Computational Statistics and Data Analysis 155 (2021) 107094

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust variable selection with exponential squared loss for the spatial autoregressive model[☆]

Yunquan Song [a,*], Xijun Liang [a,*], Yanji Zhu [a], Lu Lin [b]

[a] *College of Science, China University of Petroleum, Qingdao 266580, PR China*
[b] *Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250014, PR China*

ARTICLE INFO

Article history:
Received 28 August 2019
Received in revised form 3 May 2020
Accepted 7 September 2020
Available online 28 September 2020

Keywords:
Spatial autoregressive model
Exponential squared loss
Oracle property
Adaptive lasso
Variable selection

ABSTRACT

Spatial dependent data frequently occur in spatial econometrics and endemiology. In this work, we propose a class of penalized robust regression estimators based on exponential squared loss with independent and identical distributed errors for general spatial autoregressive models. A penalized exponential squared loss with the adaptive lasso penalty is employed for simultaneous model selection and parameter estimation. Under mild conditions, we establish the asymptotic and oracle property of the proposed estimators The induced nonconvex nondifferentiable mathematical programming offer challenges for solving algorithms. We specially design a block coordinate descent (BCD) algorithm equipped with CCCP procedure for efficiently solving the subproblem. Moreover, we provide a convergence guarantee of the BCD algorithm. Every limit point of the iterated solutions is proved a stationary point. We also present a convergence speed of spatial weight ρ^k. Numerical studies illustrate that the proposed method is particularly robust and applicable when the outliers or intensive noise exist in the observations or the estimated spatial weight matrix is inaccurate. All the source code could be freely downloaded from https://github.com/Isaac-QiXing/SAR.

1. Introduction

In many fields including spatial econometrics and endemiology, the data used is spatial dependent data. To deal with these types of data, spatial regression models are widely studied. Among them, the spatial autoregressive (SAR) model

$$Y = \rho WY + X\beta + \varepsilon$$

is widely studied with W a spatial weights matrix. The SAR model was proposed by Cliff and Ord (1973) and has attracted widespread attention (Anselin and Bera, 1998; Cressie, 1992). As a spatial weights matrix is usually constructed from geographical or economic information to characterize the spatial dependence, the candidates for the spatial weights matrix might not be unique. There exist two types of methods to select the spatial weights matrix W. The first type of methods select the spatial weights matrix from a set of alternative ones. Kelejian (2008) selects the true spatial weights matrix using GMM estimates. It is suggested to use a non-nested J-test for testing a null SAR model against a set of alternative models with different spatial weights matrices. Kelejian and Piras (2011) suggest a modification of Kelejian's J-test that

[☆] The researches are supported by NNSF project (11971265) of China, Natural Science Foundation(NSF) project (ZR2019MA016) of Shandong Province of China, the Fundamental Research Funds for the Central Universities (No. 20CX05012A).

[*] Corresponding authors.
E-mail addresses: syqfly1980@163.com (Y. Song), liangxijunsd@163.com (X. Liang).

https://doi.org/10.1016/j.csda.2020.107094
0167-9473/© 2020 Elsevier B.V. All rights reserved.

uses available information in a more efficient way, and Kelejian and Piras (2014) extend it to a panel data setting. The second type of methods estimate a weighting matrix by averaging different spatial weights matrices. A model averaging procedure was proposed in Zhang and Yu (2018) to reduce estimation error. This type of methods overcome the difficulty that the true spatial weights matrix is not among the candidates.

In the field of classical linear regression models, there are already many works which have made efforts to variable selection. One popular way is the penalized regression method, in which a number of choices of penalty functions can be chosen for variable selection, such as least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996), smoothly clipped absolute deviation (SCAD, Fan and Li, 2001), and adaptive LASSO (Zou, 2006). Due to the spatial dependence, the above penalized methods could be used directly in the variable selection of SAR model.

As the classical variable selection methods are heavily affected by intense noise and outliers, a number of robust approaches have been proposed. Many studies adopt Huber's loss function (Huber and Ronchetti, 1981). As Huber's method has limitations in terms of efficiency, Wang et al. (2013) proposed a class of robust estimators based on the exponential squared loss function $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$, which is widely used in boosting algorithm (Friedman et al., 2000). For instance, the parameter of the linear model $y_i = x_i^T\beta + \varepsilon_i$ can be estimated by minimizing $\sum_{i=1}^n \phi_\gamma(t_i)$, where $t_i = y_i - x_i^T\beta$ represents the residual of the ith observation, and $\gamma > 0$ controls the degree of robustness and efficiency. For a large γ, $1 - \exp(-t^2/\gamma) \approx t^2/\gamma$, which means that the proposed estimator is similar to the least squares estimator in this case. When γ is small, observations with large values of $|t_i|$ will result in empirical losses near 1.0 and therefore have a small impact on the estimation. Hence, a smaller γ would limit the influence of outliers on the estimators. A choice of γ is proposed in Wang et al. (2013). It was also pointed out that the method is more robust than the other compared robust methods, including Huber's estimator, quantile regression estimator (Koenker and Bassett, 1978), and composite quantile regression estimator (Zou and Yuan, 2008).

We focus on the variable selection for the most popular spatial autoregressive (SAR) model based on the regular lattice data. At present, there are mainly two types of variable selection methods for SAR model. One is the Bayesian method and the other is the non-Bayesian method. As far as the Bayesian method is concerned, starting with the seminal work of Bayesian model selection for the SAR model by LeSage and Parent (2007), there are a great bulk of literature contributed on variable selection and model selection for the SAR models. LeSage and Parent (2007) developed a Markov Chain Monte Carlo model composition methodology (MC3) and a Bayesian model averaging (BMA) technique for the SAR and spatial error models, and focused exclusively on model specification issues regarding the choice of explanatory variables as in conventional linear models. Since the BMA techniques for the SAR models rely on the calculation of marginal likelihoods, there is a severe computational burden when a large number of covariates are potential candidates of the specification. To reduce the time consuming, Piribauer (2016) proposed a posterior model based on the Bayesian information criterion and maximum likelihood estimates of the matrix exponential specification (see LeSage and Parent, 2007) of global spatial spillover effects. Piribauer (2016) used stochastic search variable selection (SSVS) priors to deal with the problem of variable selection in the SAR models, which avoids the complex calculation of marginal likelihoods in BMA. Recent work by Krisztin (2017) provided a Bayesian variable selection method for a semiparametric spatial autoregressive model. A good overview of related methods can be founded in Steel (2017). Although Bayesian methods have made great progress in variable and model selection for the SAR models, there are some difficulties to assess the quality of priors and choose proper priors for users in applications. On the other hand, the classical non-Bayesian penalized methods, such as LASSO and SCAD, have got success in the classical linear regression models. However, there is only some initial focus on non-Bayesian methods for SAR model. To the best of our knowledge, Guo and Wei (2015) studied variable selection of the SAR model via LASSO method. Xuan et al. (2018) developed a penalized quasi-maximum likelihood method for simultaneous model selection and parameter estimation, Ma et al. (2019) employed the naive least squares for the estimation of unknown parameters in SAR models.

These methods, however, are affected to outliers in finite samples. In fact, the outliers or intense noise bring in challenges in parameter estimation and variable selection for SAR model. In the regression setting, the robustness of the resulting estimators heavily depends on the choice of the loss function. For dealing with outliers in spatial autoregression, it is natural to consider employing robust loss functions. However, this research map is far from trivial. The main challenge comes from the spatial dependence, which is usually characterized by the spatial weight matrix. As the weight matrix itself usually could not be accurately estimated, or even has a large deviation from the true value, it is troublesome to deal with the inherent large noise at the same time.

In view of the robustness of the exponential squared loss function, we take a parametric penalized approach and assume that the errors in the SAR model are independent and have identical distribution. Our idea is to penalize all unknown parameters except error variance in penalized exponential squares loss function and achieve penalized exponential squared loss estimators (PELE) using adaptive Lasso penalty. The constructed optimization model is as follows

$$\min_{\beta\in R^p,\rho\in[0,1]} \quad L(\beta,\rho) = \frac{1}{n}\sum_{i=1}^{n} \phi_\gamma(Y_i - \rho\tilde{Y}_i - X_i\beta) + \lambda\sum_{j=1}^{p} P(|\beta_j|) \tag{1}$$

where $\lambda > 0$, $\tilde{Y} = WY$, $\sum_{j=1}^p P(|\beta_j|)$ is a penalty term, $\phi_\gamma(\cdot)$ is the exponential squared loss function: $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$. Here, γ is tuning parameter controlling the degree of robustness. A small γ could limit the impact of an outlier on the estimators, but it would also reduce the sensitivity of the model. While it seems natural to construct the variable selection model of (1), it turns out to be a challenge to analyze its statistical property and to design an efficient optimization algorithm. As ρ is a variable to be solved, and the exponential squared loss function is nonconvex, the empirical loss term is an essentially a structured nonconvex function with respect to two block of variables, $\beta \in R^p$ and $\rho \in [0, 1]$. Moreover, as many of the penalty terms, such as the Lasso or adaptive Lasso penalty, are nondifferentiable, the objective function of (1) is a nonconvex, non-differentiable and block-structured function.

In this work, we presented a robust variable selection method for spatial autoregressive based on the exponential squared loss function and the adaptive lasso penalty. The method could select important predictors and, simultaneously, estimate the regression coefficients. The main contributions of this work are as follows.

1. We construct a robust variable selection method for SAR model, equipped with the exponential squared loss function for resisting the affections of the observations with intense noise or an inaccurate spatial weight matrix.
2. We proposed a block-coordinate descent (BCD) algorithm for solving the SAR model. A DC (difference of two convex functions) decomposition of the exponential squared loss is specially designed, based on which a CCCP procedure for solving the subproblem of the BCD algorithm is constructed. We also presented a convergence analysis of the BCD algorithm. Convergence to stationary points and the convergence speed was analyzed under mild conditions.
3. It is proved that the proposed variable selection method has Oracle properties under reasonable assumptions. Moreover, we conducted detailed numerical studies, and validated robustness and effectiveness the proposed method in selecting important variables. The numerical studies show that the proposed method overwhelms the compared methods when the observations have outliers in terms of the number of correctly identified zero coefficients, the number of incorrectly identified nonzero coefficients and MedSE.

The organization of this paper is as follows. In Section 2, we introduce the SAR model with independent and identical distribution errors, and present the penalized exponential squared loss method under adaptive Lasso penalty function. In Section 3, we propose an efficient algorithm to complete the variable selection procedure. Some simulations are carried out to examine the finite sample performance in Section 4 and an example of application is presented in Section 5. we conclude the research in Section 6.

Users could freely download the Matlab source code from https://github.com/Isaac-QiXing/SAR.

The main abbreviations and notations used in this work are as follows.

SAR model:	spatial autoregressive model;
BCD algorithm:	block-coordinate descent algorithm;
DC function:	difference of two convex functions;
$\mathcal{N}(\mu, \sigma^2)$:	Gaussian distribution with the mean value of μ and the variance of σ^2;
$I_n \in \mathbb{R}^{n\times n}$:	an identity matrix in $\mathbb{R}^{n\times n}$.

2. Estimation and variable selection

2.1. Spatial autoregression model

We suppose a continuous response $Y_i \in \mathbb{R}^{1\times 1}$ and an associated p-dimensional predictors $X_i = (X_{i1}, \dots, X_{ip})$, where the p is a fixed constant. Denote the response vector $Y = (Y_1, \dots, Y_n)^T$ and the design matrix $X = (X_1, \dots, X_n)^T \in \mathbb{R}^{n\times p}$. We consider the following SAR model with covariates

$$Y = \rho WY + X\beta + \varepsilon \tag{2}$$

where $\rho \in \mathbb{R}^{1\times 1}$ is known as the network autocorrelation coefficient, $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^{p\times 1}$ is the regression coefficient vector, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T \in \mathbb{R}^{n\times 1}$ is the error vector. The SAR model is recast as $Y_i = \rho \sum_{j=1}^n w_{ij} Y_j + X_i\beta + \varepsilon_i$, which intuitively implies that the response of the ith subject is linearly depends on its neighbors and covariates. The measure of network dependence strength can be interpreted with ρ. It is usually supposed that the noise of ε_is are independent and identically distributed as $\mathcal{N}(0, \sigma^2)$. It means that the covariance $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$, with $I_n \in \mathbb{R}^{n\times n}$ is an identity matrix. Then the observations of Y could be formulated as

$$Y = (I_n - \rho W)^{-1}(X\beta + \varepsilon), \tag{3}$$

where $I_n - \rho W$ is ensured an invertible matrix. According to Banerjee et al.'s work (Banerjee et al., 2014), the matrix W has its largest singular value of 1 under certain normalization operations. Therefore, a condition $|\rho| < 1$ ensures the invertibility of $I_n - \rho W$. Based on this, we add one constraint of ρ. Moreover, we ignore the potential endogeneity issue induced by WY and ε as Ma et al. (2019).

2.2. Variable selection method for SAR model

Consider the variable selection for the SAR model (2). To ensure the model identifiability and to enhance the model fitting accuracy and interpretability, the true regression coefficient vector β^* is commonly assumed to be sparse with only a small proportion of nonzeros (Fan and Li, 2001; Tibshirani, 1996). It is natural to employ the penalized approach to detect the true model, which selects important variables and estimates the values of parameters simultaneously. The constructed model is recast as E.q. (1), where the exponential squared loss is equipped. Another issue is the choice of the penalty term. Lasso or adaptive Lasso penalty could be considered if there is no extra structured information. Suppose that $\hat{\beta}$ is a root-n-consistent estimator for β, for instance, the naive least square estimator $\hat{\beta}(ols)$. Pick $r > 0$, and define the weight vector $\eta \in \mathbb{R}^p$ with $\eta_j = 1/|\hat{\beta}_j|^r, j = 1, \ldots, p$. We set $r = 1$ as suggested by Zou (2006). An adaptive Lasso penalty is recast as

$$\sum_{j=1}^{p} P(|\beta_j|) = \sum_{j=1}^{p} \eta_j |\beta_j|.$$

The penalized robust regression with exponential squared loss and an adaptive Lasso penalty is formulated as

$$\min_{\beta \in R^p, \rho \in [0,1]} \quad L(\beta, \rho) = \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(Y_i - \rho \tilde{Y}_i - X_i \beta) + \lambda \sum_{j=1}^{p} \eta_j |\beta_j| \tag{4}$$

where $\lambda > 0$ is a regularization parameter, $\tilde{Y} = WY$, ϕ_γ is the exponential squared loss: $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$. The selection of tuning parameter γ is discussed in Section 4.3.

2.3. Estimation of the variance of the noise

Let $G = (I_n - \rho W)^{-1}$, then the variance of the noise is estimated as

$$\hat{\sigma}^2 = \frac{1}{n}(Y - GX\beta)^T (GG^T)^{-1} (y - GX\beta), \tag{5}$$

where ρ and β could be estimated by the solutions of (4). It is noted that G is a nonsingular matrix, then $(GG^T)^{-1} = (G^T)^{-1} G^{-1} = (I_n - \rho W)^T (I_n - \rho W)$. Let $u = GX\beta$. then $u = G \cdot X\beta = (I_n - \rho W)^{-1}(X\beta)$, and can be calculated by solving a linear system. Then $\hat{\sigma}^2$ defined by (5) can be calculated by

$$\hat{\sigma}^2 = \frac{1}{n} \|(I_n - \rho W) \cdot (Y - u)\|_2^2. \tag{6}$$

3. Large sample properties and oracle properties

In this section, we establish the asymptotic theory for the nonconcave penalized estimator for spatial autoregressive model.

Let $\hat{\boldsymbol{\beta}}_n = \left(\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T\right)^T$ be the resulting estimator of (4), $I(\boldsymbol{\beta}, \gamma) = \frac{2}{\gamma} \int ZZ^T e^{-r^2/\gamma} \left(\frac{2r^2}{\gamma} - 1\right) dF(Z, y)$, where $r = Y - (I_n - \rho W)^{-1} X\boldsymbol{\beta} = Y - Z\boldsymbol{\beta}$, $Z = (I_n - \rho W)^{-1}X$, $a_n = \max\left\{p'_{\lambda_{nj}}\left(|\beta_{0j}|\right) : \beta_{0j} \neq 0\right\}$, $b_n = \max\left\{p''_{\lambda_{nj}}\left(|\beta_{0j}|\right) : \beta_{0j} \neq 0\right\}$. For ease of presentation, omit the subscript of $\boldsymbol{\beta}_n$. Let the true value of $\boldsymbol{\beta}$ be $\boldsymbol{\beta}_0 = \left(\beta_{10}, \ldots, \beta_{p0}\right)^T = \left(\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T\right)^T$ and the true value of ρ be ρ_0. Thus $\theta_0 = \left(\rho_0, \beta_0^T\right)^T$. Without loss of generality, assume that $\beta_{20} = 0$. For ease of presentation, let $\beta_{10} = \rho$ and $\beta_{1j} = \beta_{1j}, j = 1, 2, \ldots, s$, then denote $\boldsymbol{\beta}_1 = (\rho, \beta_{11}, \ldots, \beta_{1s})^T$ and $\boldsymbol{\beta}_{01} = (\rho_0, \beta_{01}, \ldots, \beta_{0s})^T$.

We demonstrate the asymptotic and oracle properties of the proposed penalized estimator. The following regularity assumptions are needed for the theorems.

Assumption 1. $\Sigma = E\left(ZZ^T\right)$ is positive definite and $E\|Z\|^3 < \infty$.

Assumption 2. The matrix $\boldsymbol{I}_n - \rho \boldsymbol{W}$ is nonsingular with $|\rho| < 1$

Assumption 3. The row and column sums of the matrices W_n and $I - \rho W_n$ are bounded uniformly in absolute value.

Assumption 4. For matrix $\boldsymbol{G}_n = \boldsymbol{W}\left(\boldsymbol{I} - \rho \boldsymbol{W}\right)^{-1}$, there exists a constant $\tilde{\lambda}_c$ such that $\tilde{\lambda}_c \boldsymbol{I}_n - \boldsymbol{G}_n \boldsymbol{G}_n^T$ is positive semidefinite for all n

Assumption 5. $1/\min_{s+1 \leq j \leq p} \lambda_j = o_p(1)$. And with probability 1, $\liminf_{n \to \infty} \liminf_{t \to 0^+} \left\{\min_{s+1 \leq j \leq p} \frac{p'_{\lambda_j}(t)}{\lambda_j}\right\} > 0$.

Assumption 6. $\sqrt{n} a_n = o_p(1)$, $b_n = o_P(1)$.

Assumption 7. $(\gamma_n - \gamma_0) = o_p(1)$ for some $\gamma_0 > 0$.

Assumption 8. There are constants C_1 and C_2 such that, when $\theta_1, \theta_2 > C_1 \lambda_j \left|p''_{\lambda_j}(\theta_1) - p''_{\lambda_j}(\theta_2)\right| \leq C_2 |\theta_1 - \theta_2|$, for $j = 0, 1, \ldots, p$

Assumption 1 ensures that the main term dominates the remainder in the Taylor expansion. It warrants further examination as to whether this condition can be weakened. Assumptions $2-4$ are required in the setting of spatial autoregressive model. Assumption 5 makes the penalty function singular at the origin so that the penalized estimators possess the sparsity property. Assumptions 6 and 7 ensure the unbiasedness property for large parameters and the existence of the $\sqrt{n}$ consistent penalized exponential square estimator and guarantee that the penalty function does not have much more influence than the least square function on the penalized estimators. In real data analysis,if the appropriate penalty function is chosen, then by choosing λ_j appropriately, we have $a_n = o_p(1/\sqrt{n})$ and $b_n = o_p(1/\sqrt{n})$ Therefore, the tuning parameters λ_j s have to be taken depending on data, e.g., by cross validation, AIC, BIC and so on. By the way, we choose the optimal tuning parameters by BIC in our simulation study and real data analysis. Assumption 8 is the smoothness condition that is imposed on the non-concave penalty functions. With these preparations, we present the following sampling properties for our proposed estimators. The following theorem gives the consistency of the proposed estimators.

The following theorems give the large sample properties of the proposed estimators.

Theorem 1. *Suppose that Assumptions* $1-8$ *hold, there exists a local maximizer* $\hat{\boldsymbol{\theta}}$ *such that* $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2} + a_n)$.

Theorem 2 (*Oracle Property*)**.** *Suppose that Assumptions* $1-8$ *hold, and* $I\left(\boldsymbol{\beta}_0, \gamma_0\right)$ *is negative definite. If* $\gamma_n - \gamma_0 = o_p(1)$ *for some* $\gamma_0 > 0$, $\hat{\theta} = (\hat{\rho}, \hat{\beta}_1^T, \hat{\beta}_2^T)^T$ *must satisfy:*

(i) *sparsity, that is,* $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$ *with probability 1;*

(ii) *asymptotic normality:*

$$\sqrt{n}\left(I_1\left(\boldsymbol{\beta}_{01}, \gamma_0\right) + \Sigma_1\right)\left\{(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) + \left(I_1\left(\boldsymbol{\beta}_{01}, \gamma_0\right) + \Sigma_1\right)^{-1}\Delta\right\} \to N\left(\mathbf{0}, \Sigma_2\right) \quad ,$$

where $\hat{\beta}_{n1} = (\hat{\rho}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{1s})^T$, *and* $\beta_{01} = (\rho_0, \beta_{01}, \ldots, \beta_{0s})^T$,

$$\Sigma_1 = \operatorname{diag}\left\{p''_{\lambda_1}\left(|\beta_{01}|\right), \ldots, p''_{\lambda_s}\left(|\beta_{0s}|\right)\right\}$$

$$\Sigma_2 = \operatorname{cov}\left(\exp\left(-r^2/\gamma_0\right)\frac{2r}{\gamma_0}Z_{i1}\right),$$

$$\Delta = \left(p'_{\lambda_1}\left(|\beta_{01}|\right)\operatorname{sign}\left(\beta_{01}\right), \ldots, p'_{\lambda_s}\left(|\beta_{0s}|\right) \times \operatorname{sign}\left(\beta_{0s}\right)\right)^T,$$

$$I_1\left(\boldsymbol{\beta}_{01}, \gamma_0\right)$$
$$= \frac{2}{\gamma_0}E\left[\exp\left(-r^2/\gamma_0\right)\left(\frac{2r^2}{\gamma_0} - 1\right)\right] \times \left(EZ_{i1}Z_{i1}^T\right).$$

4. Block coordinate descent algorithm

In this part, we embark on designing an efficient algorithm for solving (4). The optimization model has two blocks of variables. One block of variable is $\rho \in [0, 1]$, the other block of variable is $\beta \in \mathbb{R}^p$. A block coordinate descent algorithm could be imagined to solve the two blocks of variables alternately. Unfortunately, as the objective function of (1) is a nonconvex, non-differentiable, making the convergence of the coordinate descent algorithm questionable. Moreover, the subproblem for solving β is nonconvex and non-differentiable, an efficient algorithm for this sub-problem is far from trivial. We present the block coordinate descent algorithm at first, and then conquer the mentioned challenges literally.

4.1. The algorithm framework

We present the block coordinate descent (BCD) algorithm framework in Algorithm 1.

Algorithm 1 The block coordinate descent (BCD) algorithm for solving (4)

```
1. Set initial value for $\beta^0 \in \mathbb{R}^p$ and $\rho^0 \in (0, 1)$;
2. **repeat** {For $k = 0, 1, 2, \cdots$}
3.     Solve the subproblem about $\rho$ with initial point $\rho^k$:
           $$\rho^{k+1} \leftarrow \min_{\rho \in [0,1]} L(\beta^k, \rho); \qquad (7)$$
4.     Solve the subproblem with initial value $\beta^k$,
           $$\min_{\beta \in R^p} L(\beta, \rho^{k+1}) \qquad (8)$$
       to get a solution $\beta^{k+1}$, ensuring that $L(\beta^k, \rho^{k+1}) - L(\beta^{k+1}, \rho^{k+1}) \leq 0$, and $\beta^{k+1}$ is a stationary point of $L(\beta, \rho^{k+1})$.
5. **until** convergence.
```

4.2. Solving the subproblem (7) and (8)

In this subsection, we consider efficient procedures for solving the subproblem (7) and (8), especially for the exponential squared loss function and the Lasso or adaptive Lasso penalty.

A basic observation for the subproblem (8) is that for fixed ρ^k, Lasso or adaptive Lasso is convex, $Y - \rho WY - X\beta$ is affine with respect to β, and the exponential squared loss function ϕ_γ is a DC function (difference of two convex functions). Thus, the subproblem (8) is a DC programming, and could be solved by the corresponding algorithms.

We first pursue a DC-decomposition of the exponential squared loss function $\phi_\gamma(t) = 1 - e^{-\frac{t^2}{\gamma}}$. The following decomposition can be verified by calculating the second derivatives.

Proposition 1. *The exponential squared loss function $\phi_\gamma(t)$ can be expressed as the difference of two convex functions:*

$$\phi_\gamma(t) = [\phi_\gamma(t) + v(t)] - v(t),$$

where $v(t) = e^{\frac{t^2}{\gamma}}$, $\phi_\gamma(t) + v(t)$ are convex.

Unfortunately, $e^{\frac{t^2}{\gamma}}$ may introduce computational difficulty. For instance, when $\gamma = 1, t = 10$, $e^{\frac{t^2}{\gamma}} = e^{100} \approx 2.688 \times 10^{43}$ is large and may cause computational defect. We have found another DC decomposition of the exponential square loss $\phi_\gamma(t)$ as follows.

Proposition 2. *The exponential squared loss function $\phi_\gamma(t)$ can be expressed as the difference of two convex functions:*

$$\phi_\gamma(t) := [\phi_\gamma(t) + v(t)] - v(t) := u(t) - v(t), \tag{9}$$

with $\phi_\gamma(t) = 1 - e^{-\frac{t^2}{\gamma}}$, $v(t) = \frac{1}{3\gamma^2} t^4$, $u(t) = \phi_\gamma(t) + v(t)$.

Denote

$$\begin{aligned} J_{\text{vex}}(\beta) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \; u(Y_i - \rho^k \langle w_i, Y\rangle - X_i\beta) + \lambda \sum_{j=1}^p P(|\beta_j|), \\ J_{\text{cav}}(\beta) &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \; v(Y_i - \rho^k \langle w_i, Y\rangle - X_i\beta), \end{aligned} \tag{10}$$

with $u(\cdot)$, $v(\cdot)$ defined in (9), w_i the ith row of the weight matrix W, $\sum_{j=1}^p P(|\beta_j|)$ a convex penalty with respect to β. Then, $J_{\text{vex}}(\cdot)$ and $J_{\text{cav}}(\cdot)$ are convex and concave functions respectively. The subproblem (8) is formulated as

$$\min_{\beta \in R^p} \quad L(\beta, \rho^k) = J_{\text{vex}}(\beta) + J_{\text{cav}}(\beta),$$

and can be solved by Concave–Convex Procedure algorithm framework (Yuille and Rangarajan, 2001) as shown in Algorithm 2.

Algorithm 2 The Concave-Convex Procedure (CCCP)

```
Initialize $\beta^0$. Set $k = 0$.
repeat
    $\beta^{k+1} = \text{argmin}_\beta \quad J_{\text{vex}}(\beta) + J'_{\text{cav}}(\beta^k) \cdot \beta$        (11)
until convergence of $\beta^k$.
```

The CCCP algorithm minimizes $L(\beta, \rho^k)$ by iteratively solving a series of convex subproblems (11). The efficiency of the CCCP algorithm heavily depends on the algorithm for solving these subproblems. We focus on Lasso and adaptive Lasso penalty.

As $J'_{\text{cav}}(\beta^k) \cdot \beta$ is linear about β, we have by the definition of $J_{\text{vex}}(\beta)$ in (10) that the objective function $J_{\text{vex}}(\beta) + J'_{\text{cav}}(\beta^k) \cdot \beta$ can be expressed as

$$\min_{\beta \in R^p} \quad \psi(\beta) + \lambda \sum_{i=1}^p P(|\beta_i|), \tag{12}$$

with $\psi(\beta)$ is convex and continuously differentiable function, $\sum_{i=1}^p P(|\beta_i|)$ is the Lasso penalty, $\sum_{i=1}^p |\beta_i|$, or the more general adaptive Lasso penalty, $\sum_{i=1}^p \eta_i |\beta_i|$, $\eta_i \geq 0, i = 1, \ldots, p$. Beck and Teboulle (2009) presented an efficient algorithm, ISTA and FISTA, for solving the model with structure (12) for the Lasso penalty. In the following, we show that they can naturally be generalized for solving the model with adaptive Lasso penalty.

For all $L > 0$, ISTA approximate the function $F(\beta) = \psi(\beta) + \lambda \sum_{i=1}^{p} \eta_i |\beta_i|$ at $\beta = \xi$ as:

$$Q_L(\beta, \xi) = \psi(\xi) + \langle \beta - \xi, \nabla\psi(\xi)\rangle + \frac{L}{2}\|\beta - \xi\|^2 + \lambda \sum_{i=1}^{p} \eta_i |\beta_i|.$$

This function has the following minimum point

$$\begin{aligned} \Theta_L(\xi) &= \mathrm{argmin}_{\beta\in R^p}\ Q_L(\beta, \xi) \\ &= \mathrm{argmin}_{\beta\in R^p}\ \{\lambda \sum_{i=1}^{p} \eta_i|\beta_i| + \frac{L}{2}\|\beta - (\xi - \frac{1}{L}\nabla\psi(\xi))\|^2\} \\ &= \mathcal{S}_{\lambda\eta/L}(\xi - \frac{1}{L}\nabla\psi(\xi)), \end{aligned} \tag{13}$$

with $\eta = [\eta_1, \ldots, \eta_p] \in R^p$, and for $\nu = \lambda\eta/L \in R_+^p$, $\mathcal{S}_\alpha : \mathbb{R}^p \to \mathbb{R}^p$ the vector-formed soft-thresholding operator

$$\mathcal{S}_\nu(\beta) = \bar{\beta}, \quad \bar{\beta}_i = (|\beta_i| - \nu_i)_+ \mathrm{sgn}(\beta_i),\ \ i = 1, \ldots, p.$$

Then the iterative steps of ISTA for solving the model (12) are simply as follows

$$\beta^k = \Theta_L(\beta^{k-1}).$$

FISTA, as an accelerated version of ISTA, has been proved better convergence rate in both theory and practice (Beck and Teboulle, 2009). With the backtracking technique to estimate the unknown Lipschitz constant L, the iteration formulas of FISTA algorithm for solving (12) are given by Algorithm 3 which holds the same form as in Beck and Teboulle (2009).

Algorithm 3 FISTA with Backtracking Step for solving (11)

```
Require: $A$, $\xi$, $w\lambda > 0$
Ensure: solution $\beta$
Step 0. Select $L^0 > 0$, $\eta > 1$, $\beta^0 \in \mathbb{R}^p$ Let $\xi^1 = \beta^0$, $t^1 = 1$
Step $k$ ($k \geq 1$).
    Determine the smallest nonnegative integer $i^k$ which make $\bar{L} = \eta^{i^k} L^{k-1}$ satisfy
        $F(\Theta_{\bar{L}}(\xi^k)) \leq Q_{\bar{L}}(\Theta_{\bar{L}}(\xi^k), \xi^k).$
    Let $L^k = \eta^{i^k} L^{k-1}$ according to (13), calculate:
        $\beta^k = \Theta_{L^k}(\xi^k)$
        $t^{k+1} = \frac{1}{2}\left[1 + \sqrt{1 + 4(t^k)^2}\right]$;
        $\xi^{k+1} = \beta^k + \frac{t^k - 1}{t^{k+1}}(\beta^k - \beta^{k-1})$;
Output $\beta := \beta^k$.
```

The usual termination criterion of ISTA and FISTA is $\frac{\|\beta^k - \beta^{k-1}\|}{\max\{\|\beta^k\|, 1\}} \leq tol_\beta$, with $tol_\beta > 0$ a tolerance near zero. It is a relief to be finished to solve the main subproblem, i.e., (8). We have generalized the FISTA algorithm for efficiently solving the convex subproblem (11), ensuring numerical efficiency.

We now turn to consider solving the other subproblem (7) to update ρ^k. As the problem (7) minimizes a univariate function on the interval [0,1], the classical golden section search algorithm based on parabolic interpolation can be employed. See Forsythe et al. (1977) for details about the algorithm. Moreover, in Section 5 we would compare the proposed method with the model equipped with the square loss:

$$\min_{\rho\in(0,1)} L(\rho, \beta^k) := \frac{1}{n}\|y - X\beta^k - \rho W y\|_2^2 + \lambda \sum_{i=1}^{p} P(|\beta_i|). \tag{14}$$

It can be verified that

$$\rho^* = \mathrm{Proj}_{[0,1]} \frac{\langle y - X\beta^k, Wy\rangle}{\|Wy\|_2^2} \tag{15}$$

is the optimal solution of (14), with $\mathrm{Proj}_{[0,1]}(t) = \begin{cases} t, & t \in [0, 1], \\ 1, & t > 1, \\ 0, & t < 0. \end{cases}$ We will employ this closed-form solution for the square loss in Section 5.

Algorithm 1 terminates when either the criterion $\|[\beta^k, \rho^k]^T - [\beta^{k+1}, \rho^{k+1}]^T\| \leq \epsilon_1$, or $\|L(\beta^k, \rho^k) - L(\beta^{k+1}, \rho^{k+1})\| \leq \epsilon_2$, is satisfied.

The computational complexity of the proposed algorithm is analyzed as follows. As shown in Beck and Teboulle (2009), the iterated function values generated by FISTA for solving the subproblem (11) of CCCP converge to the optimal function value in speed of $O(1/k^2)$, with k the iteration steps. Thus, to obtain an ϵ-optimal solution, the required iterations of FISTA algorithm are $O(1/\sqrt{\epsilon})$ with each iteration to calculate the gradient $\nabla\phi(\beta)$ of (12). As the $O(np)$ computation is needed to calculate the gradient $\nabla\phi(\beta)$, the computation paid for an ϵ-optimal solution of the subproblem (11) is $O(np/\sqrt{\epsilon})$. Assume that BCD algorithms converge at specified (relatively small) number of iterations and in each iteration the CCCP algorithm terminates at most ite_{CCCP} iterations. Then the total computation complexity is $O(\text{ite}_{CCCP} \cdot np/\sqrt{\epsilon})$.

To implement our methodology, we need to select both tuning parameters λ and γ. Since λ and γ depend on each other, it could be treated as a bivariate optimization problem. In this article, we consider a simple selection method for λ and a data-driven procedure for γ.

4.3. The choice of the tuning parameter γ

The tuning parameter γ controls the degree of robustness and efficiency of the proposed robust regression estimators. A choice of γ for normal regression is proposed in Wang et al. (2013). We follow its procedure and generalize it for the spatial autoregressive model. At first a set of the tuning parameters are determined such that the proposed penalized robust estimators have asymptotic breakdown point at 1/2 and then the tuning parameter is selected with the maximum efficiency. The whole procedures are described in the following steps:

Step 1. Initialize $\hat{\rho} = \rho^{(0)}$ and $\hat{\beta} = \beta^{(0)}$. Set $\rho^{(0)} = \frac{1}{2}$, $\beta^{(0)}$ a robust estimator. Rewrite the model $Y = \rho WY + X\beta + \epsilon$ as $Y^* = X\beta + \epsilon$, where $Y^* = Y - \rho WY$.

Step 2. Find the pseudo outlier set of the sample:

Let $D_n = \{(X_1, Y_1^*), \ldots, (X_n, Y_n^*)\}$. Calculate $r_i(\hat{\beta}) = Y_i^* - X_i\hat{\beta}, i = 1, \ldots, n$ and $S_n = 1.4826 \times \text{median}_i|r_i(\hat{\beta}) - \text{median}_j(r_j(\hat{\beta}))|$. Then, there exist the pseudo outlier set $D_m = \{(X_i, Y_i) : |r_i(\hat{\beta})| \geq 2.5S_n\}$, set $m = \sharp\{1 \leq i \leq n : |r_i(\hat{\beta})| \geq 2.5S_n\}$, and $D_{n-m} = D_n \backslash D_m$.

Step 3. Select the tuning parameter γ_n: Construct $\hat{V}(\gamma) = \{\hat{I}(\hat{\beta})\}^{-1}\tilde{\Sigma}_2\{\hat{I}(\hat{\beta})\}^{-1}$, in which

$$\hat{I}(\hat{\beta}) = \frac{2}{\gamma}\{\frac{1}{n}\sum_{i=1}^{n}\exp(-r_i^2(\hat{\beta})/\gamma)(\frac{2r_i(\hat{\beta})}{\gamma} - 1)\} \cdot (\frac{1}{n}\sum_{i=1}^{n}X_iX_i^T)$$

$$\tilde{\Sigma}_2 = \text{Cov}\left\{\exp(-r_1^2(\hat{\beta})/\gamma)\frac{2r_1(\hat{\beta})}{\gamma}X_1, \ldots, \exp(-r_n^2(\hat{\beta})/\gamma)\frac{2r_n(\hat{\beta})}{\gamma}X_n\right\}.$$

Next, let γ_n be the minimizer of $\det(\hat{V}(\gamma))$ in the set $G = \{\gamma : \zeta(\gamma) \in (0, 1]\}$, where $\zeta(\cdot)$ enjoys the common definition with that in Wang et al. (2013).

Step 4. Update $\hat{\rho}$ and $\hat{\beta}$ as the optimal solution of $\min_{\beta\in R^p, \rho\in[0,1]} \frac{1}{n}\sum_{i=1}^{n}\phi_\gamma(Y_i - \rho\tilde{Y}_i - X_i\beta)$, where $\tilde{Y} = WY$. Go to Step 2 until convergence.

An initial robust estimator $\beta^{(0)}$ is needed in the initial step above. In practice, we use the estimator of the LAD loss as the initial estimator. In this sense, the selection of γ do not depend on λ in essential. However, one can also select the two parameters γ and λ jointly by cross-validation as discussed in Wang et al. (2013). But this approach requires huge computation. Moreover, the candidate interval of γ is $\{\gamma : \zeta(\gamma) \in (0, 1]\}$. In practice, we find the threshold of γ_1 such that $\zeta(\gamma_1) = 1$. The chosen value of γ is usually located in the interval of $[5\gamma_1, 30\gamma_1]$.

4.4. The choice of the regularization parameter λ and η_j

Consider the choice of the regularization parameter λ and η_j in (4). As the parameter λ can be unified with η_j, we set $\lambda_j = \lambda \cdot \eta_j$. In general, many methods can be used to select λ_j, such as Akaike information criterion (AIC), BIC and cross-validation. To guarantee consistent variable selection and reduce intensive computation, we consider the regularization parameter by minimizing a BIC-type objective function as Wang et al. (2007):

$$\sum_{i=1}^{n}\left[1 - \exp\left\{-\left(Y_i^* - X_i\beta\right)^2 \quad /\gamma_n\}\right] + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right| - \sum_{j=1}^{p}\log\left(0.5n\lambda_j\right)\log(n),$$

where $Y_i^* = Y_i - \rho W_n Y_i$. This leads to $\lambda_j = \frac{\log(n)}{n|\beta_j|}$. For a practical implementation, we do not know the values of β_j; however, they can be easily estimated by the unpenalized exponential squares loss estimator $\tilde{\beta}_j$, where the parameter value of γ has been estimated as described in Section 4.3. Note that this simple choice satisfies the conditions both $\sqrt{n}\hat{\lambda}_j \to 0$ for $j \leq p_0$, and $\sqrt{n}\hat{\lambda}_j \to \infty$ for $j > p_0$, with p_0 the number of nonzeros in the true value of β. Hence the consistent variable selection is guaranteed by the final estimator.

5. Numerical examples

In this section, we evaluate the performance of the proposed variable selection method on various synthetic cases, including the cases that outliers existed in the observations, and inaccurate spatial weight matrix W.

5.1. Simulation sampling

The data generating process is based on the model (2). We consider the covariates following a $(q+3)$-dimensional normal distribution with zero mean and covariance matrix (σ_{ij}) where $\sigma_{ij} = 0.5^{|i-j|}$. Thus, X_n is an $n \times (q+3)$ matrix. We set sample sizes $n \in \{30, 120, 360\}$ and the number of insignificant covariates $q \in \{5, 20, 80, 200\}$.

The spatial autoregressive coefficient ρ is generated from a uniform distribution on the interval $[\rho_1 - 0.1, \rho_1 + 0.1]$, where $\rho_1 \in \{0.8, 0.5, 0.2\}$. For comparison, we also consider the setting $\rho = 0$ which implies that there is no spatial dependent in the regression model and the model (2) degenerates to normal linear regression.

Let the spatial weight matrix $W_n = I_R \otimes B_m$, where $B_m = (1/(m-1))(\mathbf{l}_m \cdot \mathbf{l}_m^T - I_m)$, $\otimes$ is the Kronecker product and $\mathbf{l}_m$ is an m-dimensional column vector of ones. We take $m = 3$ and different values of R, where $R = 10,\ 40,\ 120$.

The regression coefficients are set to $\beta = (\beta_1, \beta_2, \beta_3, 0_q)^T$, where $(\beta_1, \beta_2, \beta_3)$ is generated from a 3-dimensional normal distribution with mean vector $(3, 2, 1.6)$ and covariance matrix $0.01 \cdot I_3$, with $I_3 \in \mathbb{R}^{3\times 3}$ the unit matrix, 0_q is a zero vector of q dimension. The dependent variable Y_n is given by (3),

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, σ^2 is generated from a uniform distribution on the interval $[\sigma_1 - 0.1, \sigma_1 + 0.1]$ with $\sigma_1 \in \{1, 2\}$. The above settings of the observed noise are completely Gaussian, which enjoys well theoretical property. In practice, however, this is not always the case. We consider the case that there exist outliers in the response. The error term follows a mixture normal distribution $(1-\delta_1) \cdot \mathcal{N}(0, 1) + \delta_1 \cdot \mathcal{N}(10, 6^2)$, where $\delta_1 \in \{0.01,\ 0.05\}$.

Another important issue for spatial auto-regression is the estimation of the weighting matrix W. As $W \in \mathbb{R}^{n\times n}$ is made up of the correlations of each two observations, it is usually hard to get an accurate estimation of the weighting matrix W in practice. To verify the impact of inaccurate estimations of the matrix W, we specially construct the following inaccurate weighting matrix.

1. Remove 30%, 50% and 80% nonzero weights randomly from each row of the true weighing matrix W, respectively.
2. Add 50% nonzero elements randomly in each row of the true weighing matrix W, i.e., change some zero elements to nonzeros in each row of W.
3. Add 100%, 200% nonzero elements randomly in 10% rows of the true weighing matrix W, respectively, the other rows of the weighting matrix remain unvaried.

The perturbed weighting matrix is denoted as $\hat{W}$. Then the matrix $\hat{W}$ is normalized to ensure that the elements of each row have the summation of 1. The perturbed weighting matrix $\hat{W}$ would replace W and put into the spatial auto-regression model.

5.2. Simulation results

For each case, the following reference is based on 100 simulations. The average number of zero coefficients which are correctly chosen is labeled as "Correct". The label "Incorrect" depicts the average number of non-zero coefficients incorrectly identified zero. As in Liang and Li (2009), we compare estimation accuracy using the median of squared error (MedSE) defined as $\|\beta - \hat{\beta}\|^2$, where $\|\beta\| = \sqrt{\sum_{i=1}^n \beta_i^2}$, $\beta = (\beta_1, \ldots, \beta_n)$, $\hat{\beta}$ is the estimator of β.

Table 1 illustrates the results of the estimated coefficients of β by the SAR model with $q = 5$, null penalty term and Gaussian noise in y, where "Exp", "Square" and "LAD" indicate the exponential squared loss, the square loss and the LAD loss, respectively. It is shown that (1) both of the three loss functions bring nonzero estimates of β_1, β_2 and β_3, which are close to the true values (the mean of the true values of β_1, β_2 and β_3 are 3.0, 2.0, 1.6, resp.). Comparatively, the model with the square loss brings the most accurate estimation. (2) In terms of MedSE, the model with the square loss performs the best. (3) In terms of the estimated value of ρ, which is indicated by $\hat{\rho}$ in Table 1, the proposed method with all of the three loss functions put out accurate estimations. (4) In terms of the estimated value of the variance of the noise of the observations, which is denoted by $\hat{\sigma}^2$, the estimations with the three loss functions are inaccurate when $n = 30$, $q = 5$, and get accurate as n increases. When $n = 120$, and $n = 360$, all the three loss functions bring accurate estimate of $\hat{\sigma}^2$.

Table 2 illustrates the results of the estimated coefficients of β on normal data when the dimension is comparatively close to the sample size. Similar results of Table 1 have been observed, except that the estimations of $\hat{\rho}$ and $\hat{\sigma}^2$ are not as accurate as that in the case of $q = 5$. As the sample size is not enough compared with the dimension, these results are as expected.

Moreover, we employ Friedman test and Nemenyi test to quantitatively compare the estimation performance of the SAR model with the three loss functions. In each case listed in Tables 1 and 2, we order the results in terms of MedSE. The loss function that gets the lowest MedSE values receives rank 1, the one that gets the second lowest MedSE receives rank 2, etc. We calculate the mean ranks of the three loss functions in all the tested cases. Under the null-hypothesis, which states that all the algorithms are equivalent, and so their mean ranks should be equal, we calculate F_F statistic (Iman and

Table 1
Estimation with no regularizer on normal data ($q = 5$).

	$n = 30,\ q = 5$			$n = 120,\ q = 5$			$n = 360,\ q = 5$		
	Exp	Square	LAD	Exp	Square	LAD	Exp	Square	LAD
$\rho_1 = 0.8,\ \sigma_1 = 1$									
β_1	4.33	2.96	3.16	2.97	3.08	3.13	3.00	3.04	3.05
β_2	2.14	2.11	2.08	2.01	2.02	2.07	2.04	2.02	1.99
β_3	1.17	1.59	1.80	1.60	1.61	1.64	1.80	1.64	1.65
$\hat{\rho}$	0.75	0.81	0.75	0.82	0.81	0.80	0.81	0.81	0.81
$\hat{\sigma}^2$	1.67	0.74	1.57	1.04	1.00	1.05	1.10	1.03	1.03
MedSE	1.61	0.69	1.06	0.23	0.37	0.47	0.21	0.20	0.24
$\rho_1 = 0.5,\ \sigma_1 = 1$									
β_1	3.10	2.93	2.96	2.89	3.03	3.05	2.93	2.97	3.00
β_2	1.40	2.10	1.92	2.00	1.98	2.00	1.96	1.98	1.98
β_3	1.56	1.58	1.75	1.56	1.56	1.59	1.73	1.61	1.61
$\hat{\rho}$	0.51	0.52	0.50	0.53	0.50	0.50	0.51	0.50	0.50
$\hat{\sigma}^2$	4.03	0.72	0.87	1.03	0.98	0.99	1.06	1.00	1.01
MedSE	0.52	0.71	0.91	0.24	0.35	0.41	0.17	0.17	0.21
$\rho_1 = 0.2,\ \sigma_1 = 1$									
β_1	3.45	2.94	2.92	2.89	3.03	3.03	2.93	2.97	3.00
β_2	0.46	2.11	2.00	2.01	1.98	2.01	1.96	1.98	1.97
β_3	1.82	1.60	1.72	1.57	1.57	1.57	1.73	1.61	1.62
$\hat{\rho}$	0.29	0.24	0.34	0.25	0.22	0.23	0.22	0.22	0.22
$\hat{\sigma}^2$	7.02	0.72	0.97	1.04	0.98	0.99	1.06	1.01	1.02
MedSE	3.29	0.69	0.98	0.24	0.33	0.40	0.17	0.17	0.22
$\rho_1 = 0,\ \sigma_1 = 1$									
β_1	3.29	2.95	2.94	2.90	3.04	3.03	2.94	2.97	3.00
β_2	0.05	2.12	1.97	2.01	1.98	2.00	1.96	1.98	1.98
β_3	1.54	1.60	1.80	1.57	1.58	1.57	1.73	1.61	1.63
$\hat{\rho}$	0.10	0.06	0.15	0.05	0.02	0.04	0.01	0.01	0.02
$\hat{\sigma}^2$	10.77	0.73	0.97	1.04	0.98	1.00	1.06	1.01	1.02
MedSE	3.50	0.69	0.93	0.24	0.33	0.42	0.17	0.17	0.21
$\rho_1 = 0.8,\ \sigma_1 = 2$									
β_1	4.34	2.88	3.16	2.84	3.06	3.18	3.02	2.99	3.05
β_2	2.22	2.19	1.94	1.91	2.01	2.05	2.02	2.01	1.96
β_3	1.13	1.57	1.87	1.64	1.58	1.58	1.84	1.63	1.66
$\hat{\rho}$	0.75	0.83	0.74	0.84	0.82	0.81	0.82	0.83	0.82
$\hat{\sigma}^2$	3.40	2.79	4.29	3.91	3.82	3.95	4.17	3.92	3.97
MedSE	1.84	1.40	1.97	0.48	0.68	0.82	0.38	0.34	0.45
$\rho_1 = 0.5,\ \sigma_1 = 2$									
β_1	3.19	2.88	2.98	2.82	3.03	3.08	2.98	2.94	3.00
β_2	1.90	2.18	1.81	1.91	1.98	2.00	1.94	1.98	1.95
β_3	1.72	1.58	1.86	1.63	1.55	1.54	1.77	1.60	1.62
$\hat{\rho}$	0.50	0.55	0.50	0.57	0.53	0.52	0.53	0.53	0.53
$\hat{\sigma}^2$	2.10	2.81	3.46	3.98	3.84	3.89	4.16	3.94	3.98
MedSE	0.63	1.39	1.74	0.48	0.66	0.74	0.36	0.34	0.42
$\rho_1 = 0.2,\ \sigma_1 = 2$									
β_1	3.03	2.89	2.90	2.83	3.04	3.06	2.99	2.95	3.02
β_2	1.84	2.20	1.91	1.93	1.98	1.99	1.95	1.99	1.95
β_3	1.85	1.58	1.85	1.64	1.56	1.54	1.78	1.61	1.63
$\hat{\rho}$	0.25	0.27	0.38	0.31	0.23	0.26	0.23	0.23	0.24
$\hat{\sigma}^2$	2.14	2.83	3.65	4.05	3.86	3.91	4.20	3.96	4.00
MedSE	0.67	1.39	1.99	0.48	0.68	0.71	0.36	0.33	0.42
$\rho_1 = 0,\ \sigma_1 = 2$									
β_1	3.12	2.91	2.92	2.85	3.05	3.07	2.99	2.96	3.02
β_2	1.91	2.20	1.87	1.94	1.99	1.98	1.96	1.99	1.95
β_3	1.75	1.60	1.91	1.65	1.57	1.54	1.79	1.61	1.65
$\hat{\rho}$	0.02	0.09	0.21	0.10	0.03	0.06	0.01	0.02	0.03
$\hat{\sigma}^2$	2.06	2.87	3.74	4.09	3.88	3.93	4.21	3.97	4.01
MedSE	0.51	1.35	2.06	0.47	0.68	0.77	0.36	0.33	0.44

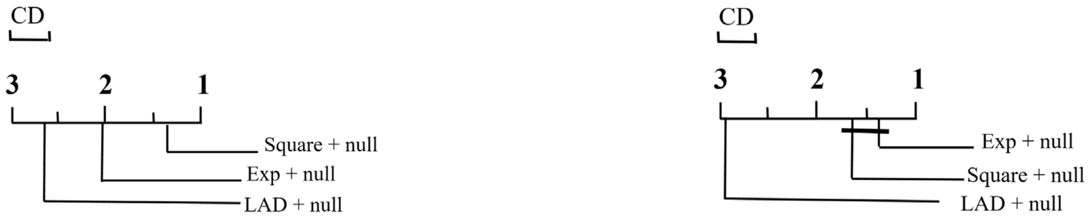
Davenport, 1980) as 38.790. The critical value for $\alpha = 0.05$ is 3.093, so we reject the null-hypothesis. We also operate Nemenyi test for pairwise comparisons, and depict the results in Fig. 1(a), where it shows that the square loss ("Square") gets the highest average rank, 1.31, the exponential squared loss ("Exp") gets the second best average rank, 2.03, and the LAD loss gets the third best average rank, 2.66. As the critical value is 0.4783 under $\alpha = 0.05$, the above results indicate that the three loss functions have different estimation performance in terms of MedSE.

Table 3 illustrates the results of the estimated coefficients of β when the observations of y have outliers. Compared with the results with normal data (Table 1), the SAR model with the exponential squared loss performs better than the square loss in terms of MedSE. We calculate the mean ranks of the three loss functions in all the tested cases in Table 3. We

Table 2
Estimation with no regularizer on normal data when the dimension is close to the sample size.

	$n = 30,\ q = 20$			$n = 120,\ q = 80$			$n = 360,\ q = 200$		
	Exp	Square	LAD	Exp	Square	LAD	Exp	Square	LAD
$\rho_1 = 0.8,\ \sigma_1 = 1$									
β_1	1.53	3.22	4.28	1.96	3.35	3.60	3.46	3.29	3.79
β_2	1.15	2.44	1.93	2.65	2.22	2.54	2.19	2.05	2.60
β_3	0.38	1.74	2.77	1.90	1.92	2.37	1.75	1.72	1.84
$\hat{\rho}$	0.52	0.65	0.50	0.53	0.70	0.50	0.74	0.75	0.50
$\hat{\sigma}^2$	21.03	0.90	8.45	8.35	1.02	10.25	1.06	0.84	12.69
MedSE	8.38	4.71	8.47	10.82	3.95	7.48	2.43	2.32	5.87
$\rho_1 = 0.5,\ \sigma_1 = 1$									
β_1	1.09	2.94	3.02	1.54	2.95	3.00	3.16	3.04	3.02
β_2	1.95	1.96	1.80	1.68	2.06	1.93	2.05	1.94	1.99
β_3	(0.37)	1.75	1.81	0.79	1.67	1.58	1.64	1.58	1.54
$\hat{\rho}$	0.49	0.49	0.50	0.49	0.50	0.50	0.49	0.50	0.50
$\hat{\sigma}^2$	17.40	0.19	0.57	9.10	0.34	0.64	0.44	0.47	0.76
MedSE	9.73	2.50	2.76	5.72	1.96	2.27	1.55	1.52	1.85
$\rho_1 = 0.2,\ \sigma_1 = 1$									
β_1	1.40	2.98	2.94	2.30	2.93	2.98	3.14	3.01	2.97
β_2	2.16	1.90	1.86	1.98	2.04	1.93	2.04	1.93	1.94
β_3	(0.39)	1.77	1.64	1.24	1.65	1.46	1.64	1.57	1.55
$\hat{\rho}$	0.47	0.32	0.50	0.34	0.31	0.50	0.24	0.27	0.50
$\hat{\sigma}^2$	13.29	0.18	0.73	2.42	0.35	0.98	0.43	0.47	1.17
MedSE	11.39	2.79	3.63	2.96	2.02	2.97	1.53	1.53	2.35
$\rho_1 = 0,\ \sigma_1 = 1$									
β_1	1.81	3.02	2.98	2.09	2.95	3.06	3.15	3.02	3.01
β_2	1.74	1.91	1.98	1.81	2.05	1.99	2.05	1.94	1.98
β_3	0.96	1.81	1.62	1.07	1.66	1.45	1.64	1.58	1.57
$\hat{\rho}$	0.44	0.20	0.50	0.17	0.10	0.50	0.04	0.06	0.50
$\hat{\sigma}^2$	10.00	0.19	0.94	4.16	0.35	1.43	0.43	0.46	1.63
MedSE	7.35	2.90	4.05	3.89	2.00	3.58	1.54	1.51	2.85
$\rho_1 = 0.8,\ \sigma_1 = 2$									
β_1	0.90	3.15	3.77	2.06	3.23	3.59	3.52	3.23	3.82
β_2	1.72	2.34	2.17	2.47	2.24	2.46	2.18	1.97	2.59
β_3	0.60	1.91	2.67	2.03	1.97	2.46	1.55	1.72	1.75
$\hat{\rho}$	0.55	0.67	0.50	0.53	0.72	0.50	0.73	0.77	0.50
$\hat{\sigma}^2$	22.44	1.46	15.42	6.34	2.04	13.85	3.50	2.12	17.79
MedSE	15.18	6.51	10.72	9.44	5.01	9.92	4.33	3.37	7.42
$\rho_1 = 0.5,\ \sigma_1 = 2$									
β_1	1.28	2.90	2.99	2.11	2.90	3.04	3.18	3.04	3.05
β_2	2.38	1.90	1.67	2.15	2.10	1.84	2.13	1.90	1.99
β_3	0.27	1.89	1.93	1.51	1.73	1.56	1.56	1.60	1.51
$\hat{\rho}$	0.49	0.50	0.50	0.48	0.51	0.50	0.50	0.51	0.50
$\hat{\sigma}^2$	8.98	0.64	2.18	2.45	1.29	2.40	2.05	1.79	2.88
MedSE	9.56	4.89	5.51	3.61	3.79	4.53	3.14	2.90	3.55
$\rho_1 = 0.2,\ \sigma_1 = 2$									
β_1	2.26	2.91	2.97	2.05	2.87	3.06	3.12	3.00	3.00
β_2	2.00	1.87	1.56	2.17	2.08	1.84	2.17	1.89	1.93
β_3	0.86	1.89	1.88	1.36	1.71	1.44	1.63	1.59	1.54
$\hat{\rho}$	0.45	0.31	0.50	0.42	0.32	0.50	0.28	0.28	0.50
$\hat{\sigma}^2$	7.13	0.63	2.04	5.00	1.30	2.77	2.08	1.80	3.26
MedSE	6.85	5.25	6.51	4.80	3.78	4.88	3.08	2.95	3.84
$\rho_1 = 0,\ \sigma_1 = 2$									
β_1	2.36	2.96	3.02	2.04	2.90	3.11	3.14	3.02	3.04
β_2	1.59	1.88	1.65	2.14	2.10	1.88	2.17	1.90	1.97
β_3	1.60	1.93	1.83	1.28	1.72	1.39	1.62	1.60	1.57
$\hat{\rho}$	0.46	0.22	0.50	0.31	0.11	0.50	0.05	0.06	0.50
$\hat{\sigma}^2$	6.09	0.67	2.35	5.98	1.30	3.20	1.97	1.79	3.80
MedSE	7.50	5.32	6.92	5.20	3.80	5.61	3.05	2.94	4.17

calculate F_F statistic as 54.815, much greater than the critical value of 3.200 for $\alpha = 0.05$, so we reject the null-hypothesis. I.e., the three loss functions have different performance. We also operate Nemenyi test for pairwise comparisons, and depict the results in Fig. 1(b), where it shows that the exponential squared loss ("Exp") gets the highest average rank, 1.40, the square loss ("Square") gets the second best average rank, 1.65, and the LAD loss gets the third best average rank, 2.96. These results are in line with expectations, as the exponential squared loss is more robust with outliers in the response. However, when the critical value is 0.6764 under $\alpha = 0.05$, the above results indicating that the exponential squared loss and the square loss functions have no evident difference in terms of MedSE $(1.65 - 1.40 < 0.6764)$.



(a) with normal data, corresponding to Table 1-2.

(b) with outliers in y, corresponding to Table 3.

Fig. 1. Nemenyi test for the estimation results in various cases, confidence level: $\alpha = 0.05$.

We list the results of the estimated coefficients of β with inaccurate weight matrix W in Table 4. All the simulated data are generated with $\rho_1 = 0.5$, $\sigma_1 = 1$. Compared with the results with normal data (Table 1), the MedSE values increase, and the estimations of $\hat{\sigma}^2$ and $\hat{\rho}$ become worse for each loss functions in total. Particularly, (1) For removing a certain part (30%, 50% and 80%) of nonzero weights of the matrix W, MedSE increases as the moving nonzeros increase, and decrease as the sample size n increases for each of the three loss functions. The exponential squared loss has lowest MedSE among the three loss functions. (2) For adding 50% nonzero weights in each row of W, both of the three methods have comparatively high MedSE among all the tested cases of inaccurate W. Note that in this case, the added nonzeros are randomized in each row of W, the produced weight matrix has large variations compared with the true W. (3) For adding 100%, 200% nonzero weights in 10% rows of W, the estimations are comparatively accurate. The exponential squared loss achieves lower MedSE than the other two loss functions, especially when the sample size n is much larger than the dimension. In brief, we find that the SAR estimation with the exponential squared loss is robust for inaccurate estimations of W.

Table 5 lists the variable section results of with Lasso and adaptive Lasso regularizer on normal data with $q = 5$, where E+ℓ_1, E+$\tilde{\ell}_1$ indicate the SAR model with the **E**xponential loss and Lasso penalty, the **E**xponential loss and adaptive Lasso penalty, respectively, S+ℓ_1, S+$\tilde{\ell}_1$ indicate the SAR model with the **S**quare loss and Lasso penalty, the **S**quare loss and adaptive Lasso penalty, respectively, L+ℓ_1, L+$\tilde{\ell}_1$ indicate the SAR model with the **L**AD loss and Lasso penalty, the **L**AD loss and adaptive Lasso penalty, respectively. We do not list the results of the case $n = 120, q = 5$ for layout, as we find the results are similar to the case of $n = 360, q = 5$. It is shown that almost in all the tested cases, the SAR model with the exponential squared loss and the Lasso penalty or the adaptive Lasso penalty (i.e., E+ℓ_1, E+$\tilde{\ell}_1$) identifies much more numbers of true zero coefficient of β ("Correct") and much lower MedSE. Moreover, all the compared penalties and loss functions achieve near-zero "Incorrect" numbers, accurate estimation of $\hat{\rho}$.

Similar results have been observed when the sample dimension is close to the sample size, which is listed in Table 6. The difference with the results of $q = 5$ (Table 5) is that the superiority of the exponential squared loss and the Lasso penalty and the adaptive Lasso penalty (i.e., E+ℓ_1, E+$\tilde{\ell}_1$) is more evident. In the tested cases of $n = 360, q = 200$, the SAR model with E+ℓ_1, E+$\tilde{\ell}_1$ almost correctly identifies all the 200 zero coefficients of β and with none incorrect nonzero coefficients of β, the MedSE of E+$\tilde{\ell}_1$ is merely $1/15 \sim 1/4$ of the best MedSE with S+ℓ_1, S+$\tilde{\ell}_1$, L+ℓ_1, L+$\tilde{\ell}_1$. The above variable selection performance of the proposed exponential squared loss and Lasso or adaptive Lasso penalty is beyond expectations.

Statistically, we calculate F_F statistic as 58.927 for "Correct", and 98.336 for "MedSE" much greater than the common critical value of 1.8719 for $\alpha = 0.05$, so we reject the null-hypothesis. I.e., the compared methods have different performance in the cases of Tables 5–6. We also operate Nemenyi test for pairwise comparisons, and depict the results in Fig. 2, where it verifies that the SAR model with the exponential squared loss and adaptive Lasso penalty ("Exp+ adaptive-ℓ_1") and with the exponential squared loss and Lasso penalty ("Exp+ℓ_1") have evident higher average ranks, (1.22, 1.84 resp. for 'Correct', and 1.31, 2.18 resp. for 'MedSE') than the other four methods. The above statistical analysis includes results in the cases of $n = 120, q = 5$ which are not listed in Tables 5–6.

Table 7 lists the variable selections results with outliers in the observations of y. It illustrates that almost in all the tested cases, the SAR model with the exponential squared loss and the Lasso penalty or the adaptive Lasso penalty (i.e., E+ℓ_1, E+$\tilde{\ell}_1$) identifies much more numbers of true zero coefficient of β ("Correct") and has much lower MedSE. Compared with the results in the normal cases (Table 5), the superiority of E+ℓ_1 and E+$\tilde{\ell}_1$ is more evident. The SAR model with E+ℓ_1, E+$\tilde{\ell}_1$ almost correctly identifies all the 5 zero coefficients of β in most tested cases, the MedSE of E+$\tilde{\ell}_1$ is about $1/5 \sim 1/3$ of the best MedSE achieved by S+ℓ_1, S+$\tilde{\ell}_1$, L+ℓ_1, L+$\tilde{\ell}_1$ in the cases of $n = 360, q = 5$. These results prove that the proposed SAR method with exponential squared loss and Lasso or adaptive Lasso penalty is more robust and efficient for variable selection in the cases of outliers existed in the observations of y.

Statistically, we also operate Nemenyi test for pairwise comparisons, and depict the results in Fig. 3, where it verifies that the SAR model with the exponential squared loss and adaptive Lasso penalty ("Exp+ adaptive-ℓ_1") and with the exponential squared loss and Lasso penalty ("Exp+ℓ_1") have evident higher average ranks, (1.33, 1.87 resp. for 'Correct', and 1.19, 1.90 resp. for 'MedSE') than the other four methods. The above statistical analysis includes results in the cases of $n = 120, q = 5$ which are not listed in Table 7.

Table 3
Estimation with no regularizer when the observations of y have outliers.

	$n = 30,\ q = 5$			$n = 120,\ q = 5$			$n = 360,\ q = 5$		
	Exp	Square	LAD	Exp	Square	LAD	Exp	Square	LAD
$\rho_1 = 0.8,\ \sigma_1 = 1,\ \delta_1 = 0.01$									
β_1	3.34	3.02	3.08	2.96	3.05	3.06	2.98	3.07	3.07
β_2	1.73	2.03	2.04	1.97	2.03	2.11	2.02	2.00	2.05
β_3	1.68	1.66	1.63	1.55	1.56	1.63	1.80	1.60	1.62
$\hat{\rho}$	0.79	0.79	0.78	0.81	0.80	0.80	0.80	0.80	0.80
$\hat{\sigma}^2$	0.97	0.92	1.21	1.02	1.09	1.12	1.19	1.09	1.14
MedSE	0.55	0.55	0.68	0.29	0.33	0.43	0.22	0.22	0.28
$\rho_1 = 0.5,\ \sigma_1 = 1,\ \delta_1 = 0.01$									
β_1	3.12	2.98	2.99	2.92	2.99	3.01	2.92	3.01	2.99
β_2	1.67	1.99	1.95	1.95	2.01	2.07	1.94	1.96	2.02
β_3	1.74	1.62	1.59	1.54	1.53	1.58	1.73	1.57	1.58
$\hat{\rho}$	0.50	0.49	0.51	0.54	0.50	0.51	0.50	0.50	0.50
$\hat{\sigma}^2$	0.94	0.92	1.13	1.05	1.12	1.12	1.19	1.11	1.15
MedSE	0.49	0.55	0.64	0.32	0.34	0.41	0.18	0.19	0.27
$\rho_1 = 0.2,\ \sigma_1 = 1,\ \delta_1 = 0.01$									
β_1	3.07	2.97	2.97	2.91	2.99	3.01	2.92	3.00	2.98
β_2	1.66	1.98	2.00	1.96	2.01	2.06	1.94	1.96	2.03
β_3	1.77	1.62	1.57	1.54	1.53	1.60	1.73	1.57	1.58
$\hat{\rho}$	0.22	0.21	0.28	0.28	0.21	0.24	0.21	0.21	0.22
$\hat{\sigma}^2$	0.92	0.91	1.11	1.04	1.09	1.11	1.18	1.10	1.14
MedSE	0.49	0.52	0.65	0.33	0.33	0.43	0.18	0.20	0.26
$\rho_1 = 0.5,\ \sigma_1 = 2,\ \delta_1 = 0.01$									
β_1	3.43	2.96	2.94	2.85	2.97	3.00	2.96	2.99	2.97
β_2	1.33	1.95	1.89	1.87	2.03	2.03	1.93	1.93	2.04
β_3	2.13	1.66	1.60	1.59	1.48	1.62	1.79	1.56	1.56
$\hat{\rho}$	0.51	0.51	0.51	0.58	0.51	0.53	0.52	0.52	0.52
$\hat{\sigma}^2$	3.77	3.30	3.97	3.82	3.95	3.97	4.24	3.94	4.11
MedSE	1.50	0.94	1.30	0.55	0.62	0.72	0.35	0.39	0.53
$\rho_1 = 0.8,\ \sigma_1 = 1,\ \delta_1 = 0.05$									
β_1	3.55	2.98	3.02	2.98	2.98	3.04	2.84	3.01	2.96
β_2	1.61	2.04	1.96	1.82	1.98	2.07	1.93	1.94	2.08
β_3	1.39	1.55	1.56	1.49	1.56	1.70	1.78	1.54	1.60
$\hat{\rho}$	0.79	0.82	0.81	0.85	0.83	0.81	0.81	0.83	0.82
$\hat{\sigma}^2$	2.94	3.61	4.10	3.82	4.10	4.19	4.33	4.31	4.31
MedSE	0.80	0.99	1.23	0.76	0.55	0.91	0.39	0.37	0.45
$\rho_1 = 0.5,\ \sigma_1 = 1,\ \delta_1 = 0.05$									
β_1	3.35	2.96	3.02	2.96	2.96	3.00	2.81	2.98	2.95
β_2	1.60	2.02	1.89	1.82	1.98	2.03	1.89	1.93	2.05
β_3	1.46	1.55	1.57	1.50	1.56	1.68	1.75	1.54	1.58
$\hat{\rho}$	0.50	0.53	0.54	0.60	0.55	0.54	0.50	0.55	0.55
$\hat{\sigma}^2$	2.85	3.65	4.09	3.93	4.19	4.21	4.38	4.39	4.40
MedSE	0.71	1.00	1.22	0.77	0.53	0.90	0.37	0.36	0.47
$\rho_1 = 0.2,\ \sigma_1 = 1,\ \delta_1 = 0.05$									
β_1	3.31	2.95	2.97	2.96	2.98	2.99	2.81	3.00	2.94
β_2	1.60	2.02	1.92	1.83	1.98	2.03	1.90	1.94	2.05
β_3	1.48	1.54	1.59	1.51	1.56	1.67	1.76	1.54	1.58
$\hat{\rho}$	0.21	0.25	0.39	0.36	0.27	0.30	0.19	0.27	0.28
$\hat{\sigma}^2$	2.82	3.66	4.17	4.01	4.21	4.26	4.37	4.42	4.42
MedSE	0.71	0.97	1.20	0.81	0.53	0.89	0.38	0.36	0.46
$\rho_1 = 0.5,\ \sigma_1 = 2,\ \delta_1 = 0.05$									
β_1	3.46	2.95	2.95	2.93	2.95	3.01	2.84	2.98	2.93
β_2	1.43	1.98	1.85	1.71	1.99	2.08	1.88	1.91	2.03
β_3	1.76	1.58	1.55	1.50	1.50	1.65	1.84	1.52	1.60
$\hat{\rho}$	0.48	0.55	0.55	0.63	0.56	0.55	0.52	0.57	0.56
$\hat{\sigma}^2$	4.93	5.79	6.66	6.14	6.83	6.74	7.28	6.99	7.10
MedSE	0.98	1.31	1.64	0.94	0.75	1.01	0.45	0.49	0.63

Table 8 lists the variable selections results with inaccurate weight matrix W. We are particularly interested in this case as the weight matrix is practically hard to be accurately estimated. (1) Almost in all the tested cases, the SAR model with the exponential squared loss and the Lasso penalty or the adaptive Lasso penalty (i.e., E+ℓ_1, E+$\tilde{\ell}_1$) identifies much more numbers of true zero coefficient of β ("Correct") and has much lower MedSE. Compared with the estimation results with no penalty (Table 4), the advantage in terms of MedSE is more obvious in all the tested cases, including removing a certain part (30%, 50% and 80%) of nonzero weights of the matrix W, adding a certain number of nonzero weights in each row of W, adding 100%, 200% nonzero weights in 10% rows of W. (2) We checked results in the cases of $n = 120, q = 5$, which are not listed in the table for layout. The results are similar to that of $n = 360, q = 5$ and support the superiority

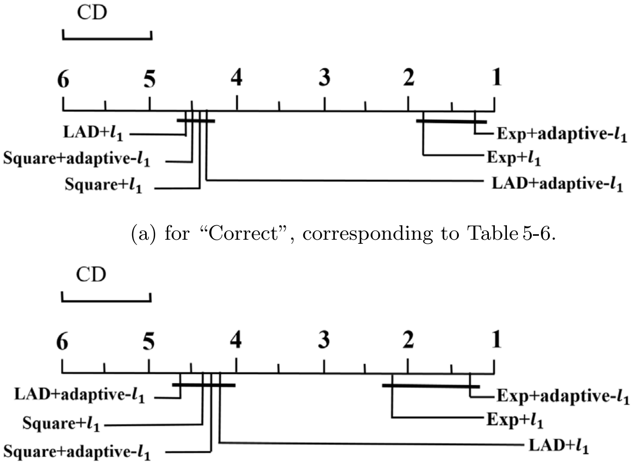
Table 4
Estimation with no regularizer with noisy weighting matrix W.

	$n = 30,\ q = 5$			$n = 120,\ q = 5$			$n = 360,\ q = 5$		
	Exp	Square	LAD	Exp	Square	LAD	Exp	Square	LAD
Remove 30% nonzero weights									
β_1	3.11	2.97	3.06	2.91	3.07	2.93	3.01	3.01	2.99
β_2	1.92	2.04	1.97	1.91	2.00	1.97	1.94	2.03	2.00
β_3	1.57	1.59	1.52	1.63	1.59	1.70	1.80	1.63	1.61
$\hat{\rho}$	0.44	0.43	0.44	0.46	0.43	0.43	0.43	0.43	0.43
$\hat{\sigma}^2$	1.10	1.02	1.13	1.18	1.07	1.10	1.15	1.10	1.10
MedSE	0.69	0.59	0.73	0.40	0.36	0.55	0.21	0.25	0.29
Remove 50% nonzero weights									
β_1	3.22	3.00	3.04	2.93	3.09	2.93	2.92	3.03	3.03
β_2	2.10	2.06	2.02	1.81	2.03	2.00	1.95	2.05	2.02
β_3	1.38	1.58	1.49	1.68	1.61	1.65	1.82	1.64	1.64
$\hat{\rho}$	0.37	0.36	0.38	0.39	0.37	0.37	0.38	0.36	0.37
$\hat{\sigma}^2$	1.35	1.28	1.42	1.48	1.32	1.34	1.43	1.36	1.34
MedSE	0.57	0.71	0.85	0.38	0.41	0.62	0.28	0.28	0.37
Remove 80% nonzero weights									
β_1	3.34	3.09	2.97	2.85	3.11	3.00	3.01	3.09	3.12
β_2	1.90	2.02	2.02	1.82	2.11	1.98	2.00	2.08	2.08
β_3	1.50	1.64	1.67	1.86	1.60	1.66	1.77	1.68	1.65
$\hat{\rho}$	0.29	0.24	0.27	0.32	0.26	0.27	0.25	0.24	0.26
$\hat{\sigma}^2$	1.76	2.03	2.13	2.15	2.09	2.07	2.32	2.17	2.11
MedSE	0.67	0.81	0.99	0.56	0.56	0.70	0.32	0.36	0.48
Add 50% nonzero weights									
β_1	4.17	3.19	3.13	3.54	3.24	3.17	3.03	3.19	3.21
β_2	1.84	2.13	2.05	1.98	2.22	2.07	2.29	2.16	2.15
β_3	1.58	1.58	1.81	1.80	1.63	1.77	1.96	1.71	1.70
$\hat{\rho}$	0.10	0.28	0.34	0.31	0.28	0.33	0.27	0.29	0.30
$\hat{\sigma}^2$	3.91	1.92	2.00	2.38	2.02	1.93	2.20	1.91	1.96
MedSE	1.52	0.96	1.18	0.94	0.77	1.00	0.58	0.45	0.61
Add 100% nonzero weights on 10% rows									
β_1	3.30	3.07	3.01	2.91	3.05	3.04	2.99	2.99	3.02
β_2	1.58	1.97	2.04	1.99	2.01	2.04	1.92	2.04	2.05
β_3	1.72	1.58	1.52	1.59	1.59	1.56	1.81	1.61	1.61
$\hat{\rho}$	0.47	0.48	0.50	0.52	0.49	0.50	0.51	0.50	0.49
$\hat{\sigma}^2$	1.03	0.92	1.02	1.02	1.01	1.00	1.06	0.98	0.96
MedSE	0.83	0.67	0.61	0.27	0.40	0.39	0.25	0.25	0.23
Add 200% nonzero weights on 10% rows									
β_1	3.23	3.08	2.98	2.93	3.04	3.02	2.97	2.99	3.02
β_2	1.56	1.97	2.07	1.99	2.02	2.04	1.91	2.04	2.06
β_3	1.76	1.59	1.54	1.57	1.59	1.57	1.82	1.61	1.61
$\hat{\rho}$	0.49	0.48	0.50	0.53	0.49	0.50	0.51	0.50	0.50
$\hat{\sigma}^2$	1.00	0.92	1.02	1.02	1.02	1.00	1.05	0.98	0.96
MedSE	0.80	0.70	0.59	0.27	0.40	0.42	0.23	0.25	0.24

of SAR with E+ℓ_1 and E+$\tilde{\ell}_1$ in terms of "Correct" and MedSE. (3) For adding 50% nonzero weights in each row of W, SAR with E+ℓ_1 and E+$\tilde{\ell}_1$ are still better than the baseline methods in terms of "Correct" and MedSE. Note that in the case of $n = 30, q = 5$, although E+ℓ_1 and E+$\tilde{\ell}_1$ have higher MedSE (1.06, 1.10, resp.) than S+ℓ_1, with the MedSE value of 0.95, E+ℓ_1 and E+$\tilde{\ell}_1$ have much better variable selection performance in fact: they have correctly identified 4.23, 5.00 zero coefficients of β ("Correct") out of 5 nonzeros in total, compared with the 1.87, 1.97, 1.97 and 1.70 for S+ℓ_1, S+$\tilde{\ell}_1$, L+ℓ_1 and L+$\tilde{\ell}_1$, respectively. Statistically, we also operate Nemenyi test for comparisons of the methods, and depict the results in Fig. 4, where it verifies that the SAR model with the exponential squared loss and adaptive Lasso penalty ("Exp+ adaptive-ℓ_1") and with the exponential squared loss and Lasso penalty ("Exp+ℓ_1") have evident higher average ranks, (1.33, 1.88 resp. for 'Correct', and 1.18, 1.90 resp. for 'MedSE') than the other four methods. The above statistical analysis includes results in the cases of $n = 120, q = 5$ which are not listed in Table 8. These encouraging results illustrate that the proposed SAR method with exponential squared loss and Lasso or adaptive Lasso penalty is robust and particularly applicable for variable selection in spatial auto-regression when the weight matrix W could not be accurately estimated.

6. Real data example

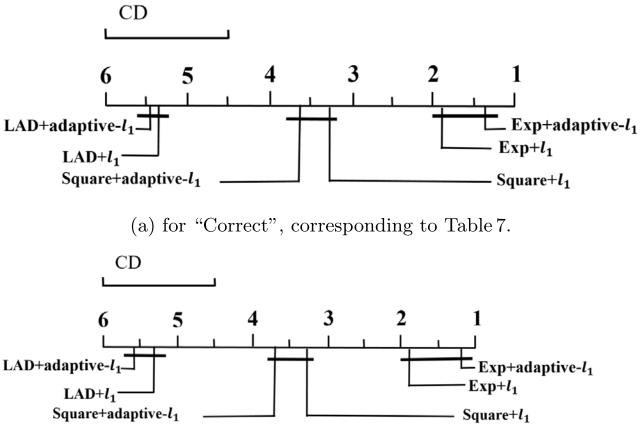
In this section, we give a real example to validate the performance of proposed variable selection procedures for the SAR model.



(a) for “Correct”, corresponding to Table 5-6.

(b) for “MedSE”, corresponding to Table 5-6.

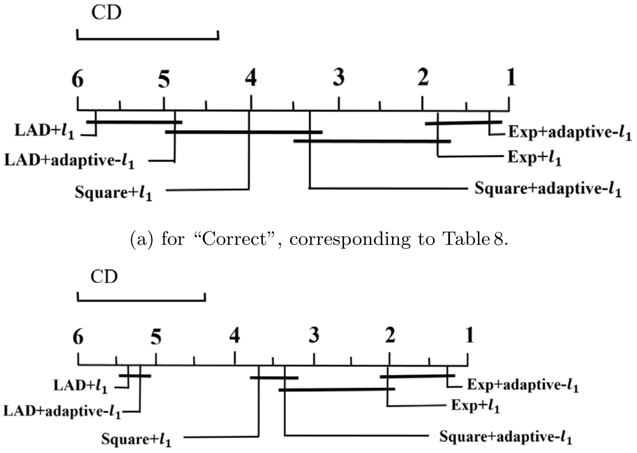
Fig. 2. Nemenyi test with regularizers in normal case, confidence level: $\alpha = 0.05$.



(a) for “Correct”, corresponding to Table 7.

(b) for “MedSE”, corresponding to Table 7.

Fig. 3. Nemenyi test with regularizers with outliers in y, confidence level: $\alpha = 0.05$.



(a) for “Correct”, corresponding to Table 8.

(b) for “MedSE”, corresponding to Table 8.

Fig. 4. Nemenyi test with regularizers with inaccurate weighting matrix W, confidence level: $\alpha = 0.05$.

Table 5
Variable section with regularizer on normal data ($q = 5$), E: the Exponential loss, S: the Square loss, L: the LAD loss, $\tilde{\ell}_1$: the adaptive Lasso penalty.

	$n = 30,\ q = 5$						$n = 360,\ q = 5$					
	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$
$\rho_1 = 0.8,\ \sigma_1 = 1$												
Correct	4.27	4.77	2.50	2.47	1.83	1.97	5.00	5.00	4.93	4.93	4.83	4.80
Incorrect	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.78	0.77	0.79	0.80	0.78	0.77	0.81	0.81	0.81	0.81	0.80	0.81
MedSE	0.92	0.95	0.78	0.72	1.02	0.91	0.16	0.14	0.20	0.21	0.28	0.24
$\rho_1 = 0.5,\ \sigma_1 = 1$												
Correct	4.03	4.63	3.13	2.50	2.37	2.23	5.00	5.00	4.97	4.97	4.80	4.90
Incorrect	0.13	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.50	0.49	0.50	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.50	0.50
MedSE	0.31	0.15	0.63	0.73	0.81	0.86	0.12	0.05	0.18	0.18	0.22	0.21
$\rho_1 = 0.2,\ \sigma_1 = 1$												
Correct	3.23	3.17	2.97	2.47	2.37	2.37	5.00	5.00	4.97	4.97	4.73	4.90
Incorrect	0.10	0.27	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.25	0.24	0.23	0.23	0.32	0.34	0.23	0.22	0.21	0.21	0.22	0.22
MedSE	0.18	1.56	0.65	0.75	0.78	0.81	0.12	0.05	0.18	0.18	0.20	0.21
$\rho_1 = 0,\ \sigma_1 = 1$												
Correct	4.83	4.10	3.03	2.43	2.40	2.30	5.00	5.00	4.97	4.97	4.80	4.90
Incorrect	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.01	0.02	0.06	0.05	0.15	0.15	0.01	0.01	0.01	0.01	0.02	0.02
MedSE	0.15	0.10	0.64	0.73	0.84	0.83	0.12	0.05	0.18	0.18	0.22	0.21
$\rho_1 = 0.8,\ \sigma_1 = 2$												
Correct	4.27	4.33	1.60	1.50	1.20	1.40	4.90	5.00	4.03	4.30	3.70	4.27
Incorrect	0.00	0.20	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.79	0.78	0.81	0.82	0.78	0.77	0.83	0.82	0.82	0.82	0.82	0.82
MedSE	0.93	1.42	1.34	1.45	1.68	1.50	0.26	0.17	0.37	0.37	0.49	0.43
$\rho_1 = 0.5,\ \sigma_1 = 2$												
Correct	4.90	5.00	1.60	1.67	1.40	1.60	4.27	5.00	4.10	4.27	3.73	4.20
Incorrect	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.51	0.49	0.51	0.54	0.51	0.53	0.53	0.53	0.53	0.53	0.53	0.53
MedSE	0.31	0.29	1.19	1.41	1.33	1.45	0.25	0.13	0.35	0.36	0.42	0.42
$\rho_1 = 0.2,\ \sigma_1 = 2$												
Correct	4.83	5.00	1.50	1.63	1.53	1.57	4.37	5.00	4.07	4.23	3.77	4.17
Incorrect	0.00	0.00	0.00	0.00	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.23	0.20	0.23	0.27	0.37	0.40	0.23	0.23	0.23	0.23	0.24	0.24
MedSE	0.27	0.29	1.20	1.43	1.31	1.49	0.25	0.14	0.36	0.36	0.44	0.43
$\rho_1 = 0,\ \sigma_1 = 2$												
Correct	5.00	5.00	1.47	1.60	1.37	1.60	4.77	5.00	4.07	4.23	3.80	4.23
Incorrect	0.00	0.00	0.00	0.00	0.10	0.03	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.01	0.00	0.08	0.10	0.23	0.25	0.01	0.00	0.02	0.02	0.03	0.03
MedSE	0.27	0.32	1.25	1.44	1.42	1.51	0.25	0.14	0.36	0.35	0.43	0.41

6.1. The sample data

We consider a popular data set in the Boston Standard Metropolitan Statistical Area, which has been used by many authors, for example, Harrison Jr. and Rubinfeld (1978), Pace and Gilley (1997), Tang (2014) and Xuan et al. (2018), and so on.

The database can be found in the spdep library of R. Descriptions of the variables we employed are given in Table 6 of Xuan et al. (2018). Our interest is to investigate the relationship between the home value and the other variables and select several important variables to explain the home value. Following the analysis of Harrison Jr. and Rubinfeld (1978), the dependent variable is the log(home value) and the other variables are independent variables, where the distance to the employment, highway access index, lower status population proportion are processed with logarithm and the number of rooms, nitrogen oxide concentrations are processed with the square operation. For convenient analysis, all the variables are centralized so that their sample means become zero. Xuan et al. (2018) test spatial dependence by a Moran's I statistic (Moran I) which was widely used in spatial data and the value of test statistics suggested that the SAR model was a better choice than the spatial error model to fit the Boston sample. Thus, the data can serve the purpose of our analysis and the SAR model will be used to perform variable selection procedure.

The spatial weight matrix is generally set by two kinds of information. One is set by latitude and longitude coordinates, and the other is through the relative location of the region. In this work, we construct both types of the weight matrices and particularly interested to verify the impact of the weight matrix for the variable selection results. Following Xuan et al. (2018), the weight of the spatial weight matrix is set 1 if they share a common boundary, and zero otherwise. In addition, the spatial weight matrix is row-normalized as usually done in the practice. Moreover, we also remove 50% nonzero weights in each row of the weight matrix, or add 100% nonzero weights in 10% rows of the matrix W generated

Table 6
Variable section with regularizer on normal data when the dimension is close to the sample size, E: the Exponential loss, S: the Square loss, L: the LAD loss, $\tilde{\ell}_1$: the adaptive Lasso penalty.

	$n = 30,\ q = 20$						$n = 360,\ q = 200$					
	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$
$\rho_1 = 0.8,\ \sigma_1 = 1$												
Correct	10.63	14.57	3.17	3.43	5.10	5.50	196.80	200.00	160.67	159.40	148.10	139.07
Incorrect	0.87	1.10	0.07	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.59	0.63	0.64	0.64	0.58	0.59	0.78	0.81	0.75	0.75	0.66	0.65
MedSE	3.90	4.23	5.01	4.36	2.70	2.68	0.78	0.14	2.18	2.15	2.41	2.72
$\rho_1 = 0.5,\ \sigma_1 = 1$												
Correct	13.90	17.33	5.50	5.40	10.13	9.50	200.00	200.00	187.70	186.90	193.67	192.60
Incorrect	0.07	0.07	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.47	0.42	0.51	0.50	0.50	0.51	0.49	0.50	0.50	0.50	0.50	0.50
MedSE	1.18	0.91	2.72	2.43	1.49	1.56	0.69	0.10	1.54	1.52	1.33	1.36
$\rho_1 = 0.2,\ \sigma_1 = 1$												
Correct	13.93	17.87	4.67	5.63	9.10	9.03	200.00	200.00	186.03	187.63	186.37	186.40
Incorrect	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.16	0.10	0.40	0.39	0.47	0.46	0.22	0.21	0.27	0.26	0.48	0.47
MedSE	1.18	0.84	3.08	2.57	1.55	1.66	0.68	0.11	1.54	1.52	1.57	1.58
$\rho_1 = 0,\ \sigma_1 = 1$												
Correct	13.43	16.30	4.57	5.30	8.50	8.30	200.00	200.00	186.67	188.43	182.47	182.13
Incorrect	0.10	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.01	0.00	0.28	0.30	0.42	0.41	0.00	0.00	0.06	0.04	0.39	0.37
MedSE	1.20	0.85	3.26	2.71	1.69	1.79	0.69	0.10	1.53	1.51	1.70	1.67
$\rho_1 = 0.8,\ \sigma_1 = 2$												
Correct	11.23	15.77	2.27	2.17	4.80	4.53	177.93	197.17	119.00	119.53	124.57	114.53
Incorrect	0.23	0.33	0.13	0.07	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.77	0.74	0.69	0.71	0.60	0.60	0.80	0.82	0.77	0.77	0.65	0.63
MedSE	2.52	2.36	7.43	6.19	3.13	3.24	1.68	0.56	3.42	3.38	3.29	3.55
$\rho_1 = 0.5,\ \sigma_1 = 2$												
Correct	11.00	14.27	2.60	2.77	7.03	6.43	180.63	197.63	130.50	134.13	158.17	158.70
Incorrect	0.00	0.00	0.13	0.07	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.54	0.49	0.55	0.55	0.50	0.51	0.51	0.53	0.52	0.51	0.50	0.50
MedSE	2.48	2.18	5.75	4.85	2.11	2.26	1.61	0.55	2.91	2.96	2.28	2.24
$\rho_1 = 0.2,\ \sigma_1 = 2$												
Correct	10.87	14.93	2.97	2.63	7.43	6.70	182.60	198.03	129.13	134.77	152.63	153.60
Incorrect	0.00	0.00	0.20	0.10	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.30	0.09	0.39	0.42	0.47	0.48	0.23	0.24	0.29	0.27	0.50	0.50
MedSE	2.50	2.22	5.50	4.67	2.19	2.43	1.59	0.55	2.98	2.95	2.41	2.39
$\rho_1 = 0,\ \sigma_1 = 2$												
Correct	10.63	13.97	2.67	2.87	6.33	6.17	180.03	197.87	129.93	134.50	147.10	146.97
Incorrect	0.33	0.07	0.17	0.07	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.05	0.01	0.27	0.34	0.43	0.42	0.00	0.02	0.06	0.04	0.49	0.47
MedSE	2.51	2.17	5.69	5.01	2.26	2.45	1.61	0.55	2.95	2.95	2.56	2.54

by the method of relative location (Method 1). Then the spatial weight matrix is row-normalized and put into the SAR model to verify the effect of different estimations of weight matrix W.

6.2. Model selection and estimation

Table 9 presents the variable selection results of and SAR model with the exponential squared loss, the squared loss and the LAD loss estimate via the Lasso, adaptive Lasso and null penalties. It shows that all the three estimators of the SAR with null penalties (E+ null, S+ null, L+ null) agree that the residential land proportion, tax rate and black proportion are unimportant (the absolute value of the coefficient is less than 0.0006) and some important variables (the absolute value of the coefficient is greater than 0.05) include the nitrogen oxide concentrations, distance to the employment center, highway access index and lower status population proportion. The estimated values of the penalized exponential squared loss with the Lasso and adaptive Lasso penalty (E+ℓ_1, E+$\tilde{\ell}_1$) for the SAR model are similar that the coefficients of the location contiguous to the Charles River shrink to zero, except that there are two more zeros under the adaptive Lasso than the Lasso, that is the value of the residential land proportion, the proportion of nonretail business location.

We are interested to study the effects of an inaccurate weight matrix W in the SAR model. We list in Table 10 the variable selection results with the weight matrix W removed 50% nonzeros on each row, and in Table 11 the results with the weight matrix W removed 100% nonzeros on 10% rows. In both of the two perturbations of the matrix W, the BIC values of the estimators increase than the normal case, indicating that the weight matrix W indeed affects the performance of the SAR model in a certain degree. But the identified important and unimportant variables hold almost the same. In both Tables 10 and 11, exponential squared loss with adaptive Lasso (i.e., E+$\tilde{\ell}_1$) achieves minimum BIC values, indicating the robustness and efficiency in spatial regression when the estimated weighting matrix is not accurate.

Table 7
Variable selection with regularizer when the observations y have outliers, E: the Exponential loss, S: the Square loss, L: the LAD loss, $\tilde{\ell}_1$: the adaptive Lasso penalty.

	$n = 30,\ q = 5$						$n = 360,\ q = 5$					
	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$
$\rho_1 = 0.8,\ \sigma_1 = 1,\ \delta_1 = 0.01$												
Correct	5.00	5.00	3.83	3.83	2.87	3.30	5.00	5.00	4.97	5.00	4.90	4.77
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.80	0.79	0.80	0.80	0.79	0.79	0.80	0.80	0.80	0.80	0.80	0.80
MedSE	0.30	0.27	0.48	0.50	0.56	0.59	0.13	0.07	0.21	0.22	0.26	0.26
$\rho_1 = 0.5,\ \sigma_1 = 1,\ \delta_1 = 0.01$												
Correct	5.00	5.00	3.70	3.77	3.33	3.50	5.00	5.00	4.97	5.00	4.90	4.80
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.50	0.50	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.50	0.51	0.50
MedSE	0.29	0.25	0.45	0.49	0.55	0.54	0.11	0.04	0.19	0.20	0.23	0.25
$\rho_1 = 0.2,\ \sigma_1 = 1,\ \delta_1 = 0.01$												
Correct	5.00	5.00	3.87	3.77	3.23	3.40	5.00	5.00	5.00	5.00	4.87	4.87
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.22	0.22	0.23	0.23	0.28	0.26	0.22	0.21	0.21	0.22	0.22	0.21
MedSE	0.28	0.25	0.48	0.50	0.57	0.55	0.11	0.04	0.19	0.19	0.23	0.24
$\rho_1 = 0.5,\ \sigma_1 = 2,\ \delta_1 = 0.01$												
Correct	2.57	5.00	2.47	2.43	2.20	2.10	5.00	5.00	4.33	4.17	3.70	3.87
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.51	0.51	0.52	0.53	0.53	0.52	0.53	0.52	0.51	0.53	0.53	0.51
MedSE	0.80	0.49	0.88	0.90	0.97	1.02	0.24	0.10	0.36	0.37	0.45	0.44
$\rho_1 = 0.8,\ \sigma_1 = 1,\ \delta_1 = 0.05$												
Correct	4.93	5.00	2.43	2.27	1.97	1.93	4.70	5.00	4.03	4.20	3.63	4.07
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.79	0.79	0.83	0.82	0.81	0.82	0.81	0.81	0.82	0.83	0.83	0.82
MedSE	0.71	0.78	0.85	0.91	1.02	1.16	0.28	0.13	0.40	0.36	0.54	0.46
$\rho_1 = 0.5,\ \sigma_1 = 1,\ \delta_1 = 0.05$												
Correct	4.93	5.00	2.33	2.30	2.10	2.03	4.97	5.00	3.93	4.33	3.57	3.93
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.49	0.48	0.55	0.54	0.54	0.54	0.51	0.51	0.54	0.55	0.55	0.54
MedSE	0.55	0.57	0.83	0.92	0.96	1.02	0.27	0.15	0.38	0.36	0.49	0.45
$\rho_1 = 0.2,\ \sigma_1 = 1,\ \delta_1 = 0.05$												
Correct	4.93	5.00	2.33	2.27	2.33	2.07	4.97	5.00	3.97	4.27	3.53	4.03
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.20	0.18	0.27	0.26	0.32	0.35	0.19	0.20	0.26	0.26	0.28	0.27
MedSE	0.54	0.55	0.83	0.98	0.99	1.05	0.28	0.14	0.38	0.36	0.51	0.44
$\rho_1 = 0.5,\ \sigma_1 = 2,\ \delta_1 = 0.05$												
Correct	5.00	5.00	2.00	1.73	1.63	1.77	4.30	5.00	3.67	3.80	2.87	3.20
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.49	0.48	0.56	0.56	0.54	0.56	0.53	0.53	0.56	0.56	0.56	0.56
MedSE	0.66	0.70	1.10	1.30	1.29	1.24	0.31	0.15	0.49	0.44	0.64	0.57

Particularly, in Table 10, the value of the proportion of nonretail business shrinkages further (almost to zero) for all the estimators compared with the normal case, indicating the proportion of nonretail business should be an unimportant variable. Another change of the coefficient is the location contiguous to the Charles River. In the normal case (Table 9), all the estimators output the absolute values of this coefficient less than 0.005. But the estimated values of this variable increase for many estimators when removed 50% nonzeros on each row of the weight matrix W. However, SAR via exponential squared loss and adaptive Lasso penalty (i.e., E+$\tilde{\ell}_1$) still identifies its value as zero, indicating the robustness of SAR with E+$\tilde{\ell}_1$.

7. Conclusions and discussions

In this paper, we propose a robust variable selection for the spatial autoregressive model. A penalized exponential squared loss method is applied to select significant covariates and estimate unknown parameters simultaneously. The main conclusions are as follows:

- The penalized exponential squared loss method is efficient to select useful covariates. The performance of all the tested methods are heavily affected when the outliers or intensive noise exist in the observations or a large number of the elements of the spatial weight matrix are inaccurately estimated. In these cases, however, the proposed method is particularly robust and applicable. Specially, we should note the cases of inaccurate spatial weight matrices, which frequently occur in practical spatial autoregressive applications.
- Under some regular conditions, we establish the theoretical properties of the proposed estimators, including consistency and the oracle property.

Table 8
Variable selection with regularizer and noisy weighting matrix W, E: the Exponential loss, S: the Square loss, L: the LAD loss, $\tilde{\ell}_1$: the adaptive Lasso penalty.

	$n = 30,\ q = 5$						$n = 360,\ q = 5$					
	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$	E+ℓ_1	E+$\tilde{\ell}_1$	S+ℓ_1	S+$\tilde{\ell}_1$	L+ℓ_1	L+$\tilde{\ell}_1$
Remove 30% nonzero weights												
Correct	3.97	5.00	3.00	3.63	2.53	3.00	5.00	5.00	4.97	4.93	4.60	4.77
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.44	0.44	0.40	0.43	0.44	0.44	0.43	0.43	0.43	0.43	0.43	0.43
MedSE	0.34	0.16	0.58	0.58	0.75	0.64	0.17	0.13	0.23	0.24	0.29	0.29
Remove 50% nonzero weights												
Correct	4.93	5.00	2.73	2.93	2.33	2.60	4.80	5.00	4.73	4.83	4.43	4.63
Incorrect	0.10	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.38	0.38	0.34	0.37	0.39	0.40	0.38	0.37	0.37	0.36	0.37	0.36
MedSE	0.37	0.39	0.65	0.77	0.92	0.73	0.24	0.15	0.28	0.26	0.34	0.34
Remove 80% nonzero weights												
Correct	3.43	5.00	2.33	2.40	2.00	2.20	5.00	5.00	4.43	4.47	3.73	4.07
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.30	0.29	0.24	0.26	0.29	0.28	0.25	0.25	0.25	0.25	0.26	0.25
MedSE	0.47	0.39	0.78	0.76	1.13	1.01	0.22	0.10	0.34	0.36	0.46	0.42
Add 50% nonzero weights												
Correct	4.23	5.00	1.87	1.97	1.97	1.70	4.47	5.00	4.20	4.23	3.70	3.67
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.17	0.17	0.25	0.33	0.37	0.37	0.27	0.27	0.29	0.29	0.29	0.29
MedSE	1.06	1.10	0.95	1.05	1.30	1.16	0.42	0.33	0.47	0.46	0.55	0.55
Add 100% nonzero weights on 10% rows												
Correct	4.53	5.00	3.27	3.17	3.07	2.90	5.00	5.00	4.83	4.90	4.80	4.93
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.48	0.48	0.50	0.50	0.50	0.50	0.51	0.51	0.49	0.49	0.50	0.50
MedSE	0.44	0.34	0.56	0.65	0.58	0.58	0.17	0.15	0.25	0.22	0.25	0.24
Add 200% nonzero weights on 10% rows												
Correct	4.50	5.00	3.10	3.17	2.90	3.03	5.00	5.00	4.77	4.83	4.77	4.87
Incorrect	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\rho}$	0.50	0.49	0.50	0.50	0.50	0.50	0.51	0.51	0.49	0.49	0.50	0.50
MedSE	0.43	0.33	0.59	0.60	0.59	0.57	0.17	0.16	0.23	0.22	0.24	0.24

Table 9
Variable section on Boston data set with weight matrix of relative location of the region.

	E+ℓ_1	E+$\tilde{\ell}_1$	E+ null	S+ℓ_1	S+$\tilde{\ell}_1$	S+ null	L+ℓ_1	L+$\tilde{\ell}_1$	L+ null
Crime rate	−0.0066	−0.0062	−0.0066	−0.0059	−0.0059	−0.0070	−0.0035	−0.0034	−0.0063
Residential land proportion	0.0003	0	0.0003	−0.0004	−0.0004	0.0004	−0.0008	−0.0007	0.0002
Proportion nonretail business	0.0014	0	0.0015	0.0017	0.0017	0.0013	0.0012	0.0010	0.0018
Location contiguous to the Charles River	0	0	−0.0017	0.0025	0.0025	0.0048	0.0001	0.0001	−0.0182
Nitrogen oxide concentrations	−0.1573	−0.1232	−0.1745	−0.0003	−0.0003	−0.2578	0.0000	0.0000	−0.1514
Number of rooms	0.0080	0.0077	0.0079	0.0146	0.0145	0.0067	0.0200	0.0200	0.0123
Old units rate	−0.0005	−0.0002	−0.0004	−0.0012	−0.0012	−0.0003	−0.0015	−0.0014	−0.0012
Distance to the employment center	−0.1355	−0.1201	−0.1396	−0.0107	−0.0107	−0.1571	−0.0006	−0.0006	−0.1419
Highways access index	0.0619	0.0505	0.0639	0.0103	0.0103	0.0701	0.0004	0.0004	0.0484
Tax rate	−0.0004	−0.0003	−0.0004	0.0000	0.0000	−0.0004	−0.0001	−0.0001	−0.0003
Pupil teacher ratio	−0.0104	−0.0090	−0.0106	−0.0171	−0.0166	−0.0115	−0.0034	−0.0032	−0.0081
Black proportion	0.0003	0.0003	0.0003	0.0004	0.0004	0.0003	0.0006	0.0006	0.0004
Lower status population proportion	−0.2080	−0.2202	−0.2110	−0.0249	−0.0249	−0.2279	−0.0014	−0.0014	−0.1503
$\hat{\rho}$	0.5187	0.5219	0.5174	0.6177	0.6190	0.5190	0.5149	0.5150	0.5000
$\hat{\sigma}^2$	0.0193	0.0193	0.0192	0.0247	0.0247	0.0192	0.0294	0.0294	0.0206
BIC	−566.69	−563.90	−567.69	−446.13	−446.90	−564.23	−355.95	−355.89	−540.99

Table 10
Variable section on Boston data set with the weight matrix W removed 50% nonzeros on each row.

	E+ℓ_1	E+$\tilde{\ell}_1$	E+ null	S+ℓ_1	S+$\tilde{\ell}_1$	S+ null	L+ℓ_1	L+$\tilde{\ell}_1$	L+ null
Crime rate	−0.0061	−0.0055	−0.0061	−0.0053	−0.0053	−0.0062	−0.0039	−0.0041	−0.0056
Residential land proportion	0.0002	0	0.0002	−0.0006	−0.0005	0.0002	−0.0005	−0.0006	0.0000
Proportion nonretail business	0.0000	0	0.0000	0.0001	0.0002	−0.0003	0.0002	0.0004	0.0024
Location contiguous to the Charles River	0.0235	0	0.0249	0.0040	0.0040	0.0283	0.0001	0.0001	0.0068
Nitrogen oxide concentrations	−0.1917	−0.0890	−0.1930	−0.0001	−0.0002	−0.2202	0.0000	0.0000	−0.2459
Number of rooms	0.0070	0.0068	0.0069	0.0147	0.0147	0.0065	0.0185	0.0187	0.0114
Old units rate	−0.0004	−0.0001	−0.0004	−0.0013	−0.0013	−0.0003	−0.0017	−0.0017	−0.0014
Distance to the employment center	−0.1412	−0.1068	−0.1419	−0.0103	−0.0103	−0.1473	−0.0004	−0.0004	−0.1322
Highways access index	0.0663	0.0550	0.0667	0.0112	0.0112	0.0729	0.0003	0.0003	0.0433
Tax rate	−0.0003	−0.0003	−0.0003	−0.0001	−0.0001	−0.0003	−0.0001	−0.0001	−0.0002
Pupil teacher ratio	−0.0082	−0.0052	−0.0082	−0.0144	−0.0140	−0.0083	−0.0021	−0.0022	−0.0110
Black proportion	0.0003	0.0003	0.0003	0.0005	0.0005	0.0003	0.0006	0.0006	0.0004
Lower status population proportion	−0.2299	−0.2457	−0.2308	−0.0261	−0.0261	−0.2398	−0.0012	−0.0012	−0.1684
$\hat{\rho}$	0.4658	0.4748	0.4655	0.5276	0.5286	0.4686	0.4327	0.4367	0.3980
$\hat{\sigma}^2$	0.0228	0.0231	0.0228	0.0301	0.0301	0.0227	0.0340	0.0339	0.0246
BIC	−481.50	−476.23	−481.61	−346.47	−346.56	−475.99	−289.42	−290.53	−454.30

Table 11
Variable section on Boston data set with the weight matrix W added 100% nonzeros on 10% rows.

	E+ℓ_1	E+$\tilde{\ell}_1$	E+ null	S+ℓ_1	S+$\tilde{\ell}_1$	S+ null	L+ℓ_1	L+$\tilde{\ell}_1$	L+ null
Crime rate	−0.0077	−0.0077	−0.0076	−0.0072	−0.0071	−0.0082	−0.0045	−0.0040	−0.0063
Residential land proportion	0.0004	0.0004	0.0004	−0.0004	−0.0003	0.0005	−0.0007	−0.0007	0.0002
Proportion nonretail business	0.0004	0	0.0001	0.0007	0.0008	0.0001	0.0007	0.0006	0.0015
Location contiguous to the Charles River	−0.0003	0	−0.0019	0.0026	0.0026	0.0045	0.0001	0.0001	−0.0170
Nitrogen oxide concentrations	−0.2149	−0.2021	−0.1204	−0.0001	−0.0002	−0.2528	0.0001	0.0001	−0.1441
Number of rooms	0.0077	0.0076	0.0078	0.0148	0.0147	0.0068	0.0226	0.0201	0.0117
Old units rate	−0.0004	−0.0003	−0.0004	−0.0011	−0.0011	−0.0002	−0.0017	−0.0016	−0.0010
Distance to the employment center	−0.1544	−0.1520	−0.1409	−0.0115	−0.0115	−0.1662	−0.0007	−0.0006	−0.1340
Highways access index	0.0628	0.0605	0.0603	0.0100	0.0100	0.0681	0.0003	0.0003	0.0485
Tax rate	−0.0003	−0.0003	−0.0003	0.0000	0.0000	−0.0003	0.0000	−0.0001	−0.0003
Pupil teacher ratio	−0.0140	−0.0138	−0.0126	−0.0208	−0.0204	−0.0144	−0.0037	−0.0035	−0.0097
Black proportion	0.0003	0.0003	0.0003	0.0004	0.0004	0.0003	0.0006	0.0006	0.0004
Lower status population proportion	−0.2186	−0.2200	−0.2206	−0.0256	−0.0256	−0.2337	−0.0016	−0.0014	−0.1571
$\hat{\rho}$	0.4965	0.4972	0.5000	0.5930	0.5945	0.4991	0.4712	0.5110	0.4924
$\hat{\sigma}^2$	0.0210	0.0210	0.0211	0.0274	0.0274	0.0209	0.0343	0.0321	0.0224
BIC	−519.30	−519.15	−523.33	−395.28	−395.88	−520.94	−289.70	−313.54	−500.03

- The proposed BCD algorithm has proved convergence and is efficient for minimizing the penalized exponential squared loss function. In fact, we found that the algorithm merely needs $2 \sim 4$ iterations of ρ^k before termination.
- We have compared the variable selection procedures by the Lasso and adaptive Lasso penalty, and found that the performance of adaptive Lasso outperforms Lasso penalty in almost all the tested cases.
- The proposed method could complete selection and estimation procedure effectively even if there exists no spatial autoregressive effect (i.e., $\rho = 0$). This brings convenience to users.

This work mainly focused on the SAR model by Lasso-type penalized exponential squared loss function. In fact, there are many other spatial regressive models, including parametric, nonparametric and semiparametric spatial regressive models. Moreover, there are many other robust loss functions and sparsity-inducing penalties. Nevertheless, we have developed a solid foundation and will continue to study these issues in the future.

Appendix. Proof of theorems

Proof of Theorem 1. Let $\xi_n = n^{-1/2} + a_n$ and set $\|\mathbf{u}\| = C$, where where $\mathbf{u}$ is d-dimensional vector and C is a large enough constant. Similar to Fan and Li (2001), we first show that $\|\hat{\beta} - \beta_0\| = O_p(\xi_n)$. It suffices to show that for any given $\epsilon > 0$, there is a large constant C such that, for large n,

$$P\left\{\sup_{|\mathbf{u}|=C} \ell_n\left(\boldsymbol{\theta}_0+\xi_n \mathbf{u}\right)<\ell_n\left(\boldsymbol{\theta}_0\right)\right\} \geq 1-\epsilon. \tag{A.1}$$

Denote $Z=(I-\rho W)^{-1}X^T$ and $\varepsilon^*=(I-\rho W)^{-1}\varepsilon_n$, we can rewrite model (1) as

$$Y=(I-\rho W)^{-1}X^T\beta+(I-\rho W)^{-1}\varepsilon=Z^T\beta+\varepsilon^*. \tag{A.2}$$

We know that minimizing (4) is the same thing as maximizing

$$\ell_n(\boldsymbol{\theta})=\sum_{i=1}^{n}\exp\{-(Y_i-Z_i\boldsymbol{\beta})/\gamma_n\}-n\sum_{j=1}^{p}p_{\lambda_j}(|\beta_j|). \tag{A.3}$$

Let $D_n(\boldsymbol{\theta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\exp\left\{-\left(Y_i-Z_i\boldsymbol{\beta}\right)^2/\gamma\right\}\frac{2(Y_i-Z_i\boldsymbol{\beta})}{\gamma}Z_i$. Since $p_{\lambda_j}(0)=0$ for $j=1,2,\ldots,p$, we have

$$\begin{aligned}
&\ell_n(\boldsymbol{\theta}_0+\xi_n\mathbf{u})-\ell_n(\boldsymbol{\theta}_0)\\
&=\sum_{i=1}^{n}\exp\{-(Y_i-Z_i(\boldsymbol{\beta}_0+\xi_n\mathbf{u}))/\gamma_n\}-\sum_{i=1}^{n}\exp\{-(Y_i-Z_i\boldsymbol{\beta}_0)/\gamma_n\}-\sum_{j=1}^{p}\{p_{\lambda_j}(|\beta_{j0}+\xi_{nu_j}|)-p_{\lambda_j}(|\beta_{j0}|)\}\\
&\leq\sum_{i=1}^{n}\exp\{-(Y_i-Z_i(\beta_0+\xi_n\mathbf{u}))/\gamma_n\}-\sum_{i=1}^{n}\exp\{-(Y_i-Z_i\beta_0)/\gamma_n\}-\sum_{j=1}^{s}\{p_{\lambda_j}(|\beta_{j0}+\xi_{nu_j}|)-p_{\lambda_j}(|\beta_{j0}|)\}\\
&=S_n(\mathbf{u})+K_n(\mathbf{u})
\end{aligned} \tag{A.4}$$

Note that

$$\begin{aligned}
S_n(\|\mathbf{u}\|)&=\sum_{i=1}^{n}\exp\{-\frac{(Y_i-Z_i(\beta_0+\xi_n u))^2}{\gamma_n}\}-\sum_{i=1}^{n}\exp\{-\frac{(Y_i-Z_i\beta_0)^2}{\gamma_n}\}\\
&=\xi_n\sum_{i=1}^{n}\{\exp\{-\frac{(Y_i-Z_i\beta_0)^2}{\gamma_n}\}\frac{2(Y_i-Z_i\beta)}{\gamma_n}Z_i^T\}^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T\{-\frac{2}{\gamma_n}\int ZZ^Te^{-(Y-Z\beta_0)^2/\gamma_n}\\
&\times(\frac{2(Y-Z\beta_0)^2}{\gamma_n}-1)dF(Z,y)\}\mathbf{u}n\xi_n^2\{1+o_p(1)\}\\
&=\xi_nD_n(\beta_0,\gamma_n)^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T\{-I(\beta_0,\gamma_n)\}\mathbf{u}n\xi_n^2\{1+o_p(1)\}
\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}
K_n(\mathbf{u})&=n\sum_{j=0}^{s}\{p_{\lambda_j}(|\beta_{j0}+\xi_nu_j|)-p_{\lambda_j}(|\beta_{j0}|)\}=n\xi_n\sum_{j=0}^{s}p'_{\lambda_j}(|\beta_{j0}|)sign(\beta_{j0})u_j+n\xi_n^2\sum_{j=0}^{s}p''_{\lambda_j}(|\beta_{j0}|)u_j^2\{1+o(1)\}\\
&\leq a_nn\xi_n\sum_{j=0}^{s}|u_j|+b_nn\xi_n^2\sum_{j=0}^{s}u_j^2\{1+o(1)\}\leq a_nn\xi_n\sum_{j=0}^{s}|u_j|+2b_nn\xi_n^2\|\mathbf{u}\|^2\\
&\leq\sqrt{s}a_nn\xi_n\sum_{j=0}^{s}|u_j|+b_nn\xi_n^2\|\mathbf{u}\|^2
\end{aligned} \tag{A.6}$$

Since $\gamma_n-\gamma_0=o_p(1)$, by Taylor's expansion, we have

$$\begin{aligned}
&\ell_n\left(\boldsymbol{\theta}_0+\xi_n\mathbf{u}\right)-\ell_n\left(\boldsymbol{\theta}_0\right)\\
&\leq\xi_nD_n\left(\boldsymbol{\theta}_0,\gamma_n\right)^T\mathbf{u}-\frac{1}{2}\mathbf{u}^T\left[-I\left(\boldsymbol{\theta}_0,\gamma_n\right)\right]\mathbf{u}n\xi_n^2\left\{1+o_p(1)\right\}-\sqrt{s}a_nn\xi_n\sum_{j=0}^{s}|u_j|+b_nn\xi_n^2\|\mathbf{u}\|^2.
\end{aligned} \tag{A.7}$$

Note that $n^{-1/2}D_n\left(\boldsymbol{\theta}_0,\gamma_0\right)=O_P(1)$. Therefore, the order of the first term on the right side is equal to $O_p\left(n^{1/2}\xi_n\right)=O_p\left(n\xi_n^2\right)$ in the last equation of (A.7). By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\|=C$. Since $b_n=o_p(1)$, the third term is also dominated by the second term of (A.7). Therefore, (A.1) holds by choosing a sufficiently large C.

Proof of Theorem 2 (i). We now show the sparsity. It is sufficient to show that with probability tending to one as $n\to\infty$ for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}=O_p\left(n^{-1/2}\right)$, and for some small $\epsilon_n=Cn^{-1/2}$ and $j=s+1,\ldots,p$, we have

$$\frac{\partial\ell_n(\boldsymbol{\beta})}{\partial\beta_j}=\begin{cases}>0, & \text{for } 0<\beta_j<\epsilon_n\\ <0, & \text{for } -\epsilon_n<\beta_j<0\end{cases}$$

Let

$$Q_n(\boldsymbol{\beta},\gamma)=\sum_{i=1}^{n}\exp\left\{-\left(Y_i-Z_i^T\boldsymbol{\beta}\right)^2/\gamma\right\}. \tag{A.8}$$

By Taylor's expansion, we have

$$\begin{aligned}
\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j}=&\frac{\partial Q_n\left(\boldsymbol{\beta},\gamma_n\right)}{\partial \beta_j}-np'_{\lambda_j}\left(\left|\beta_j\right|\right)\operatorname{sign}\left(\beta_j\right)\\
=&\frac{\partial Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j}+\sum_{l=1}^{p}\frac{\partial^2 Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j\partial\beta_l}\left(\beta_l-\beta_{l0}\right)\\
&+\sum_{l=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 Q_n\left(\boldsymbol{\beta}^*,\gamma_n\right)}{\partial \beta_j\partial\beta_l\partial\beta_k}\left(\beta_l-\beta_{l0}\right)\left(\beta_k-\beta_{k0}\right)\\
&-np'_{\lambda_j}\left(\left|\beta_j\right|\right)\operatorname{sign}\left(\beta_j\right)\\
&=R_{11}+R_{12}+R_{13}-np'_{\lambda_j}\left(\left|\beta_j\right|\right)\operatorname{sign}\left(\beta_j\right)
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that

$$n^{-1}\frac{\partial Q_n\left(\boldsymbol{\beta}_0,\gamma_0\right)}{\partial \beta_j}=O_p\left(n^{-1/2}\right)$$

and

$$n^{-1}\frac{\partial^2 Q_n\left(\boldsymbol{\beta}_0,\gamma_0\right)}{\partial \beta_j\partial\beta_l}=E\left\{\frac{\partial^2 Q_n\left(\boldsymbol{\beta}_0\right)}{\partial \beta_j\partial\beta_l}\right\}+o_p(1).$$

We can conclude that $R_{11}=O_p(\sqrt{n})$, $R_{12}=O_p(\sqrt{n})$, $R_{13}=O_p(\sqrt{n})$. Since $b_n=o_p(1)$ and $\sqrt{n}a_n=o_p(1)$, we obtain $\boldsymbol{\beta}-\boldsymbol{\beta}_0=\boldsymbol{O}_p\left(\boldsymbol{n}^{-1/2}\right)$, we have

$$\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j}=O_p(\sqrt{n})-np'_{\lambda_j}\left(\left|\beta_j\right|\right)\operatorname{sign}\left(\beta_j\right)=n\lambda_j\left\{-\lambda_j^{-1}p'_{\lambda_j}\left(\left|\beta_j\right|\right)\operatorname{sign}\left(\beta_j\right)+O_p\left(n^{-1/2}/\lambda_j\right)\right\}.$$

Since $1/\min_{s+1\le j\le d}\left(\sqrt{n}\lambda_j\right)=o_p(1)$ and $\lim_{n\to\infty}\inf\lim_{t\to 0+}\inf\left\{\min_{s+1\le j\le d}p_{\lambda_j}(|t|)/\lambda_j\right\}>0$ with probability 1, the sign of the derivative is completely determined by that of β_j. This completes the proof of Theorem 1(i).

Proof of Theorem 2 (ii). We now prove (ii), namely showing the asymptotic normality of $(\hat{\rho},\hat{\boldsymbol{\beta}}_1^T)^T$. For ease of presentation, let $\beta_{10}^*=\rho$ and $\beta_{1j}^*=\beta_{1j}, j=1,\dots,s$, then denote $\boldsymbol{\beta}_1^*=(\rho,\beta_{11},\dots,\beta_{1s})^T$ and $\boldsymbol{\beta}_0^*=(\rho_0,\beta_{10},\dots,\beta_{0s})^T$. We known that $\hat{\boldsymbol{\theta}}$ minimizes $Q_n(\boldsymbol{\theta})$. We have shown that there exists a $\sqrt{n}$ -consistent local maximizer of $\ell_n\left\{\left(\boldsymbol{\beta}_1,0\right)\right\}$ satisfying that

$$\frac{\partial \ell_n\left\{\left(\hat{\boldsymbol{\beta}}_1,0\right)\right\}}{\partial \beta_j}=0,\quad \text{for } j=1,\dots,s$$

Since $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator, we have

$$\begin{aligned}
&\frac{\partial Q_n\left\{\left(\hat{\boldsymbol{\beta}}_1,0\right),\gamma_n\right\}}{\partial \beta_j}-np'_{\lambda_j}\left(\left|\hat{\beta}_j\right|\right)\operatorname{sign}\left(\hat{\beta}_j\right)\\
&\qquad=\frac{\partial Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j}+\sum_{l=1}^{s}\left\{\frac{\partial^2 Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j\partial\beta_l}+o_p(1)\right\}\left(\hat{\beta}_l-\beta_{01}\right)\\
&\qquad-n\left[p'_{\lambda_j}\left(\left|\beta_{0j}\right|\right)\operatorname{sign}\left(\beta_{0j}\right)+\left\{p''_{\lambda_j}\left(\left|\beta_{0j}\right|\right)+o_p(1)\right\}\left(\hat{\beta}_j-\beta_{0j}\right)\right]=0
\end{aligned}$$

where $Q_n(\boldsymbol{\beta},\gamma)$ is defined in (A.3).

The above equation can be rewritten as follows

$$\frac{\partial Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j}=\sum_{l=1}^{s}\left\{E\left\{\frac{\partial^2 Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial \beta_j\partial\beta_l}\right\}+o_p(1)\right\}\left(\hat{\beta}_l-\beta_{01}\right)+n\Delta+n[\Sigma_1+O_p(1)](\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})$$

$$\begin{aligned}
&nI_1(\boldsymbol{\beta}_{01},\gamma_0)(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})+n\Delta+n[\Sigma_1+O_p(1)](\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})\\
&=n[I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1](\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})+n\Delta\\
&=n[I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1](\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})+n[I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1][I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1]^{-1}\Delta\\
&=n[I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1]\left\{(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})+n[I_1(\boldsymbol{\beta}_{01},\gamma_0)+\Sigma_1]^{-1}\Delta\right\}\\
&=-\frac{\partial Q_n\left(\boldsymbol{\beta}_0,\gamma_n\right)}{\partial\beta_j}+o_p(1).
\end{aligned}$$

Since $\sqrt{n}\,(\gamma_n-\gamma_0) = o_p(1)$, invoking the Slutsky's lemma and the Lindeberg–Feller central limit theorem, we have $\sqrt{n}\left(I_1\left(\boldsymbol{\beta}_{01},\gamma_0\right)+\Sigma_1\right)\left\{(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})+\left(I_1\left(\boldsymbol{\beta}_{01},\gamma_0\right)+\Sigma_1\right)^{-1}\Delta\right\}\rightarrow N\left(\mathbf{0},\Sigma_2\right)$, where $\hat{\beta}_{n1}=(\hat{\rho},\hat{\beta}_{11},\ldots,\hat{\beta}_{1s})^T$, and $\beta_{01}=(\rho_0,\beta_{01},\ldots,\beta_{0s})^T$,

$\Sigma_1=\mathrm{diag}\left\{p''_{\lambda_1}\left(|\beta_{01}|\right),\ldots,p''_{\lambda_s}\left(|\beta_{0s}|\right)\right\}$,

$\Sigma_2=\mathrm{cov}\left(\exp\left(-r^2/\gamma_0\right)\frac{2r}{\gamma_0}Z_{i1}\right)$,

$\Delta=\left(p'_{\lambda_1}\left(|\beta_{01}|\right)\mathrm{sign}\left(\beta_{01}\right),\ldots,p'_{\lambda_s}\left(|\beta_{0s}|\right)\ \times\mathrm{sign}\left(\beta_{0s}\right)\right)^T$, $I_1\left(\boldsymbol{\beta}_{01},\gamma_0\right)=\frac{2}{\gamma_0}E\left[\exp\left(-r^2/\gamma_0\right)\left(\frac{2r^2}{\gamma_0}-1\right)\right]\times\left(EZ_{i1}Z_{i1}^T\right)$. Then the proof of part (ii) is completed.

References

Anselin, L., Bera, A.K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: Ullah, A., Giles, D.E.A. (Eds.), Handbook of Applied Economic Statistics. Marcel Dekker, New York.

Banerjee, Sudipto, Carlin, Bradley P., Gelfand, Alan E., 2014. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC.

Beck, Amir, Teboulle, Marc, 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2 (1), 183–202.

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

Cressie, Noel, 1992. Statistics for spatial data. Terra Nova 4 (5), 613–617.

Fan, Jianqing, Li, Runze, 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.

Forsythe, George Elmer, Moler, Cleve B., Malcolm, Michael A., 1977. Computer Methods for Mathematical Computations.

Friedman, Jerome, Hastie, Trevor, Tibshirani, Robert, 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. 28 (2), 337–407.

Guo, Shuang, Wei, Chuan Hua, 2015. Variable selection for spatial autoregressive models. J. Minzu Univ. China.

Harrison Jr., David, Rubinfeld, Daniel L., 1978. Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manag. 5 (1), 81–102.

Huber, Peter J., Ronchetti, Elvezio M., 1981. Robust Statistics. John Wiley & Sons, New York.

Iman, Ronald L., Davenport, James M., 1980. Approximations of the critical region of the fbietkan statistic. Commun. Stat. 9 (6), 571–595.

Kelejian, Harry H., 2008. A spatial j-test for model specification against a single or a set of non-nested alternatives. Lett. Spat. Resour. Sci. 3–11.

Kelejian, Harry H., Piras, Gianfranco, 2011. An extension of kelejian's j-test for non-nested spatial models. Reg. Sci. Urban Econ. 41 (3), 281–292.

Kelejian, Harry H., Piras, Gianfranco, 2014. An extension of the j-test to a spatial panel data framework. J. Appl. Econometrics 31 (2), 387–402.

Koenker, Roger, Bassett, Gilbert, 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Krisztin, Tamás, 2017. The determinants of regional freight transport: A spatial, semiparametric approach. Geograph. Anal. 49, 268–308.

LeSage, James P., Parent, Olivier, 2007. Bayesian Model averaging for spatial econometric models. Geograph. Anal. 39 (3), 241–267.

Liang, H., Li, R., 2009. Variable selection for partially linear models with measurement errors. J. Amer. Statist. Assoc. 104 (485), 234–248.

Ma, Yingying, Pan, Rui, Zou, Tao, Wang, Hansheng, 2019. A naive least squares method for spatial autoregression with covariates. Statist. Sinica.

Pace, R. Kelley, Gilley, Otis W., 1997. Using the spatial configuration of the data to improve estimation. J. Real Estate Finance Econ. 14 (3), 333–340.

Piribauer, Philipp, 2016. Heterogeneity in spatial growth clusters. Empir. Econ. 51 (2), 659–680.

Steel, Mark F.J., 2017. Model averaging and its use in economics. arXiv preprint arXiv:1709.08221.

Tang, Qingguo, 2014. Robust estimation for functional coefficient regression models with spatial data. Statistics 48 (2), 388–404.

Tibshirani, Robert, 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Wang, Xueqin, Jiang, Yunlu, Huang, Mian, Zhang, Heping, 2013. Robust variable selection with exponential squared loss. J. Amer. Statist. Assoc. 108 (502), 632–643.

Wang, Hansheng, Li, Guodong, Jiang, Guohua, 2007. Robust regression shrinkage and consistent variable selection through the lad-lasso. J. Bus. Econom. Statist. 25 (3), 347–355.

Xuan, Liu, Chen, Jianbao, Cheng, Suli, 2018. A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. Spat. Stat. 25, 86–104.

Yuille, Alan L., Rangarajan, Anand, 2001. The concave-convex procedure (CCCP). In: Advances in Neural Information Processing Systems. NIPS, pp. 1033–1040, http://papers.nips.cc/paper/2125-the-concave-convex-procedure-cccp.

Zhang, Xinyu, Yu, Jihai, 2018. Spatial weights matrix selection and model averaging for spatial autoregressive models. J. Econometrics 1–18.

Zou, Hui, 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.

Zou, Hui, Yuan, Ming, 2008. Composite quantile regression and the oracle model selection theory. Ann. Statist. 36 (3), 1108–1126.