Statistica Sinica **31** (2021), 1239-1259
doi:https://doi.org/10.5705/ss.202018-0400

FEATURE SCREENING FOR NETWORK AUTOREGRESSION MODEL

Danyang Huang, Xuening Zhu, Runze Li and Hansheng Wang

Renmin University of China, Fudan University, Pennsylvania State University and Peking University

Abstract: Network analyses are becoming increasingly popular in a wide range disciplines, including social science, finance, and genetics. In practice, it is common to collect numerous covariates along with the response variable. Because the network structure means the responses at different nodes are no longer independent, existing screening methods may not perform well for network data. Therefore, we propose a network-based sure independence screening (NW-SIS) method that explicitly considers the network structure. The strong screening consistency property of the NW-SIS method is rigorously established. Furthermore, we estimate the network effect and establish the $\sqrt{n}$-consistency of the estimator. The finite-sample performance of the proposed method is assessed using a simulation study and an empirical analysis of a data set from the Chinese stock market.

Key words and phrases: Feature screening, network autoregression, network structure, strong screening consistency.

1. Introduction

A network data analysis is an important tool used to explore data that have a dependency structure by incorporating the network structure into the modeling framework. Network analyses have been successfully applied in a wide range of disciplines, including social science (Leenders (2002); Newman (2010)), finance (LeSage and Pace (2009); Diebold and Yilmaz (2014)), and genetics (Monnier et al. (2013); Taylor-Teeples et al. (2015)). In the field of social network analysis, network modeling is used to study users' social behavior, where researchers have found positive dependencies between users through network links (Lee, Li and Lin (2010); Chen, Chen and Xiao (2013); Zhu et al. (2017)). In the area of empirical finance, network analyses are used to study the stock returns of financial institutions. Here, studies have found that financial contagion could spread via network relationships, which is a key indicator for financial risk management (Hautsch, Schaumburg and Schienle (2014); Zou et al. (2017); Zhu et al. (2018)).

Corresponding author: Xuening Zhu, School of Data Science, Fudan University, Shanghai, China. E-mail: xueningzhu@fudan.edu.cn.

Along with the responses, researchers often collect numerous predictors. Consider, for example, a financial network of firms. One can collect firms' fundamentals from balance sheets, income statements, and the cash flow statements. However, these might contain hundreds of predictors that are closely related to the firms' financial performance (Fama and French (2015)). As another example, on a social network platform, a user's profile is collected from user-created labels. In particular, the network labels are mostly short keywords created by the user to describe his/her personal characteristics, career, life status, and so on (Huang et al. (2016)). Accordingly, the total number of keywords could be of ultrahigh dimension. However, to the best of our knowledge, the ultrahigh dimensionality of predictors has not been adequately addressed in network modeling literature.

To deal with high dimensionality, a popular solution is to consider a sparse structure of the regression coefficients. That is, we assume that not all predictors make a significant contribution to the model prediction. In this case, the predictors are screened based on their contributions to the model fitting. Since the seminal work of Fan and Lv (2008), sure independence screening (SIS) has received considerable attention in the literature. Many extensions have been investigated for the feature screening framework. These include the extensions to the generalized linear models and robust linear models developed by Fan, Samworth and Wu (2009) and Fan and Song (2010), respectively, the nonparametric SIS procedure designed by Fan, Feng and Song (2011) for additive models, and the correlation-based SIS procedure for linear models proposed by Li et al. (2012); see Wang (2009), Li, Zhong and Zhu (2012), He, Wang and Hong (2013), Mai and Zou (2013), Liu, Li and Wu (2014) and Huang, Li and Wang (2014) for further details.

Despite their usefulness in many scenarios, traditional screening methods may not be effective when a network structure is involved, because the network nodes are dependent through the network links. As a result, two questions emerge. First, how do we conduct feature screening while considering the network information? Second, how do we estimate the network effect after feature screening? In this work, we propose a network-based sure independence screening (NW-SIS) method that explicitly considers the network structure. Specifically, we design a screening measure by controlling the network effect. We prove that the NW-SIS method enjoys the strong screening consistency property and could be easy to compute. Lastly, the network effect is estimated after screening, and the $\sqrt{n}$-consistency of the estimator is established.

The rest of this paper is organized as follows. Section 2 introduces the proposed NW-SIS approach, and establishes its theoretical properties. Simulation

studies, including a real-data example, are given in Section 3. Section 4 concludes the paper. All theoretical proofs are relegated to the online Supplementary Material.

2. Network-Based Independent Screening

2.1. Model and notation

To describe the structure of a network with n nodes, we define an adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, where $a_{ij} = 1$ if there is a link from node i to node j $(j \neq i)$, and $a_{ij} = 0$ otherwise. Define $a_{ii} = 0$, for $1 \leq i \leq n$. Note that the network could be directed (i.e., A is asymmetric) or undirected (i.e., A is symmetric). Let $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the continuous responses and $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$ be the corresponding predictors, with $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ collected from the n nodes. In this study, we consider the case $p \gg n$, which means the predictors are of ultrahigh dimension.

To model the relationship between the response and the covariates, we consider the following network vector autoregression model:

$$Y = \rho WY + \mathbb{X}\beta + \mathcal{E}, \tag{2.1}$$

where $W = (w_{ij}) \in \mathbb{R}^{n\times n}$ is the normalized weighting matrix, with $w_{ij} = a_{ij}/\sum_{j=1}^{n} a_{ij}$, and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient. The coefficient ρ is the autocorrelation parameter representing the network influence effect. The model in (2.1) is similar in spirit to the spatial autoregression (SAR) model (Lee (2004); Anselin (2013)). However, it takes the network structure A into consideration rather than geographical distance information, and allows the dimension of the covariates to be ultrahigh. Lastly, $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ is assumed to have mean $\mathbf{0}_n = (0, \ldots, 0) \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I_n \in \mathbb{R}^{n\times n}$, where I_n is the identity matrix of dimension n. It is assumed that $\mathcal{E}$ and $\mathbb{X}$ are mutually independent.

Remark 1. Note that the weighting matrix W is row-normalized such that $\sum_j w_{ij} = 1$. This form is widely assumed in the literature (Chen, Chen and Xiao (2013); Liu (2014); Zhu et al. (2017); Cohen-Cole, Liu and Zenou (2018)). Therefore, the autocorrelation parameter in model (2.1) is viewed as the average network effect that nodes receive from their following friends. One could consider other flexible forms of W, such as the non-normalized adjacency matrix, or other weighting matrices. In those cases, the autocorrelation should be explained accordingly.

The row-normalized W leads to the simple assumption about the range of ρ. In order to ensure the invertibility of $(I_n - \rho W)$, ρW should have eigenvalues all different from one. Banerjee, Carlin and Gelfand (2004) have shown that the largest absolute eigenvalue of W is one. Consequently, it can be easily verified that $|\rho| < 1$ is a sufficient condition to make $(I_n - \rho W)$ invertible for a general W. As a matter of fact, this is also a necessary condition; refer to Banerjee, Carlin and Gelfand (2004) for a more detailed discussion. Thus, throughout this paper, we assume $|\rho| < 1$.

For convenience, define $\mathbb{X}_j = (X_{1j}, X_{2j}, \ldots, X_{nj})^\top \in \mathbb{R}^n$ as the jth column of $\mathbb{X}$, for $1 \le j \le p$. We follow convention, and normalize each predictor $\mathbb{X}_j$ and Y so that the means are zero and the marginal variances are one. In the high-dimension literature, sparsity is typically assumed. This means only the important features have a significant effect on the response (Fan and Lv (2008)). Therefore, we define the full model as $\mathcal{M}_F = \{1, 2, \ldots, p\}$ and let $\mathcal{M}_T = \{1 \le j \le p : \beta_j \ne 0\}$ be the true sparse model with non-sparsity size $|\mathcal{M}_T|$.

In model (2.1), the nodes are no longer independent. Instead, they are dependent via the network structure W. As a result, unimportant features might be correlated with the responses through their linkages with the important features. This makes the traditional marginal independence screening method unreliable. To see this, one can easily verify that $z_j = \mathbb{X}_j^\top Y = \mathbb{X}_j^\top (I_n - \rho W)^{-1}(\mathbb{X}\beta + \mathcal{E})(1 \le j \le p)$ depends on the network influence parameter ρ and the weighting matrix W. In this case, the correlation between $\mathbb{X}_j$ and Y can no longer be an appropriate measurement for the screening procedure. To obtain a feasible screening method in a network setting, we propose a network-based independence screening method.

2.2. Network-based independence screening

We consider a feature screening procedure when the dimension of the predictor is ultrahigh in model (2.1). If we define $Y^* = (I_n - \rho W)Y = (Y_1^*, Y_2^*, \ldots, Y_n^*)^\top \in \mathbb{R}^n$, the model can be written as

$$Y^* = \mathbb{X}\beta + \mathcal{E}.$$

A simple calculation reveals that $\mathrm{Cov}(Y^*|\mathbb{X}) = \sigma^2 I_n$. As a result, if the network effect ρ is known, then Y^* follows immediately. In this way, we can apply traditional screening approaches, such as the marginal correlation between $Y_i^*(1 \le i \le n)$ and $X_{ij}(1 \le j \le p)$. Unfortunately, in the model defined in (2.1) with ultrahigh-dimensional predictors, the estimator of ρ can be difficult to obtain.

To avoid having to estimate ρ, we evaluate the marginal correlation between

Y_i^* and $X_{ij}(1 \le j \le p)$ directly. This amounts to measuring the multiple correlation between (Y, WY) and $\mathbb{X}_j(1 \le j \le p)$. Specifically, we treat $\mathbb{X}_j$ as the response and (Y, WY) as the predictors. By regressing $\mathbb{X}_j$ on (Y, WY), we obtain an R-square-type statistic. This measurement can function as the multiple correlation between (Y, WY) and $\mathbb{X}_j$. As a result, it plays a role as an approximation to the the marginal correlation between Y_i^* and X_{ij}. Let $\widetilde{Y} = (Y, WY) \in \mathbb{R}^{n\times 2}$. Then $\widehat{\mathbf{R}}_j^2$ is defined as,

$$\widehat{\mathbf{R}}_j^2 = \frac{\mathbb{X}_j^\top \Big\{\widetilde{Y}(\widetilde{Y}^\top \widetilde{Y})^{-1}\widetilde{Y}^\top\Big\}\mathbb{X}_j}{\mathbb{X}_j^\top \mathbb{X}_j}, \tag{2.2}$$

for every $1 \le j \le p$. For a given constant c_γ, one can estimate $\mathcal{M}_T$ using

$$\widehat{\mathcal{M}}^R = \Big\{1 \le j \le p : \widehat{\mathbf{R}}_j^2 \ge c_\gamma\Big\}. \tag{2.3}$$

As a result, the full model $\mathcal{M}_F$ is reduced to a submodel $\widehat{\mathcal{M}}^R$ of size $|\widehat{\mathcal{M}}^R|$. The rank of $\widehat{\mathbf{R}}_j^2$s $(1 \le j \le p)$ learns the order of importance of the features based on their comprehensive correlation with (Y, WY). Consequently, it filters out those features with weak correlations to (Y, WY). This is the NW-SIS method. It is generalized from the SIS approach, but incorporates the network structure.

Remark 2. The problem can also be converted to one of feature screening with multiple responses, for example, the SIS procedure based on the distance correlation (DC-SIS) by Li, Zhong and Zhu (2012). This approach is model-free and can handle multiple responses. However, it is not designed for models with network structure information, and thus does not work as well as $\widehat{\mathbf{R}}_j^2$. We compare the performance of each in the numerical studies in Section 3.

2.3. Theoretical properties

In this subsection, we study the theoretical properties of the NW-SIS method. Intuitively, we wish to have $\mathcal{M}_T \subset \widehat{\mathcal{M}}^R$ with a large probability. In fact, this is satisfied if we always define $\widehat{\mathcal{M}}^R = \mathcal{M}_F = \{1, \ldots, p\}$, which is the full model. However, by doing so, a large number of irrelevant features are introduced. To achieve a desirable screening result, two properties should be satisfied. First, it should include all relevant features consistently. Second, it should simultaneously control the screening model size.

To facilitate the development of the theory, we first provide some notation related to network structures. For convenience, define $\kappa_1^{(n)} = n^{-1}\mathrm{tr}\big\{(I_n - \rho W)^{-1}(I_n - \rho W^\top)^{-1}\big\}$, $\kappa_2^{(n)} = n^{-1}\mathrm{tr}\{W(I_n - \rho W)^{-1}(I_n - \rho W^\top)^{-1}\}$, $\kappa_3^{(n)} =$

$n^{-1}\mathrm{tr}\{(I_n - \rho W^\top)^{-1}W^\top W(I_n - \rho W)^{-1}\}$, $\kappa_4^{(n)} = n^{-1}\mathrm{tr}\{(I_n - \rho W)^{-1}\}$, $\kappa_5^{(n)} = n^{-1}\mathrm{tr}\{W(I_n - \rho W)^{-1}\}$, and $\kappa_6^{(n)} = n^{-1}\mathrm{tr}[\{(I_n - \rho W)^{-1}W\}^2]$. Moreover, let $\nu_0 = \beta^\top\Sigma\beta + \sigma^2$ and $\nu_j = \beta^\top\Sigma_{\cdot j}$, where $\Sigma = \mathrm{Cov}(\mathbb{X}) \in \mathbb{R}^{n\times n}$ and $\Sigma_{\cdot j} \in \mathbb{R}^{n\times 1}$ denotes the jth column of Σ. In addition, for an arbitrary semi-positive-definite matrix M, let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the smallest and largest eigenvalues, respectively, of matrix M. Lastly, define $\mathbf{R}_j^2 = (c_\kappa^{(n)})^{-1}(\kappa_1^{(n)}\kappa_5^{(n)2} - 2\kappa_2^{(n)}\kappa_4^{(n)}\kappa_5^{(n)} + \kappa_3^{(n)}\kappa_4^{(n)2})\nu_j^2$ and $\gamma_{\min}^* = \min_{j\in\mathcal{M}_T}\mathbf{R}_j^2$, where $c_\kappa^{(n)} = (\kappa_1^{(n)}\kappa_3^{(n)} - \kappa_2^{2(n)})\nu_0$. We show in Proposition 1 that $\max_j|\widehat{\mathbf{R}}_j^2 - \mathbf{R}_j^2| = o_p(1)$, where $\widehat{\mathbf{R}}_j^2$ is defined in (2.2).

Remark 3. Note that the population screening measure $\mathbf{R}_j^2$ is proportional to ν_j^2, where $\nu_j = \beta^\top\Sigma_{\cdot j} = \sum_{i\in\mathcal{M}_T}\beta_i\Sigma_{ij}$. This might lead to the so-called "signal cancellation" problem (Wasserman and Roeder (2009)). For instance, if $\sum_{i\neq j, i\in\mathcal{M}_T}\beta_i\Sigma_{ij}/\Sigma_{jj} \approx -\beta_j$, then $\nu_j \approx 0$, regardless of the size of β_j. This corrupts the performance of the univariate screening, especially when the signals are rare and weak (Jin, Zhang and Zhang (2014)). To solve this problem, one can either impose faithfulness assumptions, or use multivariate screening procedures (Ji and Jin (2012); Jin, Zhang and Zhang (2014)). We leave this as an important future extension to this work.

Next, to establish the two abovementioned properties of the NW-SIS estimator $\widehat{\mathcal{M}}^R$, the following technical conditions are needed.

(C1) (Sub-Gaussian Distribution) The covariates X_{ij} $(1 \le i \le n)$ and the random errors ε_i $(1 \le i \le n)$ are idependent and identically distributed (i.i.d.) mean zero sub-Gaussian random variables with scale parameters $0 < \sigma_x < \infty$ and $0 < \sigma_e < \infty$; that is, for any t, $E\{\exp(tX_{ij})\} \le \exp(\sigma_x^2t^2/2)$ and $E\{\exp(t\varepsilon_i)\} \le \exp(\sigma_e^2t^2/2)$.

(C2) (Divergence Speed) Let $\log p \le \nu n^\xi$, where $0 \le \xi < 1$ and ν is a positive finite constant.

(C3) (Convergence) The limits $\kappa_1^{(n)} \to \kappa_1$, $\kappa_4^{(n)} \to \kappa_4$, and $\kappa_6^{(n)} \to \kappa_6$ exist as $n \to \infty$.

(C4) (Sparsity) Let $|\rho| < 1$, and define $\Sigma_y = \mathrm{Cov}(Y)$ and $\mathbb{W} = WW^\top$. For finite positive constants $\tau_{\min}$ and $\tau_{\max}$, $2\tau_{\min} \le \min\{\lambda_{\min}(\Sigma), \lambda_{\min}(\Sigma_y)\} \le \max\{\lambda_{\max}(\Sigma), \lambda_{\max}(\Sigma_y), \lambda_{\max}(\mathbb{W})\} \le 2^{-1}\tau_{\max}$.

(C5) (Minimum Signal) Let $\gamma_{\min}^* = 2c_\gamma$ as $n \to \infty$, where c_γ is a positive constant, as defined in (2.3).

The following comments relate to the above technical conditions. First, Condition (C1) assumes the sub-Gaussian assumption for $\mathbb{X}_j$ $(1 \leq j \leq p)$ and $\mathcal{E}$. Note that this assumption is a more relaxed condition than the normality assumption commonly employed in the feature screening literature (Fan and Lv (2008); Wang (2009); Wang, Kim and Li (2013)). One can easily verify that the response Y, which is essentially a linear combination of $\mathbb{X}$ and $\mathcal{E}$, also follows a sub-Gaussian distribution (Bartlett (2013)). Second, Condition (C2) restricts the divergence rate of p with respect to the sample size n. Specifically, the feature dimension p can be allowed to grow exponentially fast with the sample size n. Third, Condition (C3) contains a series of convergence conditions. These conditions are easily satisfied as $n \to \infty$ if the whole network admits certain uniformity properties. In addition, the following values also converge: $\kappa_2^{(n)} = \rho^{-1}(\kappa_1^{(n)} - \kappa_4^{(n)}) \to \rho^{-1}(\kappa_1 - \kappa_4) \stackrel{\text{def}}{=} \kappa_2$, $\kappa_3^{(n)} = \rho^{-2}(\kappa_1^{(n)} - 2\kappa_4^{(n)} + 1) \to \rho^{-2}(\kappa_1 - 2\kappa_4 + 1) \stackrel{\text{def}}{=} \kappa_3$, and $\kappa_5^{(n)} = \rho^{-1}(\kappa_4^{(n)} - 1) \to \rho^{-1}(\kappa_4 - 1) \stackrel{\text{def}}{=} \kappa_5$. Thus, Condition (C3) is sufficient to ensure the convergence of all $\kappa_1^{(n)}$ to $\kappa_6^{(n)}$. Subsequently, Conditions (C4) and (C1) ensure the sparse Riesz condition (SRC), which controls the eigenvalues of a fixed subset of the design matrix. See Zhang and Huang (2008), Wang (2009), and Pan, Wang and Li (2015) for definitions of the SRC and further discussions. In addition, Condition (C4) sets constraints on the network structure W, which guarantees uniformity (Zhu et al. (2017)). Lastly, Condition (C5) sets a constraint on the minimal signal of the true model $\mathcal{M}_T$. We then have the following proposition that $\mathbf{R}_j^2$ is a good approximation to $\widehat{\mathbf{R}}_j^2$.

Proposition 1. *Assume Conditions* (C1)–(C4) *hold. Then, we have that* $\max_j |\widehat{\mathbf{R}}_j^2 - \mathbf{R}_j^2| \to_p 0$.

The proof of Proposition 1 is given in Section S2 of the Supplementary Material. Condition (C5) essentially requires that the signal of $\mathbf{R}_j^2$ in the true model must stay away from zero by a good margin. Note that this is a crucial condition that guarantees that the signal of the true model will be strong enough to be detected. Thus, the screening consistency property holds. Similar conditions are widely assumed in the ultrahigh-dimensional regression literature; see Fan and Lv (2008).

Under the above technical conditions, the following screening properties can be established for the proposed NW-SIS method.

Theorem 1. *Let* $m_{\max} = c_\beta \gamma_{\min}^{*-1} \tau_{\max}^2 |\mathcal{M}_T|$, *where* c_β *is a finite positive constant. Under Conditions* (C1)–(C5), *it holds that*

$$P(\mathcal{M}_T \subset \widehat{\mathcal{M}}^R) \to 1, \tag{2.4}$$

$$P(|\widehat{\mathcal{M}}^R| \leq m_{\max}) \to 1, \tag{2.5}$$

as $n \to \infty$.

The proof of Theorem 1 is given in Section S3 of the Supplementary Material. The first conclusion in (2.4) reveals that under appropriate conditions, the NW-SIS method selects all relevant features consistently. As a result, the proposed approach enjoys the screening consistency property. Next, the model size should be controlled. As discussed earlier, if $\widehat{\mathcal{M}}^R = \mathcal{M}_F = \{1, \ldots, p\}$, the conclusion in (2.4) holds. However, the model will be overfitted in this case. In contrast, from the second conclusion in (2.5), we conclude that the overfitting effect is controlled. The conclusions in (2.4) and (2.5) are referred to as **strong screening consistency**.

Remark 4. The $m_{\max}$ in (2.5) can be treated as the upper bound for the estimated model size. From its form, we conclude that the estimated model size will be smaller if (a) the minimal signal of the true model is stronger (i.e., larger $\gamma^*_{\min}$), (b) the covariates and the responses are not highly correlated (i.e., lower $\tau_{\max}$), and (c) the true model is sparse (i.e., smaller $|\mathcal{M}_T|$).

Note that the upper bound of the model size $m_{\max}$ in Theorem 1 involves the minimal signal $\gamma^*_{\min}$. However, if the minimal signal is too small, this will result in a very high upper bound. In this case, the method may fail to select a compact model. However, if in the true model the signal of the other features is sufficiently large, the proposed screening measure is still able to detect them using a compact screened model size. See the corollary to Theorem 1, together with the detailed discussion in Section S4 of the Supplementary Material.

2.4. Parameter estimation

By Theorem 1, we know that the true model $\mathcal{M}_T$ can be consistently covered by a finite selected model using the NW-SIS procedure. Assume $\mathcal{M}$ is a model covering the true model (i.e., $\mathcal{M}_T \subset \mathcal{M}$). In this subsection, we estimate the unknown parameters of model (2.1), given $\mathcal{M}$. For convenience, we first define some notation. Let $\mathcal{M} = \{j_1, \ldots, j_s\}$ with $\mathcal{M}_T \subset \mathcal{M}$ and $|\mathcal{M}| = s$, where $j_1, \ldots, j_s \in \{1, \ldots, p\}$. Correspondingly, define $\mathbb{X}_{\mathcal{M}} = (\mathbb{X}_{j_1}, \ldots, \mathbb{X}_{j_s})^\top \in \mathbb{R}^{n \times s}$ and $\beta_{\mathcal{M}} = (\beta_{\mathcal{M},j_1}, \ldots, \beta_{\mathcal{M},j_s})^\top \in \mathbb{R}^s$. Therefore, $\beta_{\mathcal{M}}$ contains the nonzero coefficients (i.e., $\beta_{\mathcal{M}}$) and the zero coefficients.

We next give the estimation procedure. Note that the response Y in (2.1) takes the form $Y = (I - \rho W)^{-1}(\mathbb{X}\beta + \mathcal{E})$. Therefore, Y explicitly contains information on $\mathcal{E}$. Consequently, a direct least squares-type estimation (i.e., minimizing

$\|Y - \rho WY - \mathbb{X}\beta\|^2$) may introduce endogeneity and, thus, may be biased (Lee (2004)). As an alternative, we write the quasi-loglikelihood function as $\ell(\rho, \beta_{\mathcal{M}}) =$

$$\log\big|I - \rho W\big| - \frac{n}{2}\log\Big[\big\{(I-\rho W)Y - \mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}}\big\}^{\top}\big\{(I-\rho W)Y - \mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}}\big\}\Big], \quad (2.6)$$

ignoring some constants. Note that the quasi-loglikelihood (2.6) has often been studied using spatial econometrics (Lee (2004); Anselin (2013)). The corresponding asymptotic properties are established, which suit the spatial data set very well. However, some conditions might be stringent (e.g., the bounded column summation of W) when applied to the network data, especially when the network is large.

Moreover, note that in (2.6), the dimension of $\beta_{\mathcal{M}}$ diverges slowly according to the screening model size. Given $\mathcal{M}$, it is interesting to study the asymptotic behavior of the autocorrelation coefficient estimator $\widehat{\rho}$. To this end, we first maximize (2.6) with respect to $\beta_{\mathcal{M}}$, which yields,

$$\widehat{\beta}_{\mathcal{M}} = \big(\mathbb{X}_{\mathcal{M}}^{\top}\mathbb{X}_{\mathcal{M}}\big)^{-1}\big\{\mathbb{X}_{\mathcal{M}}^{\top}(I-\rho W)Y\big\}. \quad (2.7)$$

Here $\widehat{\beta}_{\mathcal{M}}$ takes an explicit form for a fixed ρ. Next, substituting (2.7) into (2.6), we have the quasi-loglikelihood as a function of ρ,

$$\ell_1(\rho) = \log\big|I - \rho W\big| - \frac{n}{2}\log\Big[Y^{\top}(I-\rho W^{\top})(I-P_X)(I-\rho W)Y\Big], \quad (2.8)$$

where $P_X = \mathbb{X}_{\mathcal{M}}(\mathbb{X}_{\mathcal{M}}^{\top}\mathbb{X}_{\mathcal{M}})^{-1}\mathbb{X}_{\mathcal{M}}^{\top}$ is the projection matrix. By maximizing $\ell_1(\rho)$, we obtain $\widehat{\rho} = \arg\max_{\rho}\ell_1(\rho)$. To study the asymptotic properties of $\widehat{\rho}$ obtained in a network, even in a large-scale network, we require the following conditions.

(C6) (Network Structure)

(C6.1) (Connectivity) Let the set of all nodes $\{1, \ldots, n\}$ be the state space of a Markov chain, with the transition probability given by W. It is assumed the Markov chain is irreducible and aperiodic. In addition, define $\pi = (\pi_i)^{\top} \in \mathbb{R}^n$ as the stationary distribution vector of the Markov chain (i.e., $\pi_i \geq 0$, $\sum_i \pi = 1$, and $W^{\top}\pi = \pi$). It is assumed that $\sum_{i=1}^{n}\pi_i^2 \to 0$ as $n \to \infty$.

(C6.2) (Uniformity) Assume $|\lambda_{\max}(W^*)| = O(\log n)$, where W^* is defined to be a symmetric matrix as $W^* = W + W^{\top}$.

Condition (C6) sets a constraint on the network structure. Similar assumptions are assumed for the recent network vector autoregression model proposed by

Zhu et al. (2018). Specifically, (C6.1) requires that a certain connectivity holds for the network structure. This essentially assumes that each node in the network is reachable from any other node. Thus, two arbitrary nodes should be connected by a finite path in the network, which fits the well-known six degrees of separation theory (Newman, Barabasi and Watts (2006)). The second condition assumes a certain type of uniformity for the network. In particular, it requires that the diverging speed of $\lambda_{\max}(W^*)$ should be sufficiently slow. Consequently, we have the following theorem.

Theorem 2. *Assume Conditions* (C1)–(C4) *and* (C6) *hold. In addition, let* $|\mathcal{M}| = o(n^{(1-\xi)/3})$. *Then, we have* $\widehat{\rho} - \rho = O_p(n^{-1/2})$.

The proof of Theorem 2 is given in Section S5 of the Supplementary Material. By Theorem 2, we conclude that under the condition that $|\mathcal{M}|$ is slowly diverging (i.e., $|\mathcal{M}| = o(n^{(1-\xi)/3})$), the estimator $\widehat{\rho}$ is $\sqrt{n}$-consistent. Subsequently, $\beta_{\mathcal{M}}$ can be estimated using (2.7). The finite performance of $\widehat{\rho}$ and $\widehat{\beta}_{\mathcal{M}}$ is illustrated using a number of simulation studies in the next section.

3. Numerical Studies

3.1. Data generation

We consider four examples. In the first three examples, the adjacency matrix A is generated from a stochastic block model with block number $K = 50$. We randomly assign each node i a block label ($k = 1, \ldots, K$) with equal probability $1/K$. Next, let $P(a_{ij} = 1) = 0.6$ if i and j are in the same block, and $P(a_{ij} = 1) = 0$ otherwise. In all the examples, the covariance matrix of $\mathcal{E}$ is set to $\sigma^2 I_n$, with $\sigma^2 = 1$; ρ is set to 0.8. We illustrate the generation of $\mathbb{X}$ in each example; the responses can be generated using model (2.1) accordingly. In each example, n is fixed as 500 and p =2000, 5000.

Example 1. (Independent Predictors). This example is adopted from Fan and Lv (2008) with $\mathcal{M}_T = \{1, 2, \ldots, d_0\}$, where $d_0 = 8$. Each predictor $\mathbb{X}_j$ is generated independently according to a standard multivariate normal distribution. Therefore, the predictors are mutually independent. Next, the jth ($1 \le j \le d_0$) nonzero coefficient of β is given by $\beta_j = (-1)^{U_j}(4 \log n/\sqrt{n} + |Z_j|)$, where U_j is a binary random variable with $P(U_j = 1) = 0.4$, and Z_j follows a standard normal distribution.

Example 2. (Autoregressive Correlation). We consider here an autoregressive-type correlation structure. In this structure, predictors with large distances

are expected to be mutually independent, approximately. Specifically, we revise the example in Wang (2009) with $\mathcal{M}_T = \{1, 4, 7\}$. Each covariate $\mathbb{X}_j$ is generated from a multivariate normal distribution with mean $\mathbf{0}_p$ and $\text{Cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$, for $(1 \leq j_1, j_2 \leq p)$. The first, fourth, and seventh components of β are given by 0.3, 0.2, and 0.2, respectively. The other components of β are fixed as zero.

Example 3. (Compound Symmetry). By compound symmetry, all predictors are equally correlated with each other. We borrow the example from Fan and Lv (2008) with $\mathcal{M}_T = \{1, 2, 3\}$. Specifically, $\mathbb{X}_j$ is generated such that $\text{var}(X_{ij}) = 1$ and $\text{Cov}(X_{ij_1}, X_{ij_2}) = 0.5$, for any $j_1 \neq j_2$ and $1 \leq j_1, j_2 \leq p$. The first three coefficients of β are fixed as 0.3. The remainder are fixed as zero.

Example 4. (A Challenging Case). In this case, a network structure is involved in the generation of the predictors. Specifically, the predictor $\mathbb{X}_j$ is generated as follows. The first d_0 covariates are sampled independently from a multivariate normal distribution $N(\mathbf{0}_n, I_n)$. Next, for $d_0 < j \leq p$, the covariate $\mathbb{X}_j$ is simulated by $\mathbb{X}_j = \mathbb{X}_1 + \rho W \mathbb{X}_1 + 1.1\mathbb{E}_j$, where $\mathbb{E}_j$ independently follows the multivariate normal distribution $N(\mathbf{0}_n, I_n)$. Then, d_0 is set to three. The first d_0 coefficients of β are fixed as 0.5, and the others are fixed as zero; that is, $\mathcal{M}_T = \{1, 2, 3\}$. In this example, the network structure is adopted as $a_{i(i+1)} = 1$, for $1 \leq i \leq n$, for computational simplicity. Note that in the last example, the dependency structure between the important and unimportant covariates increases the screening difficulty.

3.2. Results of screening consistency

We compare the proposed NW-SIS method with two popular screening methods and the oracle screening procedure:

- The SIS method (Fan and Lv 2008), which uses the sample Pearson correlation between Y and $\mathbb{X}_j$ for feature screening.
- The DC-SIS method (Li, Zhong and Zhu 2012). The distance covariance between two random vectors is defined based on characteristic functions. Thus, the distance correlation is defined for multidimensional vectors. In this way, the DC-SIS method allows for a multidimensional response. In this study, we apply the method using the distance correlation between (Y, WY) and $\mathbb{X}_j$ for feature screening.
- The oracle procedure, which uses the sample Pearson correlation coefficient between $\mathbb{X}_j$ and $(I_n - \rho W)Y$ with the true value of ρ. Because ρ is unknown,

Table 1. Screening Simulation Results for Example 1. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$. In addition, the estimated average model size (MS) is reported after tuning parameter selection. The network effect ρ and model size are assumed to be known for the oracle estimator.

p	j	Oracle	NW-SIS	SIS	DC-SIS
			$\bar{r}_j$ (CSP_j^s)		
2,000	1	4.6(97.0)	6.5(99.0)	880.3(2.0)	412.5(24.5)
	2	4.6(98.0)	5.6(99.5)	933.3(2.5)	439.4(24.5)
	3	4.7(98.0)	5.1(100.0)	932.2(3.0)	453.1(18.5)
	4	4.7(97.5)	5.7(99.5)	903.0(1.5)	410.9(20.0)
	5	5.1(96.0)	6.9(99.0)	937.9(2.0)	463.3(17.5)
	6	4.7(99.0)	5.6(99.5)	919.0(2.0)	425.6(21.0)
	7	4.5(97.0)	5.4(100.0)	948.9(2.0)	406.3(26.0)
	8	5.0(98.0)	7.3(99.5)	880.4(1.5)	465.0(22.5)
MS		8.0	11.6	2.3	17.6
			$\bar{r}_j$ (CSP_j^s)		
5,000	1	4.6(99.0)	4.9(100.0)	2359.5(3.0)	862.7(17.5)
	2	4.6(98.5)	5.0(100.0)	2055.9(2.0)	835.1(17.5)
	3	4.8(96.0)	6.6(99.5)	2150.6(3.0)	862.8(17.5)
	4	10.0(98.5)	9.9(99.5)	2185.7(2.0)	932.3(15.0)
	5	4.8(98.0)	5.1(100.0)	2046.3(2.0)	892.0(15.5)
	6	5.0(97.5)	6.0(99.5)	2244.7(2.5)	874.6(13.0)
	7	4.4(98.0)	4.5(100.0)	2270.8(2.0)	881.9(18.5)
	8	4.4(98.5)	5.1(99.5)	2222.0(1.5)	818.1(17.0)
MS		8.0	11.2	3.1	14.5

we refer to this procedure as an oracle procedure, and label it as *Oracle* in Tables 1-4.

The first method is based on the traditional feature screening procedure. The second considers the model-free DC-SIS method with multiple responses (Y, WY). The third is proposed for feature screening based on known network information, because we know that $(I_n - \rho W)Y = \mathbb{X}\beta + \mathcal{E}$. The final one is an ideal estimator because in practice, ρ is unknown.

To gauge the finite-sample performance of the proposed method, we employ the following measurements. Denote the screening model in the mth replication as $\widehat{\mathcal{M}}^{(m)} = \{1 \le j \le p : \widehat{\mathbf{R}}^2_{j,(m)} \ge c_\gamma^{(m)}\}$. The tuning parameter $c_\gamma^{(m)}$ in the mth replication is selected using the EBIC-based method (Chen and Chen (2008); Wang (2009)), which is discussed in detail in Section S6 of the Supplementary Material. We first calculate the average model size after the tuning parameter selection as $\text{MS} = M^{-1}\sum_m \text{MS}^{(m)}$, where $\text{MS}^{(m)} = |\widehat{\mathcal{M}}^{(m)}|$ in the mth replication.

Table 2. Screening Simulation Results for Example 2. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$. In addition, the estimated average model size (MS) is reported after tuning parameter selection. The network effect ρ and model size are assumed to be known for the oracle estimator.

p	j	Oracle	NW-SIS	SIS	DC-SIS
			$\bar{r}_j(\mathrm{CSP}_j^s)$		
2,000	1	1.8(97.5)	1.8(99.5)	3.2(97.5)	4.5(93.0)
	4	3.3(73.0)	4.6(85.0)	6.9(80.0)	11.7(71.0)
	7	1.8(95.5)	1.8(99.5)	2.0(99.0)	2.9(95.5)
MS		3.0	3.6	4.0	3.8
			$\bar{r}_j(\mathrm{CSP}_j^s)$		
5,000	1	1.8(97.0)	1.8(99.5)	2.1(99.5)	3.3(96.0)
	4	3.4(73.0)	4.6(87.5)	10.2(80.5)	22.9(69.0)
	7	1.7(97.5)	1.7(100.0)	2.2(99.5)	3.8(95.5)
MS		3.0	4.4	4.7	4.2

Table 3. Screening Simulation Results for Example 3. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$. In addition, the estimated average model size (MS) is reported after tuning parameter selection. The network effect ρ and model size are assumed to be known for the oracle estimator.

p	j	Oracle	NW-SIS	SIS	DC-SIS
			$\bar{r}_j(\mathrm{CSP}_j^s)$		
2,000	1	2.3(98.5)	2.2(99.0)	5.5(88.0)	8.4(84.5)
	2	2.0(98.0)	2.0(100.0)	5.1(92.0)	6.1(87.5)
	3	2.0(98.5)	2.1(99.5)	4.5(94.0)	6.3(91.0)
MS		3.0	3.6	4.0	3.8
			$\bar{r}_j(\mathrm{CSP}_j^s)$		
5,000	1	2.2(95.5)	2.3(99.0)	13.8(88.0)	22.8(85.0)
	2	2.4(96.0)	2.3(99.0)	18.9(81.0)	33.7(73.0)
	3	2.2(96.0)	2.2(99.0)	10.6(84.0)	13.5(79.0)
MS		3.0	4.4	4.7	4.2

A smaller MS implies a more compact screening model. Next, we evaluate the screening performance for each predictor j. First, we record the rank of the jth $(1 \le j \le p)$ predictor as $r_j^{(m)}$ for the mth $(1 \le m \le M)$ replication of the simulation. For each j, the average rank $\bar{r}_j = M^{-1}\sum_{m=1}^{M} r_j^{(m)}$ is calculated. Next, the correctly selected probability (i.e., $\mathrm{CSP}_j^s = M^{-1}\sum_m I(j \in \widehat{\mathcal{M}}^{(m)})$) is reported to reflect the model recoverability. We repeat the experiment $M = 200$ times to evaluate a reliable result.

Detailed results for the simulations are given in Tables 1-4. The oracle pro-

Table 4. Screening Simulation Results for Example 4. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$. In addition, the estimated average model size (MS) is reported after tuning parameter selection. The network effect ρ and model size are assumed to be known for the oracle estimator.

p	j	Oracle	NW-SIS	SIS	DC-SIS
			$\bar{r}_j(\text{CSP}_j^s)$		
2,000	1	2.0(99.0)	2.7(96.0)	835.4(0.0)	841.5(0.5)
	2	10.8(85.5)	44.4(83.5)	1009.9(15.0)	966.8(10.5)
	3	10.8(85.0)	49.0(84.0)	965.0(17.0)	943.8(12.0)
MS		3.0	3.6	4.0	3.8
			$\bar{r}_j(\text{CSP}_j^s)$		
5,000	1	2.1(97.5)	4.5(95.0)	2141.0(0.0)	1921.7(1.0)
	2	21.6(86.0)	83.8(83.5)	2383.6(12.5)	2231.3(10.5)
	3	8.0(85.5)	51.6(83.5)	2187.6(14.5)	2084.4(12.0)
MS		3.0	4.4	4.7	4.2

cedure has the smallest $\bar{r}_j$ in all of the examples, as expected, mainly because we know the network effect ρ and model size in advance. Note that the proposed NW-SIS method outperforms the SIS and DC-SIS methods in terms of both $\bar{r}_j$ and CSP_j^s, which are almost as good as the oracle procedure. In addition, the NW-SIS method achieves a more compact model size (with lower MS) than those of the other two methods after the selection of the tuning parameter. In the final example, as expected, $\mathbb{X}_1$ is easier to recover than $\mathbb{X}_2$ and $\mathbb{X}_3$ for both the oracle procedure and the proposed NW-SIS. The reason can be explained as follows. Define Corr_j as the Pearson correlation coefficient between $\mathbb{X}_j$ and $(I_n - \rho W)Y$. By the design of Example 4, we can calculate explicitly that $|\text{Corr}_j/\text{Corr}_1| = |\rho|/(2.21 + \rho^2) < 1$, for $j > 3$ (because $(I_n - \rho W)^{-1}$ can be expressed explicitly in this case). Thus, the first feature is relatively easy to identify. However, owing to the correlation between $\mathbb{X}_j$ $(j > 3)$ and Y, recovering $\mathbb{X}_2$ and $\mathbb{X}_3$ is more difficult. The results show that NW-SIS method outperforms the SIS and DC-SIS methods in this case.

3.3. Results of parameter estimation

In this subsection, we examine the parameter estimation result. Specifically, s is set as 10 and $M = 200$. Let $\widehat{\mathcal{M}}^{(m)}$ denote the selected model in the mth $(1 \le m \le M)$ replication. Define the coverage probability (CP) and the root sum of squares error (RSSE) for ρ, σ^2, and β for the mth $(1 \le m \le M)$ replication as follows:

Table 5. Parameter Estimation Simulation Results with 200 Replications for Examples 2–4. The coverage probability CP(%) and root sum of squares error for ρ (RSSE_ρ), σ^2 (RSSE_{σ^2}), and β (RSSE_β) are reported.

p	n	CP(%)	RSSE_ρ	RSSE_{σ^2}	RSSE_β
Example 2					
5,000	200	27.50	0.0195	0.1717	0.4770
	500	97.00	0.0162	0.0689	0.2452
	1,000	100.00	0.0138	0.0369	0.1550
Example 3					
5,000	200	34.50	0.0174	0.1455	0.5652
	500	99.00	0.0139	0.0669	0.3038
	1,000	100.00	0.0127	0.0392	0.2086
Example 4					
5,000	200	18.00	0.0509	0.1396	0.6242
	500	84.00	0.0246	0.0717	0.2684
	1,000	99.00	0.0163	0.0394	0.1537

$$\begin{aligned}
\text{CP}^{(m)} &= I(\widehat{\mathcal{M}}^{(m)} \supset \mathcal{M}_T),\\
\text{RSSE}_\rho^{(m)} &= |\widehat{\rho}_{(\widehat{\mathcal{M}}^{(m)})} - \rho|^2,\\
\text{RSSE}_{\sigma^2}^{(m)} &= |\widehat{\sigma^2}_{(\widehat{\mathcal{M}}^{(m)})} - \sigma^2|^2,\\
\text{RSSE}_\beta^{(m)} &= ||\widehat{\beta}_{(\widehat{\mathcal{M}}^{(m)})} - \beta||,
\end{aligned}$$

where $I(\cdot)$ is the indicator function. We then average the performance measures across all replications. This leads to $\text{CP} = M^{-1}\sum_{m=1}^{M}\text{CP}^{(m)}$, $\text{RSSE}_\rho = M^{-1}\sum_{m=1}^{M}\text{RSSE}_\rho^{(m)}$, $\text{RSSE}_{\sigma^2} = M^{-1}\sum_{m=1}^{M}\text{RSSE}_{\sigma^2}^{(m)}$, and $\text{RSSE}_\beta = M^{-1}\sum_{m=1}^{M}\text{RSSE}_\beta^{(m)}$. We fix $p = 5{,}000$, and $n = \{200,\ 500,\ 1000\}$. In Example 1, β in each replication is not fixed. Therefore, to examine the reliability, we consider only Examples 2–4 in the simulation for the parameter estimation.

The simulation results are given in Table 5. We conclude the following. First, the CP values for all examples quickly increase toward 100% as the sample size n increases. This corroborates the strong screening consistency property, which we defined in (2.4) and (2.5). Second, RSSE_ρ decreases as n increases, as explained by Theorem 2. Lastly, RSSE_σ^2 and RSSE_β steadily decrease as n increases in all of the examples.

3.4. Financial feature screening for stock returns

We next illustrate a real-data example using data collected from the Chinese Stock Market in 2014. The data set consists of $n = 487$ stocks in the Chinese A

share market, which are traded in the Shanghai Stock Exchange and the Shenzhen Stock Exchange. The corresponding response Y_i is the annualized return of stock i $(1 \leq i \leq n)$ in 2014.

To construct the network relationship between the stocks, the common shareholders of the stocks are considered. First, we collect information on the top 10 shareholders for each stock, which we define as *major shareholders*. Second, for $i \neq j$, if the ith stock and jth stock share at least one major shareholder, then define $a_{ij} = a_{ji} = 1$; otherwise, $a_{ij} = a_{ji} = 0$. The resulting network density (i.e., $\sum_{j \neq i} a_{ij}/\{n(n-1)\}$) is 9.34%. In addition to the response (i.e., Y_i) and the network information (i.e., A), the firm-specific financial indices in the previous year (i.e., 2013) are considered as explanatory covariates. The financial indices are collected from the firms' financial statements (i.e., the balance sheet, income statement, and cash flow statement released in 2013). Furthermore, we consider the interaction effects between $\mathbb{X}_{j_1}$ and $\mathbb{X}_{j_2}$ within the same financial statement, which we define as $\mathbb{X}_{j_1}\mathbb{X}_{j_2}$. This yields a total of $p = 796$ predictors.

We then conduct the NW-SIS analysis. Here, $\widehat{\mathbf{R}}_j^2$ is calculated for $j = 1, \ldots, p$. Next, the covariates are ranked according to the decreasing order of the $\widehat{\mathbf{R}}_j^2$ values. The covariates with the top eight highest $\widehat{\mathbf{R}}_j^2$ are given in Figure 1. These are mostly related to the assets (i.e., Asset Impairment Loss, Capital Reserve Fund, Deferred Tax Asset, Intangible Assets), liabilities (i.e., Short Term Loan, Total Liability), liquidity (i.e., Cash Equivalents), and Financial Expense of the firm.

Next, we compare the NW-SIS method with the SIS and DS-SIS methods using model fitness levels. First, we conduct the screening procedure for all three approaches. We then compare the fitness levels of the methods while varying the model size $|\mathcal{M}| = 1, \ldots, 200$. The estimation is conducted as follows. For the SIS method, we follow Fan and Lv (2008) to estimate a linear regression model, and then obtain the resulting estimator $\widehat{\beta}_{\mathcal{M}}$. Therefore, the fitted value $\widehat{Y}$ can be calculated as $\widehat{Y} = \mathbb{X}_{\mathcal{M}}\widehat{\beta}_{\mathcal{M}}$. Next, for the other two methods, we use the estimation methods in Section 2.4 to obtain $\widehat{\rho}_{\mathcal{M}}$ and $\widehat{\beta}_{\mathcal{M}}$, because the multivariate information is considered in the screening procedure.

To eliminate the endogenous effect, the fitted value is computed as $\widehat{Y} = (I - \widehat{\rho}_{\mathcal{M}}W)^{-1}\mathbb{X}_{\mathcal{M}}\widehat{\beta}_{\mathcal{M}}$. Lastly, we compare the fitness of the three screening approaches using the adjusted R^2, as shown in Figure 2. The figure shows that, as more features are included, the adjusted R^2 increases at first for all three methods. Next, the adjusted R^2 of the NW-SIS method achieves peaks at $|\mathcal{M}| = 75$, which is 25.8%, and the highest of the three methods. Consequently, compared with

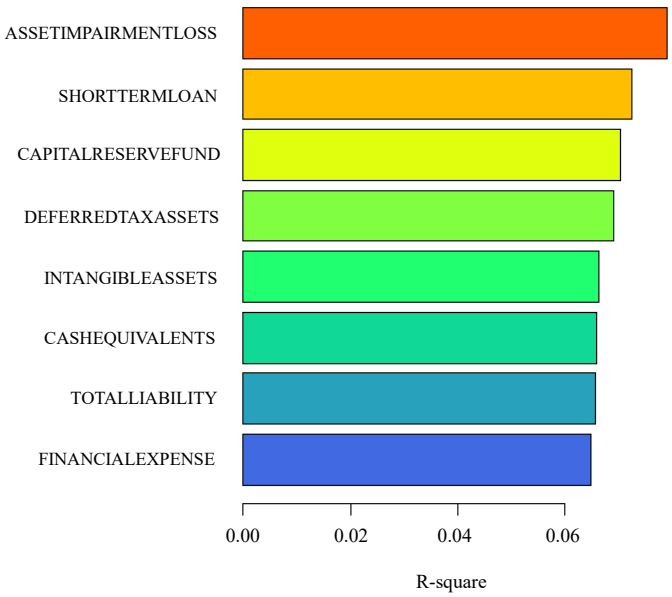

Figure 1. Covariates with top eight $\widehat{\mathbf{R}}_j^2$ related to the assets (i.e., Asset Impairment Loss, Capital Reserve Fund, Deferred Tax Asset, Intangible Assets), liabilities (i.e., Short Term Loan, Total Liability), liquidity (i.e., Cash Equivalents), and Financial Expense of the firm.

the other competing methods, the NW-SIS method obtains a better fitness level using fewer features.

4. Conclusion

We have proposed a network-based independence screening approach that incorporates the network structure. We rigorously show that the proposed NW-SIS method enjoys the strong screening consistency property. The properties of the parameter estimation are established next. Lastly, the proposed method is applied to a financial data set that screens financial indices effectively with respect to stock returns.

To conclude, we discuss several topics for future research. First, the responses considered in this study are continuous. In practice, other types of responses (i.e., discrete, mixed type) are frequently encountered. Accordingly, corresponding screening methods should be developed and studied. Second, the innovation term $\mathcal{E}$ in model (2.1) has been restricted to be independent across network nodes. This can be made more flexible to allow for more sophisticated structures (e.g., autoregressive structures). This may improve the estimation efficiency. Third, in

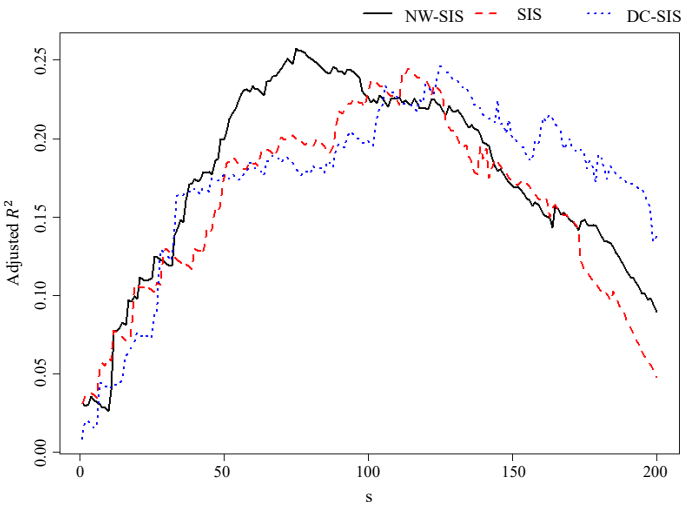

Figure 2. The fitted adjusted R^2 against the screening model size s for the three screening methods: NW-SIS, SIS, and DC-SIS. The adjusted R^2 of NW-SIS achieves the peak value first at $s = 75$, which is 25.8%, and the highest of the three methods.

the numerical study, we show that the tuning parameter selection method performs well. However, the theoretical properties of the tuning parameter selection should be investigated further. Lastly, note that unimportant features are typically included in the post-screening set because the screening technique tends to overselect the features. Consequently, appropriate variable selection methods are worth investigating after the screening procedure to precisely identify the true model.

Supplementary Material

The online Supplementary Material contains useful lemmas, the proof of Proposition 1, the proofs of Theorems 1–2, a corollary to Theorem 1, and a discussion on selecting the tuning parameter.

Acknowledgments

Danyang Huang is supported by National Natural Science Foundation of China (NSFC, 12071477, 11701560), fund for building world-class universities (disciplines) of Renmin University of China. Xuening Zhu (xueningzhu@fudan.edu.cn) was supported by the National Natural Science Foundation of China (nos. 11901105, 71991472, U1811461), the Shanghai Sailing Program for Youth Science and Technology Excellence (19YF1402700), and the Fudan-Xinzailing Joint

Research Centre for Big Data, School of Data Science, Fudan University. Runze Li was supported by the National Institute on Drug Abuse (NIDA) grant P50 DA039838, and the National Science Foundation grants DMS 1820702 and DMS 1953196. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NIDA, or NIH; Hansheng Wang's research was partially supported by the National Natural Science Foundation of China (No. 11831008, 11525101, 71532001), and partially by China's National Key Research Special Program (No. 2016YFC0207704). The corresponding author is Xuening Zhu.

References

Anselin, L. (2013). *Spatial Econometrics: Methods and Models*. Springer Science & Business Media.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Champman & Hall/CRC.

Bartlett, P. (2013). *Theoretical Statistics*. Lecture notes. Lecture 3.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model spaces. *Biometrika* **95**, 759–771.

Chen, X., Chen, Y. and Xiao, P. (2013). The impact of sampling and network topology on the estimation of social intercorrelation. *Journal of Marketing Research* **50**, 95–110.

Cohen-Cole, E., Liu, X. and Zenou, Y. (2018). Multivariate choices and identification of social interactions. *Journal of Applied Econometrics* **33**, 165–178.

Diebold, F. X. and Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics* **182**, 119–134.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics* **116**, 1–22.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **116**, 544–557.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **70**, 849–911.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* **10**, 1829–1853.

Fan, J. and Song, R. (2010). Sure independent screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604.

Hautsch, N., Schaumburg, J. and Schienle, M. (2014). Financial network systemic risk contributions. *Review of Finance* **19**, 685–738.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41**, 342–369.

Huang, D., Li, R. and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics* **32**, 237–244.

Huang, D., Yin, J., Shi, T. and Wang, H. (2016). A statistical model for social network labeling.

Journal of Business & Economic Statistics **34**, 368–374.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics* **40**, 73–103.

Jin, J., Zhang, C.H. and Zhang, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *The Journal of Machine Learning Research* **15**, 2723–2772.

Lee, L. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**, 1899–1925.

Lee, L., Li, J. and Lin, X. (2010). Specification and estimation of social interaction models with network structure. *The Econometrics Journal* **13**, 145–176.

Leenders, R. T. A. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks* **24**, 21–47.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman & Hall, New York.

Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846–1877.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association* **107**, 1129–1139.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of American Statistical Association* **109**, 266–274.

Liu, X. (2014). Identification and efficient estimation of simultaneous equations network models. *Journal of Business & Economic Statistics* **32**, 516–536.

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234.

Monnier, P., Martinet, C., Pontis, J., Stancheva, I. Ait-Si-Ali, S. and Dandolo, L. (2013). H19 lncRNA controls gene expression of the imprinted gene network by recruiting MBD1. *Proceedings of the National Academy of Sciences* **110**, 20693–20698.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M., Barabasi, A.-L. and Watts, D. J. (2006). *The Structure and Dynamics of Networks*. Princeton University Press.

Pan, R., Wang, H. and Li, R. (2015). Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association* **111**, 169–179.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T. W., Gaudinier, A. et al. (2015). An arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature* **517**, 571-575.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

Wang, L., Kim, Y. and Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics* **41**, 2505–2536.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37**, 2178-2201.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics* **36**, 1567–1594.

Zhu, X., Pan, R., Li, G., Liu, Y. and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics* **45**, 1096–1123.

Zhu, X., Wang, W., Wang, H. and Härdle, W. K. (2018). Network quantile autoregression.

Journal of Econometrics **212**, 345-358.

Zou, T., Lan, W., Wang, H. and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association* **112**, 266–281.

Danyang Huang

Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China.

E-mail: dyhuang89@126.com

Xuening Zhu

School of Data Science, Fudan University, Shanghai, China.

E-mail: xueningzhu@fudan.edu.cn

Runze Li

Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111, USA.

E-mail: rzli@psu.edu

Hansheng Wang

Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, China.

E-mail: hansheng@gsm.pku.edu.cn

(Received September 2018; accepted October 2019)