayesian regime.

By integrating the partial profile score (2.2) for selecting the covariate and EBIC (2.4) as the stopping rule, we show that the PVS procedure enjoys the property of selection consistency in the high-dimensional SAR model, as discussed in Section 3. The PVS procedure for the SAR model is summarized as follows:

(1) Initialize $s^* = \emptyset$ (or a set of pre-selected covariates).

(2) Identify the covariate with the largest absolute partial profile score among the unselected features, i.e.,
$$j^* = \arg\max_{j \in \overline{s^*}} |\psi(j \mid s^*)|.$$

(3) The current model s^* is tentatively augmented as $s^+ = s^* \cup \{j^*\}$, which is fitted by maximizing the reduced log-likelihood function:
$$\max_{\boldsymbol{\theta}_{s^+}} \ell(\boldsymbol{\theta}_{s^+}, \mathbf{0}).$$

(4) If $\mathrm{EBIC}_\gamma(s^+) < \mathrm{EBIC}_\gamma(s^*)$, update $s^* = s^+$ and return to the selecting rule; otherwise, stop the PVS procedure and output s^* as the set of selected covariates.

An R package `pboost4sar` (https://github.com/profile-boosting/pboost4sar) is available for the implementation of the PVS procedure for the SAR model.

In the PVS procedure, it requires at most $O(n^2 \log \varepsilon^{-1} + np)$ computation operations, where ε is the pre-specified accuracy for the numerical solution of ρ. Specifically, to achieve an ε-optimal estimation for ρ for a given X_s, a numerical solver requires at most $O(\log \varepsilon^{-1})$ iterations and $O(n^2 + np_0^2)$ operations in each update, where p_0 is the maximum number of selected variables in any model s during the PVS procedure. Given $\hat{\rho}$, the complexities for the MLE of β and σ^2 are, respectively, $O(np_0^2 + p_0^3)$ and $O(np_0 + n^2)$. In addition, the operations for partial profile scores of all unselected features are at most $O(n^2 + np)$.

3. Selection consistency of the PVS procedure

Denote the true coefficient vector by $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \cdots, \beta_{0p})^\top$ in the SAR model. Let $s_0 = \{j : |\beta_{0j}| \neq 0\}$ be the index set of the true model, and $p_0 = |s_0|$ be the number of true variables. For the reduced model consisting of covariates in s, we define
$$\bar{\boldsymbol{\theta}}_s = \arg\max_{\boldsymbol{\theta}_s} \mathrm{E}\left[\ell(\boldsymbol{\theta}_s, \mathbf{0}_{\bar{s}})\right].$$

Let $K > 1$ be a generic constant. The following assumptions are imposed.

(C1) $p_0 \ln p = O(n^\kappa)$ for some $0 < \kappa < 1/4$, and $p_0 = o(n^{1/6})$.

(C2) The elements in X are bounded above by K. For every s with $|s| \leq Kp_0$, it holds that
$$K^{-1}n \leq \lambda_{\min}(X_s^\top X_s) \leq \lambda_{\max}(X_s^\top X_s) \leq Kn.$$

(C3) There exists some constant $\alpha > 0$ such that
$$\inf_{s \subset s_0} \max_{j \in s_0 - s} \left| \frac{1}{n}(\boldsymbol{\eta}_0 - X_s \bar{\boldsymbol{\beta}}_s) \boldsymbol{X}_j \right| > Kn^{-\alpha},$$
$$\max_{j \in s_0^c} \left| \frac{1}{n}(\boldsymbol{\eta}_0 - X_s \bar{\boldsymbol{\beta}}_s) \boldsymbol{X}_j \right| = o(n^{-\alpha})$$

with $p_0 n^{2\alpha-1/4} \to 0$ as $n \to \infty$.

(C4) As $n \to \infty$, $\inf_{1\le k<p_0} \{n^{1/4}\|\breve{\boldsymbol{\beta}}_{s_k} - \breve{\boldsymbol{\beta}}_{s_{k+1}}\|^2 : s_k \subset s_{k+1} \subset s_0, |s_k| = k\} \to \infty$ wherein $\breve{\boldsymbol{\beta}}_s = (\bar{\boldsymbol{\beta}}_s, \mathbf{0})$.

The selection consistency of the PVS procedure for the SAR model is established by the following theorem.

Theorem 1. *Let s^* be the index set of covariates identified by the PVS procedure. Under assumptions, we have*

$$\lim_{n\to\infty} \mathrm{P}(s^* = s_0) = 1.$$

The proof of Theorem 1 is provided in Appendix A. The assumptions (C1)–(C4) are mild conditions to establish the selection consistency for the PVS procedure. Assumption (C1) allows the dimension of the covariates under consideration to have an exponential magnitude relative to the sample size and the number of relevant features can diverge as $n \to \infty$, which shows that the PVS procedure is applicable to the ultra high-dimensional SAR model. Assumption (C2) is commonly used in variable selection, and a similar condition can be found in [2]. Specifically, in the PVS procedure, it guarantees that $\ell(\boldsymbol{\theta}_s)$ is concave around $\bar{\boldsymbol{\theta}}_s$, thus the MLE for a reduced model in Eq (2.3) exists and is unique. Assumption (C3) imposes a signal strength restriction on the partial profile score for unselected true covariates, which is indeed a variant of the irrepresentability condition for model identifiability [11]. Assumption (C4) is a weak condition since the numbers of non-zero components in $\breve{\boldsymbol{\beta}}_{s_k}$ and $\breve{\boldsymbol{\beta}}_{s_{k+1}}$ are different [12].

4. Numerical simulations

In this section, we conduct extensively numerical simulations to compare the performance of the proposed PVS procedure and the penalization methods, including the adaptive lasso (aLASSO) penalty [1] and the SCAD penalty [2]. To evaluate the advantages of the proposed PVS method within a sequential strategy, we take the forward regression selection (FRS) method in [7] as a benchmark. The FRS was originally developed for linear models, and has been shown to achieve selection consistency under proper conditions.

Throughout the simulations, the sample size is fixed at $n = 200$, and the spatial weight matrix $W_0 = (w_{ij}) \in \mathbb{R}^{n\times n}$ between the subjects is set as one of the following two modes: (1) *Case*: $W_0 = I_{10} \otimes B_{20}$ with $B_m = \frac{1}{m-1}(\mathbf{1}_m^\top \mathbf{1}_m - I_m)$. (2) *Rook*: $w_{ij} = 1$ for the contiguous i-th and j-th subjects which locates in a 10×20 grid; otherwise, $w_{ij} = 0$. The spatial weight matrix is row standardized when it is used to generate the simulation data. The spatial auto-regressive coefficient is taken as $\rho_0 = 0.2, 0.5$, or 0.8.

The dimensions of potential covariates are taken from $p \in \{50, 100, 200, 500, 1000, 3000\}$, and the covariate vector is simulated by $\boldsymbol{x} \sim N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (0.7^{|i-j|})$. Furthermore, the true set of covariates related to the response variable is fixed as the first ten covariates:

$$s_0 = \{1, 2, \cdots, 10\},$$

and the corresponding coefficients are generated as $\beta_j \sim \mathrm{Unif}(1.5, 2.0)$ for $1 \le j \le 10$, thus the true coefficient vector is given by $\boldsymbol{\beta}_0 = (\beta_1, \cdots, \beta_{10}, \mathbf{0}_{p-10}^\top)^\top$. The vector of dependent variables is generated by

$$\boldsymbol{y} = (I_n - \rho_0 W_0)^{-1}(X\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}),$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma_0^2 I_n)$, and the standard derivation is taken as $\sigma_0 = 2$ in the cases where $p \leq n$, and as $\sigma_0 = 1$ in high-dimensional cases, $p > n$.

Let $\widehat{\rho}$ and $\widehat{\sigma}$ be the estimation of ρ and σ in the SAR model, and $\widehat{s}$ be the set of selected covariates by the considered variable selection methods. The performance of a variable selection method is assessed by the positive discovery rate (PDR) and the false discovery rate (FDR):

$$\text{PDR} = \frac{|\widehat{s} \cap s_0|}{|s_0|}, \ \text{FDR} = \frac{|\widehat{s} \setminus s_0|}{|\widehat{s}|}.$$

A total of 12 configurations from the above settings are simulated in the numerical studies. For each configuration, we perform 100 independent replications for both the PVS method and penalization methods, reporting the averaged values of $\widehat{\rho}$, $\widehat{\sigma}$, PDR, and FDR. The results are summarized in Tables 1 and 2. Table 1 summarizes the results of low-dimensional case ($p < n$) and edge case ($p = n$), and Table 2 corresponds to the high-dimensional case ($p > n$).

For the sequential strategies, the PDR and FDR of the PVS and FRS are generally comparable. In the cases where $p \leq n$, the PDRs are all greater than 95%, while the FDRs are controlled under 10%; see Table 1. In the high-dimensional cases reported in Table 2, the patterns of variable selection are similar to that under low-dimensional settings. Moreover, the PVS procedure has a superior PDR under the W_0 of *Rook*.

In the case where $p \leq n$, as shown in Table 1, the PVS method demonstrates near-optimal performance in variable selection. It consistently achieves a PDR close to one while controlling the FDR effectively at zero across all spatial configurations W_0 and auto-regressive coefficients ρ_0. In contrast, the penalization methods exhibit distinct limitations. Specifically, the adaptive LASSO penalty results in a sharply rising FDR although it has a comparable or slightly higher PDR compared with the PVS as p approaches n, which indicates it tends to identify many spurious variables. The SCAD penalty has a clear downward trend in the PDR accompanied by a marked upward trend in the FDR as p grows, reflecting its deteriorating selection accuracy and reliability under these settings. The estimates for ρ and σ from all methods are stable with small standard errors in most of the settings, confirming reliable estimation in this low-dimensional regime.

In the case where $p = n = 200$, it is worth noting that the auto-regressive coefficient ρ in both adaptive LASSO and SCAD penalties is estimated as 1 in all configurations of W_0 and ρ_0, demostrating that they fail to identify a valid model when the number of covariates approaches to the sample size. These results underscore that PVS provides precise and clean variable selection without sacrificing power when the number of covariates does not exceed the sample size.

When $p > n$, the adaptive LASSO and SCAD penalties suffer from numerical instability in estimating the spatial autocorrelation parameter ρ, thereby impeding their penalization procedures. To ensure comparability, we adopt the $\widehat{\rho}$ estimated from the PVS method as a common input for all approaches. As illustrated in Table 2, all three methods achieve reasonably high PDRs, with PVS and the adaptive LASSO demonstrating particularly stable performance—exhibiting both high mean PDR and low standard deviations. In contrast, SCAD yields the lowest PDR alongside the highest variability. Regarding false discovery control, adaptive LASSO incurs a substantially elevated FDR, while both PVS and SCAD maintain lower levels, with PVS being only marginally higher than SCAD. Furthermore, PVS provides the most accurate and stable estimate of the error variance σ, characterized by the smallest bias and standard deviation. Collectively, these results indicate that in ultrahigh-dimensional settings, PVS delivers superior accuracy and robustness in variable selection.

Table 1. Average PDR and FDR with $p \leq n = 200$ and $\sigma_0 = 2$ (standard deviations in parentheses).

			$p = 50$				$p = 100$				$p = 200$			
W_0	ρ_0		**PVS**	FRS	aLASSO	SCAD	**PVS**	FRS	aLASSO	SCAD	**PVS**	FRS	aLASSO	SCAD
Case	0.2	$\hat{\rho}$	0.19(0.05)	0.19(0.05)	0.19(0.06)	0.19(0.06)	0.19(0.05)	0.20(0.04)	0.19(0.06)	0.19(0.06)	0.19(0.05)	0.19(0.04)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.11)	1.93(0.13)	1.87(0.10)	1.93(0.11)	1.95(0.16)	1.93(0.09)	1.79(0.09)	1.93(0.11)	1.98(0.24)	1.92(0.12)	0.19(0.09)	3.33(0.41)
		PDR	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.08)	1.00(0.00)	1.00(0.00)	1.00(0.02)	0.94(0.11)	1.00(0.00)	0.96(0.06)	0.74(0.14)
		FDR	0.06(0.07)	0.05(0.07)	0.41(0.07)	0.08(0.10)	0.07(0.08)	0.05(0.07)	0.62(0.05)	0.15(0.13)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.26(0.16)
	0.5	$\hat{\rho}$	0.49(0.03)	0.50(0.03)	0.49(0.04)	0.49(0.04)	0.49(0.03)	0.50(0.03)	0.49(0.04)	0.49(0.04)	0.49(0.03)	0.49(0.02)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.11)	1.93(0.13)	1.87(0.10)	1.93(0.11)	1.95(0.16)	1.93(0.09)	1.79(0.09)	1.93(0.12)	1.98(0.24)	1.92(0.12)	0.19(0.09)	3.27(0.40)
		PDR	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.08)	1.00(0.00)	1.00(0.00)	1.00(0.02)	0.94(0.11)	1.00(0.00)	0.96(0.06)	0.74(0.14)
		FDR	0.06(0.07)	0.05(0.07)	0.41(0.07)	0.08(0.10)	0.07(0.08)	0.05(0.07)	0.62(0.05)	0.15(0.13)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.26(0.16)
	0.8	$\hat{\rho}$	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.02)	0.80(0.02)	0.80(0.01)	0.80(0.01)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.11)	1.93(0.13)	1.87(0.10)	1.93(0.11)	1.95(0.16)	1.93(0.09)	1.79(0.09)	1.93(0.12)	1.98(0.24)	1.92(0.12)	0.19(0.09)	3.22(0.39)
		PDR	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.08)	1.00(0.00)	1.00(0.00)	1.00(0.02)	0.94(0.11)	1.00(0.00)	0.96(0.06)	0.74(0.14)
		FDR	0.06(0.07)	0.05(0.07)	0.41(0.07)	0.08(0.10)	0.07(0.08)	0.05(0.07)	0.62(0.05)	0.15(0.13)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.26(0.16)
Rook	0.2	$\hat{\rho}$	0.20(0.02)	0.20(0.02)	0.20(0.03)	0.20(0.03)	0.20(0.02)	0.20(0.02)	0.20(0.03)	0.20(0.03)	0.20(0.03)	0.19(0.03)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.12)	1.93(0.13)	1.87(0.10)	1.93(0.11)	1.94(0.15)	1.93(0.10)	1.79(0.09)	1.93(0.11)	1.98(0.24)	1.91(0.12)	0.24(0.16)	5.49(0.50)
		PDR	0.99(0.03)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.07)	1.00(0.00)	1.00(0.00)	0.99(0.02)	0.94(0.12)	1.00(0.00)	0.97(0.05)	0.53(0.10)
		FDR	0.06(0.07)	0.05(0.07)	0.41(0.07)	0.08(0.10)	0.08(0.08)	0.05(0.07)	0.62(0.05)	0.15(0.13)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.26(0.20)
	0.5	$\hat{\rho}$	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.02)	0.50(0.03)	0.49(0.02)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.11)	1.93(0.13)	1.87(0.10)	1.93(0.10)	1.94(0.14)	1.93(0.10)	1.79(0.09)	1.93(0.12)	1.98(0.23)	1.92(0.12)	0.20(0.12)	4.29(0.41)
		PDR	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.99(0.03)	0.94(0.11)	1.00(0.00)	0.97(0.05)	0.60(0.12)
		FDR	0.06(0.07)	0.05(0.07)	0.42(0.07)	0.09(0.10)	0.08(0.08)	0.05(0.07)	0.62(0.05)	0.14(0.12)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.27(0.19)
	0.8	$\hat{\rho}$	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	0.80(0.01)	1.00(0.00)	1.00(0.00)
		$\hat{\sigma}$	1.93(0.11)	1.93(0.13)	1.87(0.10)	1.93(0.10)	1.95(0.15)	1.93(0.10)	1.79(0.09)	1.93(0.12)	1.97(0.22)	1.92(0.12)	0.19(0.08)	3.24(0.30)
		PDR	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.07)	1.00(0.00)	1.00(0.00)	0.99(0.02)	0.94(0.10)	1.00(0.00)	0.97(0.05)	0.70(0.13)
		FDR	0.06(0.07)	0.05(0.07)	0.41(0.07)	0.09(0.10)	0.08(0.08)	0.05(0.07)	0.62(0.05)	0.14(0.12)	0.10(0.10)	0.08(0.09)	0.95(0.00)	0.24(0.18)

Table 2. Average PDR and FDR with $p > n = 200$ and $\sigma_0 = 1$ (standard deviations in parentheses).

			$p = 500$				$p = 1000$				$p = 3000$			
W_0	ρ_0		**PVS**	FRS	aLASSO	SCAD	**PVS**	FRS	aLASSO	SCAD	**PVS**	FRS	aLASSO	SCAD
Case	0.2	$\hat{\rho}$	0.20(0.03)	0.20(0.03)	/	/	0.20(0.03)	0.20(0.02)	/	/	0.20(0.03)	0.19(0.03)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.93(0.06)	0.97(0.05)	1.58(0.25)	0.93(0.07)	0.93(0.06)	0.46(0.43)	1.59(0.24)	0.95(0.18)	0.90(0.07)	0.36(0.41)	1.62(0.25)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.09)
		FDR	0.10(0.09)	0.12(0.09)	0.14(0.07)	0.01(0.04)	0.11(0.09)	0.11(0.10)	0.57(0.42)	0.03(0.05)	0.14(0.12)	0.14(0.11)	0.68(0.40)	0.07(0.09)
	0.5	$\hat{\rho}$	0.50(0.02)	0.50(0.02)	/	/	0.50(0.02)	0.50(0.02)	/	/	0.50(0.02)	0.50(0.02)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.93(0.06)	0.97(0.05)	1.58(0.25)	0.93(0.07)	0.93(0.06)	0.46(0.43)	1.59(0.24)	0.95(0.18)	0.90(0.07)	0.36(0.41)	1.62(0.25)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.09)
		FDR	0.10(0.09)	0.12(0.09)	0.14(0.07)	0.01(0.04)	0.11(0.09)	0.11(0.10)	0.57(0.42)	0.03(0.05)	0.14(0.12)	0.14(0.11)	0.68(0.40)	0.07(0.09)
	0.8	$\hat{\rho}$	0.80(0.01)	0.80(0.01)	/	/	0.80(0.01)	0.80(0.01)	/	/	0.80(0.01)	0.80(0.01)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.93(0.06)	0.97(0.05)	1.58(0.25)	0.93(0.07)	0.93(0.06)	0.46(0.43)	1.59(0.25)	0.95(0.18)	0.90(0.07)	0.36(0.41)	1.62(0.25)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.09)
		FDR	0.10(0.09)	0.12(0.09)	0.14(0.07)	0.01(0.04)	0.11(0.09)	0.11(0.10)	0.57(0.42)	0.03(0.05)	0.14(0.12)	0.14(0.11)	0.68(0.40)	0.07(0.09)
Rook	0.2	$\hat{\rho}$	0.20(0.01)	0.20(0.01)	/	/	0.20(0.01)	0.20(0.01)	/	/	0.20(0.02)	0.20(0.01)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.94(0.06)	0.97(0.05)	1.58(0.25)	0.93(0.06)	0.92(0.07)	0.44(0.42)	1.60(0.25)	0.96(0.18)	0.90(0.09)	0.41(0.42)	1.61(0.26)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.09)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.09)
		FDR	0.10(0.09)	0.10(0.09)	0.14(0.08)	0.01(0.04)	0.11(0.09)	0.11(0.10)	0.60(0.42)	0.03(0.05)	0.12(0.11)	0.14(0.13)	0.64(0.41)	0.07(0.09)
	0.5	$\hat{\rho}$	0.50(0.01)	0.50(0.01)	/	/	0.50(0.01)	0.50(0.01)	/	/	0.50(0.01)	0.50(0.01)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.94(0.06)	0.97(0.05)	1.58(0.26)	0.93(0.06)	0.92(0.07)	0.40(0.42)	1.59(0.25)	0.96(0.18)	0.90(0.09)	0.38(0.41)	1.61(0.26)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.09)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.10)
		FDR	0.10(0.09)	0.10(0.09)	0.14(0.07)	0.02(0.04)	0.11(0.09)	0.12(0.11)	0.64(0.41)	0.03(0.06)	0.13(0.12)	0.15(0.14)	0.66(0.40)	0.07(0.09)
	0.8	$\hat{\rho}$	0.80(0.01)	0.80(0.01)	/	/	0.80(0.01)	0.80(0.01)	/	/	0.80(0.01)	0.80(0.01)	/	/
		$\hat{\sigma}$	0.94(0.06)	0.94(0.06)	0.97(0.05)	1.58(0.25)	0.93(0.06)	0.92(0.07)	0.44(0.42)	1.59(0.25)	0.96(0.18)	0.91(0.09)	0.36(0.41)	1.61(0.26)
		PDR	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.09)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.08)	0.98(0.07)	1.00(0.00)	1.00(0.00)	0.92(0.09)
		FDR	0.09(0.09)	0.10(0.09)	0.14(0.07)	0.02(0.04)	0.11(0.10)	0.12(0.11)	0.60(0.41)	0.03(0.06)	0.13(0.12)	0.14(0.13)	0.68(0.39)	0.07(0.09)

The computation times of all compared methods are summarized in Table 3. The time cost of PVS remains stable as the dimension p grows, and the SCAD penalty seems also scalable to the dimensionality. However, the SCAD-penalized SAR model is not applicable to the high-dimensional case directly since it requires a consistent estimator of ρ initially. In simulations, the estimation of ρ in penalized methods for high-dimensional settings is initialized as the estimation of PVS. As a sequential method, the FRS takes much more time than PVS, and its time cost grows quickly when p increases. The adaptive LASSO penalty also requires significantly increasing time when p diverges. Therefore, both the FRS and adaptive LASSO are computationally infeasible for large-scale spatial auto-regression.

Table 3. Average time (in seconds) with varying p.

Method	50	100	200	500	1000	3000
PVS	0.45	0.45	0.45	0.45	0.46	0.47
FRS	15.50	34.64	72.48	191.36	387.02	1219.84
aLASSO	4.33	15.29	173.55	207.43	346.81	499.36
SCAD	0.05	0.08	0.18	0.48	0.50	0.60

5. Real data analysis

In this section, we conduct an empirical analysis on the house price data from Chinese cities to validate the performance of the PVS method against penazlization methods, including adaptive LASSO and SCAD penalties.

5.1. Data and preprocessing

The dataset comprises cross-sectional observations from $n = 35$ major Chinese cities in the year 2020. The response variable y is the average residential housing sales price (in yuan per square meter). A set of $p = 24$ potential covariates were considered, capturing economic, demographic, social, and land-market factors hypothesized to influence urban housing prices. The variables, listed in Table 4, were all scaled as zero mean and unit variance. This standardization ensures that the estimated coefficients are comparable and reduces potential issues related to different measurement scales.

Table 4. Description of variables in the housing price analysis.

Variable	Description
y	Average residential housing sales price
x_1	Per capita gross domestic product (GDP)
x_2	Per capita disposable income
x_3	Consumer Price Index (previous year = 100)
x_4	Registered household population (in 10,000 persons)
x_5	Permanent resident population (in 10,000 persons)
x_6	Year-end employed persons in urban units
x_7	Average wage of on-post workers
x_8	Floor space of commercial residential buildings for sale (10,000 m^2)
x_9	Purchased land area (10,000 m^2)
x_{10}	Floor space under construction (10,000 m^2)
x_{11}	Completed floor space (10,000 m^2)
x_{12}	Investment in residential real estate development (100 million yuan)
x_{13}	Local general public budget revenue (100 million yuan)
x_{14}	Local general public budget expenditure (100 million yuan)
x_{15}	Year-end balance of RMB loans of financial institutions (100 million yuan)
x_{16}	Year-end balance of RMB deposits of financial institutions (100 million yuan)
x_{17}	Balance of RMB household savings deposits (100 million yuan)
x_{18}	Number of institutions of higher education
x_{19}	Number of hospitals
x_{20}	Public green area (hectares)
x_{21}	Land supply area (10,000 m^2)
x_{22}	Land transaction area (10,000 m^2)
x_{23}	Transaction price of land (100 million yuan)
x_{24}	Average transaction price of land (yuan/m^2)

5.2. Spatial weight matrix and model specification

A spatial weight matrix W is constructed based on the geographical coordinates (latitude and longitude) of the 35 city centers. The matrix follows the common *inverse distance* specification, where the weight between city i and city j ($i \neq j$) is defined as $w_{ij} = 1/d_{ij}$, with d_{ij} denoting the great-circle distance in kilometers. The diagonal elements are set to zero ($w_{ii} = 0$), and the matrix is row-standardized to ensure $\sum_j w_{ij} = 1$, facilitating the interpretation of the spatial lag term.

We conduct variable selection and model fitting to the SAR model (1.1) for the house price data, where $\mathbf{y} \in \mathbb{R}^{35}$ is the vector of standardized housing prices, $X \in \mathbb{R}^{35\times 24}$ is the matrix of standardized covariates, and W is the row-standardized inverse-distance spatial weight matrix. The objective is to identify the most relevant economic drivers from the 24 candidates while accounting for spatial spillover effects.

5.3. Variable selection results and discussion

We apply the proposed PVS procedure, implemented via the `pboost4sar` package, to this dataset. For comparison, we also implement the forward regression selection (FRS) method for the SAR model [7], and the adaptive LASSO and SCAD-penalized SAR models [1,2]. All methods use the same spatial weight matrix W. The selected variables and corresponding refitted coefficients are summarized in Table 5.

The results in Table 5 reveal clear differences among the four variable selection methods. The key findings are summarized as follows: (1) PVS procedure and SCAD penalty select the most parsimonious model with only four variables, while the FRS procedure and adaptive LASSO penalty select seven and nine variables. This suggests PVS offers stronger sparsity and better interpretability for moderate-dimensional spatial data. (2) PVS identifies four economically meaningful drivers: Local budget revenue (x_{13}, $\widehat{\beta} = 0.6572$), average land price (x_{24}, $\widehat{\beta} = 0.5819$), housing inventory (x_8, $\widehat{\beta} = -0.249$), and number of universities (x_{18}, $\widehat{\beta} = -0.1306$). These variables represent fundamental supply-demand and fiscal factors affecting housing prices. (3) PVS provides larger coefficient estimates for strong signals (e.g., x_{13}: 0.6572) compared to adaptive LASSO (0.0851), suggesting less shrinkage bias.

Table 5. Selected variables and the refitted coefficients for Chinese city housing price data (variables not selected by any method are omitted, and the values in parentheses are the standard errors).

Variable	PVS	FRS	Adaptive LASSO	SCAD
x_2 (Per capita income)	–	0.190 (0.075)	0.132 (0.079)	–
x_8 (Space for sale)	-0.249 (0.070)	-0.179 (0.051)	-0.117 (0.057)	-0.249 (0.070)
x_{10} (Under construction)	–	–	-0.061 (0.054)	–
x_{13} (Budget revenue)	0.657 (0.079)	0.145 (0.114)	0.227 (0.119)	0.657 (0.079)
x_{15} (Loans)	–	0.644 (0.138)	0.818 (0.158)	–
x_{17} (Savings)	–	–	-0.274 (0.141)	–
x_{18} (Universities)	-0.131 (0.050)	-0.181 (0.052)	-0.133 (0.055)	-0.131 (0.050)
x_{22} (Trans. area)	–	–	–	–
x_{23} (Trans. price)	–	-0.280 (0.063)	-0.272 (0.067)	–
x_{24} (Avg. land price)	0.582 (0.057)	0.459 (0.050)	0.482 (0.052)	0.582 (0.057)

To assess the prediction accuracy of the four methods, we have conducted leave-one-out cross-validation on the house price data. Specifically, one of the $n = 35$ cities is dropped out, and the other data are applied to the four methods. As a result, the significant variables identified by a method are used to fit the SAR model, and the fitted model is then used to predict the house price of the dropped-out sample. This procedure is repeated on each of the datasets, and the root mean squared error for the four methods are reported in Table 6. Compared with the other methods, PVS has the smallest prediction error, which is significantly lower than those of the others.

In summary, the PVS method demonstrates practical utility by producing a sparse, interpretable, and efficient model for spatial housing price analysis, reinforcing its advantages observed in simulation studies.

Table 6. Root mean square error of prediction error estiamted by the leave-one-out method.

Method	PVS	FRS	Adaptive LASSO	SCAD
RMSE	0.3705	0.5534	0.4446	0.4876

6. Conclusions

In this paper, we propose a sequential variable selection method, termed profiled variable selection (PVS), for SAR models. A novel measure, the partial profile score, is introduced to evaluate the importance of unselected covariates. The selection consistency of the PVS procedure is also established. Both numerical simulations and real data analysis demonstrate that the PVS method is favorable in high-dimensional variable selection, in which the existing penalization approaches fails due to the degenerated estimation of the spatial correlation coefficient.

As a sequential method, a great advantage of PVS is that it is extremely computationally efficient for ultra high-dimensional settings compared with the naive counterpart of forward regression selection. In each selection step of the PVS procedure, only one low-dimensional model is fitted, and the computation of the partial profile score for the unselected variables does not involve any model fitting. However, for forward regression selection, a model must be fitted for each of the unselected variables at each step, which entails a heavy computational burden even when p diverges.

Noting that the definition of the partial profile score is applicable to the general differentiable loss function, the PVS procedure offers a computationally efficient and theoretically sound framework for variable selection in high-dimensional spatial econometrics, and it is readily extended to generally spatial models, such as spatial panel model. These extensions will be considered elsewhere in future research.

Author contributions

Zengchao Xu: Methodology, formal analysis, software, validation, writing-original draft; Yu Liu: Data curation, writing-review & editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 12571310).

Conflict of interest

The authors declare no potential conflicts of interest.

References

1. Y. Wu, Y. Sun, Shrinkage estimation of the linear model with spatial interaction, *Metrika*, **80** (2017), 51–68. https://doi.org/10.1007/s00184-016-0590-z
2. X. Liu, J. Chen, S. Cheng, A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model, *Spat. Stat.*, **25** (2018), 86–104. https://doi.org/10.1016/j.spasta.2018.05.001
3. T. Xie, R. Cao, J. Du, Variable selection for spatial autoregressive models with a diverging number of parameters, *Stat. Pap.*, **61** (2020), 1125–1145. https://doi.org/10.1007/s00362-018-0984-2
4. G. Luo, M. Wu, Variable selection for semiparametric varying-coefficient spatial autoregressive models with a diverging number of parameters, *Commun. Stat.-Theor. M.*, **50** (2021), 2062–2079. https://doi.org/10.1080/03610926.2019.1659367
5. Y. Liu, X. Zhuang, Shrinkage estimation of semi-parametric spatial autoregressive panel data model with fixed effects, *Stat. Probabil. Lett.*, **194** (2023), 109746. https://doi.org/10.1016/j.spl.2022.109746
6. Y. Liu, Adaptive lasso variable selection method for semiparametric spatial autoregressive panel data model with random effects, *Commun. Stat.-Theor. M.*, **53** (2024), 2122–2140. https://doi.org/10.1080/03610926.2022.2119088
7. H. Wang, Forward regression for ultra-high dimensional variable screening, *J. Am. Stat. Assoc.*, **104** (2009), 1512–1524. https://doi.org/10.1198/jasa.2008.tm08516
8. Y. C. Pati, R. Rezaiifar, P. S. Krishnaprasad, *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*, Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, 1993, 40–44. https://doi.org/10.1109/ACSSC.1993.342465
9. P. Bühlmann, Boosting for high-dimensional linear models, *Ann. Stat.*, **34** (2006), 559–583. https://doi.org/10.1214/009053606000000092
10. J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, *Biometrika*, **95** (2008), 759–771. https://doi.org/10.1093/biomet/asn034
11. P. Zhao, B. Yu, On model selection consistency of Lasso, *J. Mach. Learn. Res.*, **7** (2006), 2541–2563.
12. J. Chen, Z. Chen, Extended BIC for small-n-large-p sparse GLM, *Stat. Sinica*, **22** (2012), 555–574. https://doi.org/10.5705/ss.2010.216
13. S. Luo, Z. Chen, Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters, *Stat. Interface*, **6** (2013), 275–284. https://doi.org/10.4310/SII.2013.v6.n2.a10

A. Proof of Theorem 1

For notation simplicity through the proofs, let $\widehat{\boldsymbol{\theta}}_s$ also denote the augmented vector $(\widehat{\boldsymbol{\theta}}_s, \mathbf{0}_{\bar{s}})$, where $\mathbf{0}_{\bar{s}}$ corresponds to the sub-vector $\boldsymbol{\beta}_{\bar{s}}$. The score function of the log-likelihood function $\ell(\boldsymbol{\theta})$ in Eq (2.1)

with respect to $\boldsymbol{\beta}$ is given by

$$g(\boldsymbol{\theta}) = [g_j(\boldsymbol{\theta})] = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} X^\top (A(\rho)\boldsymbol{y} - X\boldsymbol{\beta}).$$

Let $C > 0$ be a generic constant and $s_1 \subset s_2 \subset \cdots$ the model sequence identified in the PVS procedure. Recall that s_0 denotes the index set of non-zero components in true coefficient vector $\boldsymbol{\beta}_0$. To prove Theorem 1, it suffices to show the following three parts under the assumptions.

(1) There exists a k^* such that

$$\lim_{n\to\infty} \mathrm{P}(s_{k^*} = s_0) = 1. \tag{A.1}$$

(2) Uniformly for $k < p_0$ and $0 < \gamma < 1$,

$$\lim_{n\to\infty} \mathrm{P}(\mathrm{EBIC}_\gamma(s_k) > \mathrm{EBIC}_\gamma(s_{k+1})) = 1. \tag{A.2}$$

(3) For some constant $K_0 > 1$ and $\gamma > 1 - \log n/(2\log p)$,

$$\lim_{n\to\infty} \mathrm{P}(\min_{p_0<k<K_0 p_0} \mathrm{EBIC}_\gamma(s_k) > \mathrm{EBIC}_\gamma(s_0)) = 1. \tag{A.3}$$

Proof of (A.1). Note the fact that $\psi(j \mid s) = g_j(\widehat{\boldsymbol{\theta}}_s) = 0$ for all $j \in s$. We will show that

$$\lim_{n\to\infty} \mathrm{P}\Big(\forall s \subsetneqq s_0,\ \max_{j\in s_0\setminus s} |\psi(j \mid s)| > C \max_{j\notin s_0} |\psi(j \mid s)|\Big) = 1.$$

Given $s \subsetneqq s_0$, consider the following decomposition:

$$\begin{aligned}\psi(j \mid s) &= [g_j(\widehat{\boldsymbol{\theta}}_s) - g_j(\bar{\boldsymbol{\theta}}_s)] + [g_j(\bar{\boldsymbol{\theta}}_s) - \mathrm{E}\, g_j(\bar{\boldsymbol{\theta}}_s)] + \mathrm{E}\, g_j(\bar{\boldsymbol{\theta}}_s)\\ &=: \psi_1(j \mid s) + \psi_2(j \mid s) + \psi_3(j \mid s).\end{aligned}$$

Under (C2) and noting the row summation of weight matrix W being one, we have

$$\begin{aligned}|\psi_1(j \mid s)| &= \frac{1}{\widehat{\sigma}^2}\Big|\Big\{[A(\widehat{\rho}) - w_s A(\bar{\rho})]\boldsymbol{y} - X_s(\widehat{\boldsymbol{\beta}}_s - w_s\bar{\boldsymbol{\beta}}_s)\Big\}^\top \boldsymbol{X}_j\Big|\\ &= \frac{1}{\widehat{\sigma}^2}\Big\|[A(\widehat{\rho}) - w_s A(\bar{\rho})]\boldsymbol{y}\Big\|_\infty \big\|\boldsymbol{X}_j\big\|_1 + \Big\|\sum_{i=1}^n x_{ij}\boldsymbol{X}_{(i)s}\Big\|_\infty \big\|\widehat{\boldsymbol{\beta}}_s - w_s\bar{\boldsymbol{\beta}}_s\big\|_1\\ &\le C(n\big|\bar{\rho} - \widehat{\rho}\big| + np_0^{1/2}\big\|\widehat{\boldsymbol{\beta}}_s - \bar{\boldsymbol{\beta}}_s\big\|)\\ &\le C(n^{1/2} + np_0^{1/2}\big\|\widehat{\boldsymbol{\beta}}_s - \bar{\boldsymbol{\beta}}_s\big\|),\end{aligned}$$

where $w_s = \widehat{\sigma}^2/\bar{\sigma}^2$, and the last inequality is due to $\|\widehat{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\| = O(n^{-1/2}p_0^{-1})$ and $w_s = 1 + o_p(1)$, which can be obtained by following [2]. The proof is omitted to avoid a lenghty statement.

Next we examine the term $\|\widehat{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\|$. For $s \subset s_0$, we have the decomposition

$$\begin{aligned}\ell(\boldsymbol{\theta}_s) - \ell(\bar{\boldsymbol{\theta}}_s) &= [\ell(\boldsymbol{\theta}_s) - \mathrm{E}\,\ell(\boldsymbol{\theta}_s)] - [\ell(\bar{\boldsymbol{\theta}}_s) - \mathrm{E}\,\ell(\bar{\boldsymbol{\theta}}_s)] + [\mathrm{E}\,\ell(\boldsymbol{\theta}_s) - \mathrm{E}\,\ell(\bar{\boldsymbol{\theta}}_s)]\\ &=: e_1 + e_2 + e_3.\end{aligned}$$

Under (C2), by eliminating the terms only involved in the pre-estimated/specified $\bar{\rho}$ but irrelevant to $\boldsymbol{\beta}$ and applying Hoffding's inequality under independent sub-Gaussian variables, we have

$$\begin{aligned}
&\mathrm{P}(\ell(\bar{\boldsymbol{\theta}}_s) - \mathrm{E}\,\ell(\bar{\boldsymbol{\theta}}_s) \geq \sqrt{np_0}\log p) \\
&\leq \mathrm{P}((X_s\bar{\boldsymbol{\beta}}_s)^\top(A(\bar{\rho})\boldsymbol{y} - X_{s_0}\boldsymbol{\beta}_{s_0}) \geq C\sqrt{np_0}\log p) \\
&\leq \mathrm{P}((X_s\bar{\boldsymbol{\beta}}_s)^\top[A(\bar{\rho})A^{-1}(\rho_0)(X_{s_0}\boldsymbol{\beta}_{s_0} + \boldsymbol{\varepsilon}) - X_{s_0}\boldsymbol{\beta}_{s_0}] \geq C\sqrt{np_0}\log p) \\
&\leq \mathrm{P}((X_s\bar{\boldsymbol{\beta}}_s)^\top[\boldsymbol{\varepsilon} - (\bar{\rho} - \rho_0)W_0\boldsymbol{\eta}_0] \geq C\sqrt{np_0}\log p) \\
&\leq \mathrm{P}\Big(\sum_{i=1}^{n}\theta_i\varepsilon_i \geq C\sqrt{np_0}\log p\Big) \\
&\leq \exp\Big\{-C\frac{np_0(\log p)^2}{\sum_{i=1}^{n}\theta_i^2}\Big\} \\
&\leq \exp\{-Cp(\log p)^2\},
\end{aligned}$$

where $\theta_i = [X_s\bar{\boldsymbol{\beta}}_s]_i(1 + O(p_0n^{-1/2}))$ and $\boldsymbol{\eta}_0 = X\boldsymbol{\beta}_0$, and the last inequality is due to (C1). Thus we have

$$\ell(\bar{\boldsymbol{\theta}}_s) - \mathrm{E}\,\ell(\bar{\boldsymbol{\theta}}_s) = O_p(\sqrt{np_0}\log p).$$

Similarly, we also have $\ell(\boldsymbol{\theta}_s) - \mathrm{E}\,\ell(\boldsymbol{\theta}_s) = O_p(\sqrt{np_0}\log p)$ for any $\boldsymbol{\theta}_s$ such that $\|\boldsymbol{\theta}_s - \bar{\boldsymbol{\theta}}_s\| \to 0$.

Under (C2), noting that $\bar{\boldsymbol{\beta}}_s$ is the maximizer of $\mathrm{E}\,\ell(\boldsymbol{\theta}_s)$, we have

$$\begin{aligned}
\mathrm{E}\,\ell(\boldsymbol{\theta}_s) - \mathrm{E}\,\ell(\bar{\boldsymbol{\theta}}_s) &= -\frac{n}{2}\log\frac{\sigma^2}{\bar{\sigma}^2} - \log\frac{|A(\rho)|}{|A(\bar{\rho})|} - \frac{1}{2\bar{\sigma}^2}(\boldsymbol{\beta}_s - \bar{\boldsymbol{\beta}}_s)^\top H_s(\boldsymbol{\beta}_s - \bar{\boldsymbol{\beta}}_s) \\
&\leq -Cn^{1/2} - (1-\varepsilon)\lambda_{\min}(H_s)\|\boldsymbol{\beta}_s - \bar{\boldsymbol{\beta}}_s\|^2 \\
&\leq -C(n^{1/2} + n\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_s\|^2),
\end{aligned} \tag{A.4}$$

where $H_s = (X_s^\top X_s)\frac{\bar{\sigma}^2}{\sigma^2}$. Therefore, for any $\boldsymbol{\theta}_s$ such that $\|\boldsymbol{\theta}_s - \bar{\boldsymbol{\theta}}_s\|^2 = O(n^{-1/4})$, we have

$$\begin{aligned}
\ell(\boldsymbol{\beta}_s) - \ell(\bar{\boldsymbol{\beta}}_s) &\leq -Cn\|\boldsymbol{\beta}_s - \bar{\boldsymbol{\beta}}_s\|^2 + O_p(\sqrt{np_0}\log p) \\
&\leq -Cn^{3/4}.
\end{aligned}$$

Due to the concavity of $\ell(\boldsymbol{\theta}_s)$ with respect to $\boldsymbol{\beta}_s$, we have

$$\mathrm{P}(\|\widehat{\boldsymbol{\beta}}_s - \bar{\boldsymbol{\beta}}_s\|^2 \geq n^{-3/4}) \leq 2\exp\{-Cp_0(\log p)^2\}. \tag{A.5}$$

Under (C1), we have

$$\begin{aligned}
\mathrm{P}(\max_{s\subsetneq s_0}\max_{j\in\bar{s}}\big|\psi_1(j \mid s)\big| \geq Cn^{7/8}p_0^{1/2}) &\leq p^{p_0+1}\mathrm{P}(\|\widehat{\boldsymbol{\theta}}_s - \bar{\boldsymbol{\theta}}_s\| \geq Cn^{-1/4}) \\
&\leq 2\exp\Big\{-Cp_0(\log p)^2 + (p_0+1)\log p\Big\} \\
&\to 0.
\end{aligned}$$

Now we examine $\psi_2(j \mid s)$. Using Hoffding's inequality and $W_0\mathbf{1}_n = \mathbf{1}_n$, we have

$$\begin{aligned}
\mathrm{P}(|\psi_2(j \mid s)| \geq \sqrt{np_0}\log p) &= \mathrm{P}(|\frac{1}{\bar{\sigma}^2}X_s^\top[I - (\bar{\rho} - \rho_0)W_0A^{-1}(\rho_0)]\boldsymbol{\varepsilon}| \geq \sqrt{np_0}\log p) \\
&\leq \mathrm{P}\Big(|\sum_{i=1}^{n}\theta_i\varepsilon_i| \geq \sqrt{np_0}\log p\Big) \\
&\leq 2(1+\varepsilon)\exp\Big\{-\frac{np_0(\log p)^2}{2\sum_{i=1}^{n}\theta^2}\Big\} \\
&\leq 2(1+\varepsilon)\exp\{-Cp_0(\log p)^2\},
\end{aligned}$$

where $\theta_i = \|X_{s(i)}\|(1 + O_p(n^{-1/2}))$ and $\varepsilon_i \sim N(0, \sigma_0^2)$. Therefore, uniformly for $s \subsetneqq s_0$, we have

$$\mathrm{P}(\max_{s\subsetneqq s_0}\max_{j\in\bar{s}}|\psi_2(j \mid s)| \geq \sqrt{np_0}\log p) \leq 2p^{p_0+1}\exp\{-Cp_0(\log p)^2\} \to 0.$$

Combined with Assumption (C3), we have

$$\max_{j\in\bar{s}}\Big\{|\psi_1(j \mid s)|, |\psi_2(j \mid s)|\Big\} = o_p(|\psi_3(j \mid s)|), \text{ uniformly for } s \subsetneqq s_0.$$

Consequently, with probability one, we have that

$$\min_{s\subsetneqq s_0}\max_{j\in s_0 - s}|\psi(j \mid s)| > C\max_{s\subsetneqq s_0}\max_{j\in\bar{s}}|\psi(j \mid s)|,$$

which yields the desired result.

□

Proof of (A.2). The event $\mathrm{EBIC}_\gamma(s_k) > \mathrm{EBIC}_\gamma(s_{k+1})$ is equivalent to

$$\ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_{k+1}}) < -(0.5\log n + \gamma\log p).$$

Denote by $\breve{\boldsymbol{\theta}}_{s_k}$ the augmented vector of $\widehat{\boldsymbol{\theta}}_{s_k}$ with zero corresponding to the element of $s_{k+1} - s_k$. Expanding $\ell(\breve{\boldsymbol{\theta}}_{s_k})$ at $\widehat{\boldsymbol{\theta}}_{s_{k+1}}$ with respect to $\boldsymbol{\beta}$ and noting that $\widehat{\boldsymbol{\theta}}_{s_{k+1}}$ is the maximizer of $\ell(\boldsymbol{\theta}_{s_{k+1}})$, we have

$$\begin{aligned}
\ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_{k+1}}) &= \ell(\breve{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_{k+1}}) \\
&= -\frac{n}{2}\log\frac{\widehat{\sigma}^2_{s_k}}{\widehat{\sigma}^2_{s_{k+1}}} - \log\frac{|A(\widehat{\rho}_{s_k})|}{|A(\widehat{\rho}_{s_{k+1}})|} - \frac{1}{2\widehat{\sigma}^2_{s_{k+1}}}(\breve{\boldsymbol{\beta}}_{s_k} - \widehat{\boldsymbol{\beta}}_{s_{k+1}})^\top H_{s_{k+1}}(\breve{\boldsymbol{\beta}}_{s_k} - \widehat{\boldsymbol{\beta}}_{s_{k+1}}) \\
&= -Cn\log n - \frac{1}{2\widehat{\sigma}^2_{s_{k+1}}}(\breve{\boldsymbol{\beta}}_{s_{k+1}} - \widehat{\boldsymbol{\beta}}_{s_{k+1}})^\top H_{s_{k+1}}(\breve{\boldsymbol{\beta}}_{s_{k+1}} - \widehat{\boldsymbol{\beta}}_{s_{k+1}}),
\end{aligned}$$

where $H_{s_{k+1}} = X^\top_{s_{k+1}}X_{s_{k+1}}\frac{\widehat{\sigma}^2_{s_k}}{\widehat{\sigma}^2_{s_{k+1}}}[1 + o_p(1)]$. Together with (C2), we have $\boldsymbol{u}^\top H_s\boldsymbol{u} = O(n)$ for all $s \subset s_0$ and $\|\boldsymbol{u}\| = 1$. From (A.5) and (C4), we also have $\|\widehat{\boldsymbol{\beta}}_{s_k} - \bar{\boldsymbol{\beta}}_{s_k}\|^2 = o_p(\|\breve{\boldsymbol{\beta}}_{s_k} - \bar{\boldsymbol{\beta}}_{s_{k+1}}\|^2)$. Therefore, with probability going to one, we have

$$\ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_{k+1}}) \leq -Cn\|\|^2 \leq -Cn\|\|^2 \leq -Cn^{-3/4}.$$

The above inequality holds uniformly for $s_k \subset s_0$ with $k < p_0$, which completes the proof.

□

Proof of (A.3). By definition, $\mathrm{EBIC}_\gamma(s_k) > \mathrm{EBIC}_\gamma(s_{k+1})$ is equivalent to

$$\ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_{k+1}}) < (k - p_0)(0.5 \log n + \gamma \log p).$$

It suffices to show that, uniformly for $p_0 < k < K_0 p_0$, the above inequality holds with probability one. Recall that $\boldsymbol{\theta}_0 = (\rho_0, \sigma_0^2, \boldsymbol{\beta}_0)$ is the true coefficient. Using the Taylor expansion with respect to $\boldsymbol{\beta}$, we have

$$\begin{aligned}
\ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\widehat{\boldsymbol{\theta}}_{s_0}) &\le \ell(\widehat{\boldsymbol{\theta}}_{s_k}) - \ell(\boldsymbol{\theta}) \\
&= -\frac{n}{2} \log \frac{\widehat{\sigma}^2_{s_k}}{\sigma^2} + \log \frac{|A(\widehat{\rho})|}{|A(\rho)|} + (\widehat{\boldsymbol{\beta}}_{s_k} - \boldsymbol{\beta})^\top g(\boldsymbol{\beta}) - \frac{1}{2}(\widehat{\boldsymbol{\beta}}_{s_k} - \boldsymbol{\beta})^\top H (\widehat{\boldsymbol{\beta}}_{s_k} - \boldsymbol{\beta}) \\
&\le -Cn \log n + \frac{1}{2} g(\boldsymbol{\beta})^\top H^{-1} g(\boldsymbol{\beta}),
\end{aligned}$$

where $H = X_s^\top X_s \frac{\widehat{\sigma}^2}{\sigma^2}[1 + o(1)]$. By integrating (C2) and (A.5), it holds that $C^{-1} n \le \lambda_{\min}(H) \le \lambda_{\max}(H) \le Cn$. Under (C1) with $p_0 = o(n^{1/6})$ and following the proof of part (2) of Theorem 2.1 in [13], the proof is complete.

□