Statistica Sinica Preprint No: SS-2018-0400	
Title	Feature Screening for Network Autoregression Model
Manuscript ID	SS-2018-0400
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0400
Complete List of Authors	Danyang Huang Xuening Zhu Runze Li and Hansheng Wang
Corresponding Author	Xuening Zhu
E-mail	xueningzhu@fudan.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Feature Screening for Network Autoregression Model

Danyang Huang[1], Xuening Zhu[2], Runze Li[3], and Hansheng Wang[4]

[1]*Renmin University of China,* [2]*Fudan University,*

[3]*Pennsylvania State University,* [4]*Peking University*

Abstract: Network analysis has drawn great attention in recent years. It is applied to a wide range disciplines. These include but are not limited to social science, finance and genetics. It is typical that one collects abundant covariates along the response variable in practice. Since the network structure makes the responses at different nodes no longer independent, existing screening methods may not perform well for network data. We propose a network-based sure independence screening (NW-SIS) method. This approach explicitly takes the network structure into consideration. The strong screening consistency property of the NW-SIS is rigorously established. We further investigated the estimation of the network effect and establish the $\sqrt{n}$-consistency of the estimator. The finite sample performance of the proposed method is assessed by simulation study and illustrated by an empirical analysis of a dataset from Chinese stock market.

Key words and phrases: Feature Screening, Network Structure, Strong Screening Consistency, Network Autoregression.

1. Introduction

Network data analysis is an important tool to explore data with dependency structure. Particularly, it takes the valuable network structure information into the modelling framework. Network analysis has been successfully applied to a wide range of disciplines, which include but are not limited to 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, the network modelling is used to study user social behaviours. Researchers found positive dependence among users through network links (Lee et al., 2010; Chen and Xiao, 2013; Zhu et al., 2017). In the area of empirical finance, network analysis can be carried on to study the stock returns of financial institutions. It is found the financial contagion could spread through network relationships, which is a key indicator for the financial risk management (Hautsch et al., 2014; Zou et al., 2017; Zhu et al., 2018).

Meanwhile, along with the responses, one could collect abundant predictors. Consider for example the financial network of firms. One could be able to collect firms' fundamentals through balance sheet, income statement, and the cash flow statement. These sheets might contain hundreds of predictors. They could be closely related to the firms' financial perfor-

mance (Fama and French, 2015). As another example, in social network platform, a user's profile is collected with user-created labels. Particularly, the network labels are mostly short keywords created by the user themselves to describe their personal characteristics, careers, life status, and so on (Huang et al., 2016). Accordingly, the total amount of keywords could be of ultrahigh dimension. However, to our best knowledge, the ultrahigh dimensionality of the predictors has not been carefully considered and dealt with by the existing network modelling literature.

To deal with the high dimensionality, a popular solution is to consider the sparse structure of regression coefficients. That is to assume that not all the predictors would make a significant contribution to the model prediction. Consequently, the predictors should be screened according to 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 recent literature. Many extensions have been investigated for the feature screening framework. To name a few, the extensions to generalized linear models and robust linear models have been developed by Fan et al. (2009) and Fan and Song (2010) separately. A nonparametric SIS procedure is designed by Fan et al. (2011) for additive models. A correlation based SIS procedure for linear models is proposed by Li et al. (2012a). See Wang

(2009); Li et al. (2012b); He et al. (2013); Mai and Zou (2013); Liu et al. (2014); Huang et al. (2014) for more discussions.

Despite the great usefulness to many application scenarios, traditional screening methods may not be effective when the network structure is involved. That is because, the network nodes are no longer independent but dependent through network links. As a result, two questions emerge. First, how to conduct feature screening by considering the network information? Second, how to estimate the network effect after feature screening? In this work, we propose a network-based sure independence screening method, which explicitly takes the network structure into consideration. Specifically, a screening measure is designed by controlling the network effect. It is proved that the NW-SIS enjoys the strong screening consistency 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 the article is organized as follows. Section 2 introduces the proposed network-based independence screening approach, and establishes its theoretical properties. Simulation studies including a real data example are given in Section 3. Section 4 concludes the article with discussions. All the theoretical proofs are relegated to a separate supplementary material.

2. Network-Based Independent Screening

2.1 Model and Notations

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 $1 \leq i \leq n$. It is remarkable that the network could be directed (i.e., A is asymmetric) or undirected (i.e., A is symmetric). Let $Y = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ be the continuous responses and $\mathbb{X} = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n\times p}$ be the corresponding predictors with $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ collected from the n nodes. In this article, we consider the case that $p \gg n$, which means the predictors are of ultrahigh dimension.

To model the relationship between the response and 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 $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient. The coefficient ρ is the autocorrelation parameter representing the network influence effect. The model form (2.1) is in spirit similar to the spatial

autoregression (SAR) model (Lee, 2004; Anselin, 2013). Instead, it takes network structure A into consideration rather than the geographical distance information, and allows the dimension of covariates to be ultrahigh at the same time. Lastly, $\mathcal{E} = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$ is assumed to have mean $\mathbf{0}_n = (0, \cdots, 0) \in \mathbb{R}^n$ 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 $\mathcal{E}$ and $\mathbb{X}$ are mutually independent with each other.

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

In fact, 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 1. Banerjee and Gelfand (2004) have shown the largest absolute eigenvalue of W is 1. 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; we refer to Banerjee and Gelfand (2004) for more detailed discussions. As a consequence, throughout the rest of this article, we assume $|\rho| < 1$.

For convenience, define $\mathbb{X}_j = (X_{1j}, X_{2j}, \cdots, X_{nj})^\top \in \mathbb{R}^n$ to be the jth column of $\mathbb{X}$ for $1 \leq j \leq p$. We follow the convention to normalize each predictor $\mathbb{X}_j$ and Y so that the means are 0 and the marginal variances are 1. In the high dimension literature, sparsity is typically assumed. This means not all the features will make a significant influence on the response, but only the important ones do (Fan and Lv, 2008). Therefore, we define the full model to be $\mathcal{M}_F = \{1, 2, \cdots, p\}$ and $\mathcal{M}_T = \{1 \leq j \leq p : \beta_j \neq 0\}$ to be the true sparse model with non-sparsity size $|\mathcal{M}_T|$.

By model (2.1), it is noteworthy that the nodes are no longer independent. Instead, they are dependent via the network structure W. As a result, the non-important features might be correlated with the responses through the linkage with the important ones. This makes the traditional marginal independence screening method unreliable. To see this, one could easily verify that $z_j = \mathbb{X}_j^\top Y = \mathbb{X}_j^\top (I_n - \rho W)^{-1}(\mathbb{X}\beta + \mathcal{E})(1 \leq j \leq p)$ depends on 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 the network setting, we discuss a network-based independence screening method in the next section.

2.2 Network-based Independence Screening

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

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

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

To avoid the estimation of ρ, we consider the evaluation of the marginal correlation between Y_i^* and $X_{ij}(1 \le j \le p)$ directly. This amounts to measure the multiple correlation between (Y, WY) and $\mathbb{X}_j(1 \le j \le p)$. Specifically, we treat $\mathbb{X}_j$ as response and (Y, WY) as predictors. By regressing $\mathbb{X}_j$ on (Y, WY), an R-square type statistic could be obtained. This measure-

ment could work as the multiple correlation between (Y, WY) and $\mathbb{X}_j$. As a result, it plays a role as an approximate 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,

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

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

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

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

Remark 2. The problem could also be converted to feature screening with multiple responses, for example, the sure independence screening procedure based on the distance correlation (DC-SIS) provided by Li et al. (2012b). It is model-free and can handle the problem with multiple responses. How-

ever, it is not designed for the particular model with network structure information, thus could not work as well as the $\widehat{\mathbf{R}}_j^2$. We will compare their performances in 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 could be satisfied if we always define $\widehat{\mathcal{M}}^R = \mathcal{M}_F = \{1, ..., p\}$, 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 the relevant features consistently; second, it should control the screening model size simultaneously.

To facilitate the development of the theory, we first give a few notations with respect to network structures as follows. 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 of matrix M respectively. Lastly, we 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 $c_\kappa^{(n)} = (\kappa_1^{(n)}\kappa_3^{(n)} - \kappa_2^{2(n)})\nu_0$. It will be shown 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 $\nu_j \approx 0$ no matter how large β_j is. This corrupts the performance of the univariate screening especially when the signals are rare and weak (Jin et al., 2014). To solve this, one could either impose faithfulness assumptions or take multivariate screening procedures (Ji and Jin, 2012; Jin et al., 2014). We leave this as an important future research direction as an 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 random errors ε_i $(1 \le i \le n)$ are i.i.d mean zero sub-Gaussian random variable with scale parameters $0 < \sigma_x < \infty$, $0 < \sigma_e < \infty$, i.e., 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 following limits exist as $\kappa_1^{(n)} \to \kappa_1$, $\kappa_4^{(n)} \to \kappa_4$ and $\kappa_6^{(n)} \to \kappa_6$ 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 $\tau_{\max}$, $2\tau_{\min} \leq \min\{\lambda_{\min}(\Sigma), \lambda_{\min}(\Sigma_y)\} \leq \max\{\lambda_{\max}(\Sigma), \lambda_{\max}(\Sigma_y), \lambda_{\max}(\mathbb{W})\} \leq 2^{-1}\tau_{\max}$.

(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 are made for better understanding to the above technical conditions. First, Condition (C1) assumes the sub-Gaussian assumption for $\mathbb{X}_j$s $(1 \leq j \leq p)$ and $\mathcal{E}$. It is remarkable 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 et al., 2013). One can easily verify that the response Y, which essentially is a linear combination of $\mathbb{X}$ and $\mathcal{E}$, also follows 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 could be allowed to grow exponentially fast with the sample size n. Third, Condition (C3) contains a series of convergence conditions. These conditions

can be easily satisfied as $n \to \infty$ if the whole network admits certain uniformity property. In addition, it holds that 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, Condition (C4) together with Condition (C1) assures the Sparse Riesz Condition (SRC), which controls the eigenvalues of a fixed subset of the design matrix. See Zhang and Huang (2008), Wang (2009), Pan et al. (2015) for definition of the SRC and more discussions. In addition, Condition (C4) sets constrains on the network structure W, which guarantees certain form of uniformity (Zhu et al., 2017). Lastly, condition (C5) sets the constraint of 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), then it holds* $\max_j |\widehat{\mathbf{R}}_j^2 - \mathbf{R}_j^2| \to_p 0$.

The proof of Proposition 1 is given in Section S2 in the supplementary material. The condition (C5) essentially requires the signal of $\mathbf{R}_j^2$s in the true model must stay away from 0 with a good margin. Note that this is a crucial condition which guarantees the signal of the true model 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) for more discussion.

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 in the supplementary material. The first conclusion in (2.4) reveals that under appropriate conditions, the NW-SIS could select all the relevant features consistently. As a consequence, the proposed approach enjoys the screening consistency property. Next, the model size should be controlled. As we have discussed before, if $\widehat{\mathcal{M}}^R = \mathcal{M}_F = \{1, ..., p\}$, the conclusion in (2.4) holds. However, the model will be overfitted in this case. By contrast, by the second conclusion in (2.5), we could conclude that the overfitting effect is controlled. The conclusions in both (2.4) and (2.5) are referred to as the **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, one could conclude that the estimated model size will be smaller if (a) the minimal signal of true model is stronger (i.e., larger $\gamma^*_{\min}$); (b) the covariates as well as the responses are not highly correlated (i.e., lower $\tau_{\max}$); (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 that situation, the method may fail to select a compact model. However, if in the true model, the signal of the other features is large enough, the proposed screening measure is still able to detect them with a compact screened model size. See a corollary from Theorem 1, together with detailed discussion in Section S4 in 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 through 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 investigate the estimation of unknown parameters of model (2.1) given $\mathcal{M}$. For convenience, we first define some notations. Let $\mathcal{M} = \{j_1, \cdots, j_s\}$

with $\mathcal{M}_T \subset \mathcal{M}$ and $|\mathcal{M}| = s$, where $j_1, \cdots, j_s \in \{1, ..., p\}$. Correspondingly, define $\mathbb{X}_{\mathcal{M}} = (\mathbb{X}_{j_1}, \cdots, \mathbb{X}_{j_s})^\top \in \mathbb{R}^{n\times s}$ and $\beta_{\mathcal{M}} = (\beta_{\mathcal{M},j_1}, \cdots, \beta_{\mathcal{M},j_s})^\top \in \mathbb{R}^s$. Therefore, $\beta_{\mathcal{M}}$ contains the non-zero coefficients (i.e., $\beta_{\mathcal{M}}$) as well as the zero ones.

We next aim to give the estimation procedure. It is noteworthy that the response Y in (2.1) takes the form $Y = (I - \rho W)^{-1}(\mathbb{X}\beta + \mathcal{E})$. Therefore, Y explicitly contains the information of $\mathcal{E}$. Consequently, direct least squares type estimation (i.e., minimizing $\|Y - \rho WY - \mathbb{X}\beta\|^2$) may introduce endogeneity and thus can be biased (Lee, 2004). As an alternative, we write the quasi-loglikelihood function as $\ell(\rho, \beta_{\mathcal{M}}) =$

$$\log\left|I - \rho W\right| - n/2 \log\left[\left\{(I - \rho W)Y - \mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}}\right\}^\top \left\{(I - \rho W)Y - \mathbb{X}_{\mathcal{M}}\beta_{\mathcal{M}}\right\}\right] \tag{2.6}$$

by ignoring some constants. It is remarkable that the quasi-loglikelihood (2.6) is frequently studied by spatial econometrics in recent years (Lee, 2004; Anselin, 2013). The corresponding asymptotic properties are established, which suit the spatial dataset very well. However, some conditions might be stringent (e.g., bounded column summation of W) when applied to the network data, especially when the network is in large scale.

Moreover, it is worthy noting that in (2.6), the dimension of $\beta_{\mathcal{M}}$ will

be slowly diverging 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,

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

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

$$\ell_1(\rho) = \log\left|I-\rho W\right| - n/2\log\left[Y^{\top}(I-\rho W^{\top})(I-P_X)(I-\rho W)Y\right], \tag{2.8}$$

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 $\ell_1(\rho)$ one could obtain $\widehat{\rho} = \arg\max_{\rho}\ell_1(\rho)$. To study the asymptotic properties of $\widehat{\rho}$ obtained in network, even in large-scale network, we give the following conditions.

(C6) (Network Structure)

(C6.1) (Connectivity) Let the set of all the nodes $\{1,\cdots,n\}$ 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 $\pi = (\pi_i)^{\top} \in \mathbb{R}^n$ to be 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 constraint on the network structure. Similar assumptions are assumed by the recent network vector autoregression model proposed by Zhu et al. (2018). Specifically, (C6.1) requires certain connectivity holds for the network structure. It is essentially assumed that the nodes in the network are reachable to each other. This implies two arbitrary nodes should be connected with a finite path in the network, which fits the famous six degrees of separation theory (Newman et al., 2006). The second condition assumes the certain type of uniformity for the network. Particularly, it requires the diverging speed of $\lambda_{\max}(W^*)$ should be sufficiently slow. Consequently, we have the following theorem.

Theorem 2. *Assume the 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 in the supplementary material. By Theorem 2, it is concluded 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 by (2.7). The finite performance of $\widehat{\rho}$ and $\widehat{\beta}_{\mathcal{M}}$ will be illustrated by a number of simulation studies in the next section.

3. Numerical Studies

3.1 Data Generation

We consider the following 4 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 ($k = 1, \cdots, K$) 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. For all the examples, the covariance matrix of $\mathcal{E}$ is set to be $\sigma^2 I_n$ with $\sigma^2 = 1$; and ρ is set to be 0.8. We illustrate the generation of $\mathbb{X}$ in each example, and the responses can be generated by model (2.1) accordingly. In each example, n is fixed to be 500 and p =2,000, 5,000.

Example 1 (Independent Predictors). This example is adopted by Fan and Lv (2008) with $\mathcal{M}_T = \{1, 2, \cdots, d_0\}$, 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 ($1 \leq j \leq d_0$) 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 standard normal distribution.

Example 2 (Autoregressive Correlation). We consider in this example an autoregressive type correlation structure. In this structure, the 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 $(1 \le j_1, j_2 \le p)$. The 1st, 4th, and 7th components of β are given by 0.3, 0.2, 0.2, respectively. Other components of β are fixed to be 0.

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, $\mathbb{X}_j$s are 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 $1 \le j_1, j_2 \le p$. The first 3 coefficients of β are fixed to be 0.3. And others are fixed to be 0.

Example 4 (A Challenging Case). In this case, we consider a challenging case, where the network structure is involved in the generation of predictors. Specifically, the predictors $\mathbb{X}_j$s are generated as follows. First, the first d_0 covariates are sampled independently from 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 $\mathbb{E}_j$s independently follow the multivariate normal distribution $N(\mathbf{0}_n, I_n)$. Then d_0 is set to be 3. The first d_0 coefficients of β are fixed to be 0.5 and others are fixed to be 0, i.e., $\mathcal{M}_T = \{1, 2, 3\}$. In this example, the network structure is adopted as $a_{i(i+1)} = 1$ for $1 \leq i \leq n$ for computation simplicity. It is remarkable 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 with the following two popular screening methods and the oracle screening procedure.

- Sure independent screening method (SIS, Fan and Lv 2008), which use sample Pearson correlation between Y and $\mathbb{X}_j$ for feature screening.

- Sure independent screening method based on distance correlation (DC-SIS, Li et al. 2012b). The distance covariance between two random vectors could be defined based on characteristic functions. Thus distance correlation could be defined for multidimensional vectors, correspondingly. In this way, the DC-SIS method allows for multidimensional response. In this paper, we apply the method using the distance correlation between (Y, WY) and $\mathbb{X}_j$ for feature screening.

- Oracle procedure, which use the sample Pearson correlation coefficient between $\mathbb{X}_j$ and $(I_n-\rho W)Y$ with plugging-in the true value of ρ. Since ρ is unknown. We refer to this procedure as an oracle procedure, and labelled as *Oracle* in Tables 1 – 4.

The first one is based on the traditional feature screening procedure. The second one considers the model free method of DC-SIS with multiple responses (Y, WY). The third one is proposed for feature screening based on the known network information, since we know that $(I_n - \rho W)Y = \mathbb{X}\beta + \mathcal{E}$. The last one is an ideal estimator since 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 details in the Section S6 of the supplementary material. We first calculate the average model size after tuning parameter selection as MS$= M^{-1}\sum_m \text{MS}^{(m)}$, where $\text{MS}^{(m)} = |\widehat{\mathcal{M}}^{(m)}|$ in the mth replication. A smaller MS implies a more compact screening model. Next, we evaluate the screening performance in details 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 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, namely $\mathrm{CSP}_j^s = M^{-1}\sum_m I(j \in \widehat{\mathcal{M}}^{(m)})$, is reported to reflect model recover ability. We repeat the experiment for M=200 times to evaluate a reliable result.

The detailed results of the simulations are given in Tables 1 – 4. Considering of $\bar{r}_j$, the oracle procedure has the smallest for all the examples as we expect, which is mainly due to knowing the network effect ρ and model size in advance. It is remarkable that the newly proposed NW-SIS has better performance than the SIS and DC-SIS in terms of both $\bar{r}_j$ and CSP_j^s, which are almost as good as the oracle procedure. In addition, NW-SIS is able to achieve a compact model size (with lower MS) than the other two methods after the selection of tuning parameter. In the last example, as we expect, $\mathbb{X}_1$ is easier to be recovered than $\mathbb{X}_2$ and $\mathbb{X}_3$, for both the Oracle procedure and the proposed NW-SIS. The reason could be easily explained. Define Corr_j to be the Pearson correlation coefficient between $\mathbb{X}_j$ and $(I_n - \rho W)Y$. By the design of Example 4, it could be calculated explicitly that $|\mathrm{Corr}_j/\mathrm{Corr}_1| = |\rho|/(2.21 + \rho^2) < 1$ for $j > 3$ (since $(I_n - \rho W)^{-1}$ can be expressed explicitly in this case). Thus, the first feature is relatively easy to identify. However, due to the correlation between $\mathbb{X}_j$ $(j > 3)$ and Y, recovery of $\mathbb{X}_2$ and $\mathbb{X}_3$ is more difficult. We could see that the performance

of NW-SIS is better than SIS and DC-SIS in this case.

3.3 Results of Parameter Estimation

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

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

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

The simulation result is given in Table 5. We could make the following conclusions. First, the CP values for all examples quickly increase towards 100% as the sample size n increases. This corroborates with the strong screening consistency property, which we have defined in (2.4) and (2.5). Second, the RSSE_ρs are decreasing as n increases, as explained by Theorem 2. Lastly, it is also observed that the RSSE_σ^2s and RSSE_βs steadily decrease as n increases for all the examples.

3.4 Financial Feature Screening for Stock Returns

We next illustrate a real data example with 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 Shanghai Stock Exchange and Shenzhen Stock Exchange. The corresponding response Y_i is the annualized return of stock i $(1 \leq i \leq n)$ in the year 2014.

To construct the network relationship between the stocks, the common shareholders of the stocks are considered. Specifically, it takes the following steps. First, the top ten shareholders' information for each stock are collected, which are defined as *major shareholders*. Second, for $i \neq j$, if the ith stock and jth stock share at least one major common 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%. Besides the response (i.e., Y_i) and net-

work information (i.e., A), the firm specific financial indexes in the previous year (i.e., year 2013) are considered as explanatory covariates. The financial indexes are collected from financial statements of the firm, namely, the balance sheet, the income statement, and the cash flow statement released in 2013. Besides, the interaction effects between $\mathbb{X}_{j_1}$ and $\mathbb{X}_{j_2}$ within the same financial statement are also taken into consideration, which is defined as $\mathbb{X}_{j_1}\mathbb{X}_{j_2}$. This leads to a total of p =796 predictors.

We then conduct the NW-SIS analysis. Particularly, $\widehat{\mathbf{R}}_j^2$ is calculated for $j = 1, \cdots, p$. Next, the covariates are ranked according to the decreasing order of $\widehat{\mathbf{R}}_j^2$ values. Particularly, the covariates with the top 8 highest $\widehat{\mathbf{R}}_j^2$s are given in Figure 1. They are mostly related to the asset (i.e., Asset Impairment Loss, Capital Reserve Fund, Deferred Tax Asset, Intangible Assets), liability (i.e., Short Term Loan, Total Liability), liquidity (i.e., Cash Equivalents), and Financial Expense of the firm.

Next, we compare NW-SIS with SIS and DS-SIS by model fitness levels. First we conduct the screening procedure of all the three approaches. We then compare the fitness level of different methods with varying the model size $|\mathcal{M}| = 1, \cdots, 200$. The estimation is conducted as follows. For the SIS method, we follow Fan and Lv (2008) to estimate a linear regression

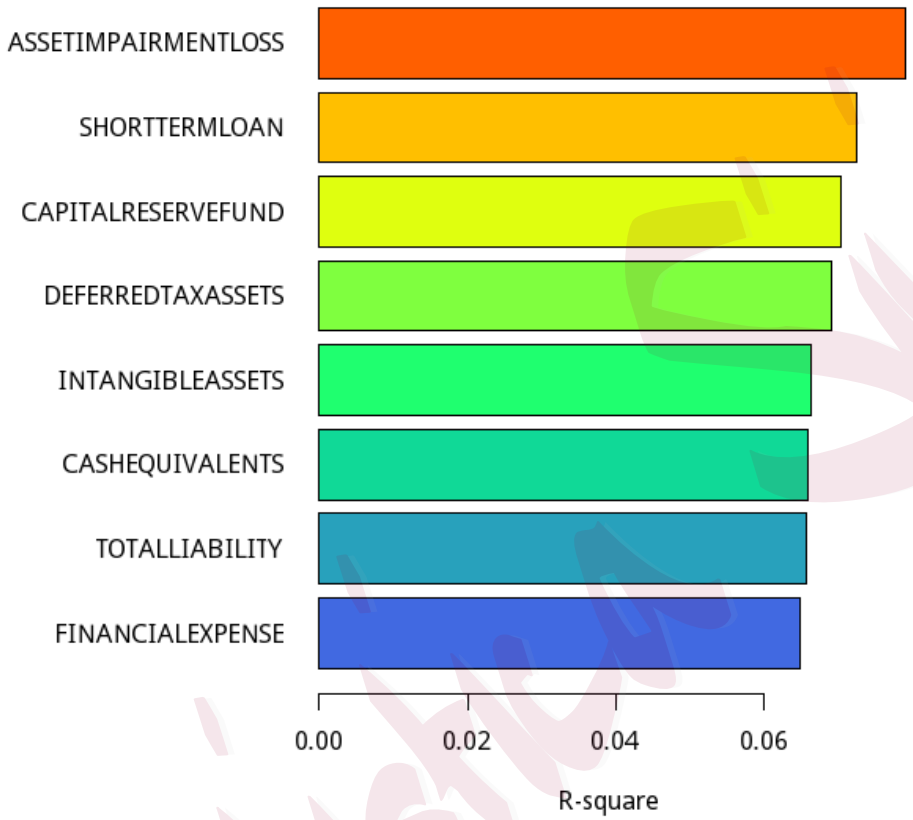

Figure 1: Covariates with top 8 $\widehat{\mathbf{R}}_j^2$. They are related to the asset (i.e., Asset Impairment Loss, Capital Reserve Fund, Deferred Tax Asset, Intangible Assets), liability (i.e., Short Term Loan, Total Liability), liquidity (i.e., Cash Equivalents), and Financial Expense of the firm.

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 $\widehat{\beta}_{\mathcal{M}}$ since the multivariate information is taken into consideration 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 by the adjusted R^2, which is displayed Figure 2 for illustration. We could see from the figure that, as more features are included, the adjusted R^2 increases at first for all the three methods. Next, the adjusted R^2 of NW-SIS achieves the peak value at $|\mathcal{M}| = 75$, which is 25.8% and the highest of the three methods. Consequently, compared to the other competing methods, it can be concluded that NW-SIS can obtain better fitness level with less features.

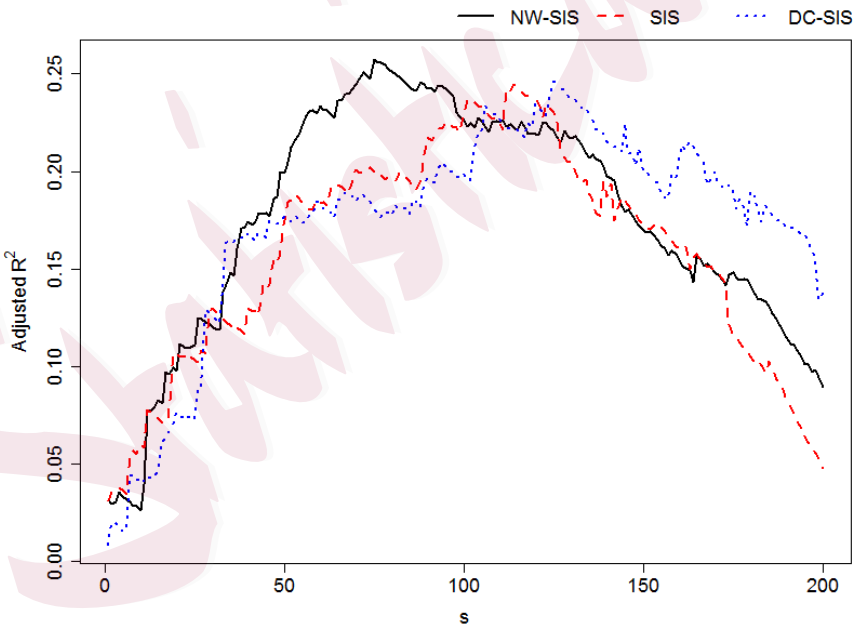

Figure 2: The fitted adjusted R^2 against the screening model size s for the three screening methods: NW-SIS, SIS, and DC-SIS. The adjusted R^2 of NW-SIS achieves the peak value first at $s = 75$, which is 25.8% and the highest of the three methods.

4. Conclusion

In this paper, we propose a network-based independence screening approach. This approach takes the network structure into consideration. We rigorously show the proposed NW-SIS method enjoys the strong screening consistency. The properties of parameter estimation are established subsequently. The proposed method is lastly applied to a financial dataset, which screens financial indexes effectively with respect to stock returns.

To conclude the article, we discuss several topics for future research. First, the responses considered in this article 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. One could make it more flexible to allow it with more sophisticated structures (e.g., autoregressive structures). As a result, the estimation efficiency could be improved. Third, in the numerical study, we show that the tuning parameter selection method performs well. Theoretically, the properties of the tuning parameter selection should be further investigated. Lastly, one should note that unimportant features are typically included in the post-screening set since the screening technique tend to over select the features. Consequently, appro-

priate variable selection methods are worth further investigation after the screening procedure and precisely identify the true model.

Supplementary Materials

Supplementary Materials are available in the attached file which contains useful lemmas, the proof of Proposition 1, the proofs of Theorems 1–2, a corollary from Theorem 1, and a discussion about the tuning parameter selection.

Acknowledgment

Danyang Huang is supported by National Natural Science Foundation of China (NSFC, 11701560,71873137), fund for building world-class universities (disciplines) of Renmin University of China; Xuening Zhu (*xueningzhu@fudan.edu.cn*) is supported by National Nature Science Foundation of China (NSFC, 11901105, U1811461, 11690014, 11690015), Shanghai Sailing Program for Youth Science and Technology Excellence (19YF1402700), Fudan-Xinzailing joint research centre for big data, School of Data Science, Fudan University. Runze Li is supported by National Institute on Drug Abuse (NIDA) grants P50 DA039838, and National Science Foundation grant, DMS 1820702. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA or the NIH; Hansheng Wang's research is partially supported by National

Natural Science Foundation of China (No. 11831008, 11525101, 71532001). It is also supported in part 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., C. 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., C. 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*. 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. *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. *New York: Chapman & Hall*.

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

Li, R., Zhong, W., and Zhu, L. (2012b). 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., a. T. T. W., Gaudinier, A., ., and Handakumbura, P. P. (2015). An arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature*. 517– 571.

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.

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

E-mail: dyhuang89@126.com

School of Data Science, Fudan University, Shanghai, China

E-mail: xueningzhu@fudan.edu.cn

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

E-mail: rzli@psu.edu

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

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

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 $\mathbb{X}_j$, respectively. 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)			
2000	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)			
5000	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

Table 2: Screening Simulation Results for Examples 2. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$, respectively. 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
		$\bar{r}_j(\text{CSP}_j^s)$			
2000	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
		$\bar{r}_j(\text{CSP}_j^s)$			
5000	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 Examples 3. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$, respectively. 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
		$\bar{r}_j(\text{CSP}_j^s)$			
2000	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
		$\bar{r}_j(\text{CSP}_j^s)$			
5000	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

Table 4: Screening Simulation Results for Examples 4. The average rank $\bar{r}_j$ and correctly selected probability CSP_j^s (%) are reported for each predictor $\mathbb{X}_j$, respectively. 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
		$\bar{r}_j(\text{CSP}_j^s)$			
2000	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
		$\bar{r}_j(\text{CSP}_j^s)$			
5000	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

Table 5: Parameter Estimation Simulation Results with 200 Replications for Examples 2-4. The coverage probability CP(%), the root of the sum squared error for ρ (RSSE$_\rho$), σ^2 (RSSE$_{\sigma^2}$), β (RSSE$_\beta$) are reported.

p	n	CP(%)	RSSE$_\rho$	RSSE$_{\sigma^2}$	RSSE$_\beta$
Example 2					
5000	200	27.50	0.0195	0.1717	0.4770
	500	97.00	0.0162	0.0689	0.2452
	1000	100.00	0.0138	0.0369	0.1550
Example 3					
5000	200	34.50	0.0174	0.1455	0.5652
	500	99.00	0.0139	0.0669	0.3038
	1000	100.00	0.0127	0.0392	0.2086
Example 4					
5000	200	18.00	0.0509	0.1396	0.6242
	500	84.00	0.0246	0.0717	0.2684
	1000	99.00	0.0163	0.0394	0.1537